ISSN 0142-6001 (PRINT)
ISSN 1477-450X (ONLINE)

APPLIED LINGUISTICS

Volume 31 Number 3 July 2010

Published in cooperation with
AAAL American Association for Applied Linguistics
AILA International Association of Applied Linguistics
BAAL British Association for Applied Linguistics

www.applij.oxfordjournals.org

OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

EDITORS

REVIEWS AND FORUM EDITOR

ADVISORY BOARD

EDITORIAL PANEL

Applied Linguistics is published five times a year in February, May, July, September and December by Oxford University Press, Oxford, UK. Annual subscription price is £254/US$457/€381.
Applied Linguistics is distributed by Mercury International, 365 Blair Road, Avenel, NJ 07001, USA.
Periodicals postage paid at Rahway, NJ and at additional entry points.
US Postmaster: send address changes to *Applied Linguistics* (ISSN 0142-6001), c/o Mercury International, 365 Blair Road, Avenel, NJ 07001, USA.

Typeset by Glyph International, Bangalore, India
Printed by Bell and Bain Ltd, Glasgow

Applied Linguistics Journal online

The full text of *Applied Linguistics* is available online to journal subscribers. Online access has a number of advantages:

- quality PDFs ensure articles look the same as the print original and are easy to print out
- access is easy—all you need is your subscription number or institutional IP address (see below)
- online access is available ahead of print publication—so view while you await your print version!
- access the text wherever you are (or from any part of your institution network if you have a library subscription)
- perform searches by word or author across the full text of the articles of any part of the journal
- download articles whenever you choose—you will be able to access past online issues as long as you have a current subscription
- free sample copy available online
- fully searchable abstracts/titles going back to volume 1
- Table of Contents email alerting service.

The print version will continue to be available as previously. Institutions may choose to subscribe to the print edition only, online only, or both. Individual subscribers automatically receive both.

CONTRIBUTORS

There is no need for contributors to format their articles any differently; online files are produced automatically from the final page proofs of the journal. However, if you know that an item in your list of references is available online, please supply the URL. If you have your own website, you are welcome to include the URL with your contact address in your biodata.

ADVANCE ACCESS

Applied Linguistics now has Advance Access articles. These are papers that have been copyedited and typeset but not yet paginated for inclusion in an issue of the journal. More information, including how to cite Advance Access papers, can be found online at http://www.applij.oxfordjournals.org.

APPLIED LINGUISTICS

SUBSCRIPTION INFORMATION

A subscription to *Applied Linguistics* comprises 5 issues.

Annual Subscription Rate
(Volume 31, 5 issues, 2010)

Institutional.
Print edition and site-wide online access:
£254.00/US$457.00/€381.00;

Print edition only:
£233.00/US$419.00/€350.00;

Site-wide online access only:
£212.00/US$382.00/€318.00.

Personal.
Print edition and individual online access:
£82.00/US$164.00/€123.00.

Please note: US$ rate applies to US & Canada, Euros applies to Europe, UK£ applies to UK and Rest of World.

Prices include postage by surface mail, or for subscribers in the USA and Canada by airfreight, or in India, Japan, Australia and New Zealand, by Air Speeded Post. Airmail rates are available on request. There are other subscription rates available for members of AAAL, BAAL, AILA, and LSA, for a complete listing please visit www.applij.oxfordjournals.org/subscriptions.

Full prepayment, in the correct currency, is required for all orders. Orders are regarded as firm and payments are not refundable. Subscriptions are accepted and entered on a complete volume basis. Claims cannot be considered more than FOUR months after publication or date of order, whichever is later. All subscriptions in Canada are subject to GST. Subscriptions in the EU may be subject to European VAT. If registered, please supply details to avoid unnecessary charges. For subscriptions that include online versions, a proportion of the subscription price may be subject to UK VAT. Personal rate subscriptions are only available if payment is made by personal cheque or credit card and delivery is to a private address.

The current year and two previous years' issues are available from Oxford Journals. Previous volumes can be obtained from the Periodicals Service Company at http://www.periodicals.com/oxford.html or Periodicals Service Company, 11 Main Street, Germantown, NY 12526, USA. Email: psc@periodicals.com. Tel: +1 (518) 537 4700. Fax: +1 (518) 537 5899.

For further information, please contact: Journals Customer Service Department, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK.
Email: jnls.cust.serv@oxfordjournals.org.
Tel (and answerphone outside normal working hours): +44 (0)1865 353907.
Fax: + 44 (0)1865 353485. In Japan, please contact: Journals Customer Services, Oxford Journals, Oxford University Press, Tokyo, 4-5-10-8F Shiba, Minato-ku, Tokyo 108-8386, Japan. Tel: +81 3 5444 5858.
Fax: +81 3 3454 2929.

SUBSCRIBE TO APPLIED LINGUISTICS

For new subscriptions and recent single issues only. Current subscribers will automatically receive a renewal form.

Please complete the form below and return it to: Journal Customer Service Department (please see above).

Please record my subscription to Applied Linguistics, starting with Volume__________
(Subscriptions start with the March issue and can be accepted for complete volumes only.)

Please send me the following single issue(s)
Volume_________ Issue_________

Name (BLOCK CAPITALS please)

____

Adresss _________________________________

____

____

City ____________________________________

Country _________________________________

Postcode ________________________________

I enclose the correct payment of (see rates above):
£/US/€ __________________________________

Please debit my credit card:

American Express / Mastercard / Visa
(delete as appropriate)

Card number:
__|__|__|__|__|__|__|__|__|__|__|__|__|__|__|__|__
Expiry date: |__|__|__|

Signature _______________________________

□ Please tick this box if you do NOT wish to receive details of related products and services of OUP and other companies that we think may be of interest.

AIMS

Applied Linguistics publishes research into language with relevance to real-world problems. The journal is keen to help make connections between fields, theories, research methods, and scholarly discourses, and welcomes contributions which critically reflect on current practices in applied linguistic research. It promotes scholarly and scientific discussion of issues that unite or divide scholars in applied linguistics. It is less interested in the *ad hoc* solution of particular problems and more interested in the handling of problems in a principled way by reference to theoretical studies.

Applied linguistics is viewed not only as the relation between theory and practice, but also as the study of language and language-related problems in specific situations in which people use and learn languages. Within this framework the journal welcomes contributions in such areas of current enquiry as: bilingualism and multilingualism; computer-mediated communication; conversation analysis; corpus linguistics; critical discourse analysis; deaf linguistics; discourse analysis and pragmatics; first and additional language learning, teaching, and use; forensic linguistics; language assessment; language planning and policies; language for special purposes; lexicography; literacies; multimodal communication; rhetoric and stylistics; and translation. The journal welcomes both reports of original research and conceptual articles.

The Journal's Forum section is intended to enhance debate between authors and the wider community of applied linguists (see Editorial in 22/1) and affords a quicker turnaround time for short pieces. Forum pieces are typically responses to a published article, a shorter research note or report, or a commentary on research issues or professional practices. The Journal also contains a Reviews section.

Applied Linguistics is covered by the following abstracting/indexing services: Bibliographie Linguistique/Linguistic Bibliography, BLonline, British Education Index, Current Index to Journals in Education, ERIC (Education Resources Information Centre), International Bibliography of the Social Sciences, ISI: Social Sciences Citation Index, Research Alert, Current Contents/Social and Behavioral Sciences, Social Scisearch, Sociological Abstracts: Language and Linguistics Behaviour Abstracts, Language Teaching, MLA Directory of Periodicals, MLA International Bibliography, PsycINFO, Sociological Abstracts, Zeitschrift für Germanistische Linguistik.

ADVERTISING

Inquiries about advertising should be sent to Linda Hann, Oxford Journals Advertising, 60 Upper Broadmoor Road, Crowthorne, RG45 7DE, UK. Email: lhann@talktalk.net. Tel/Fax: +44 (0)1344 779945.

PERMISSIONS

For information on how to request permissions to reproduce articles/information from this journal, please visit www.oxfordjournals.org/permissions.

DISCLAIMER

Statements of fact and opinion in the articles in *Applied Linguistics* are those of the respective authors and contributors and not of *Applied Linguistics* or Oxford University Press. Neither Oxford University Press nor *Applied Linguistics* make any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

NOTES TO CONTRIBUTORS

Articles submitted to *Applied Linguistics* should represent outstanding scholarship and make original contributions to the field. The Editors will assume that an article submitted for their consideration has not previously been published and is not being considered for publication elsewhere, either in the submitted form or in a modified version. Articles must be written in English and not include libelous or defamatory material. Manuscripts accepted for publication must not exceed 8,000 words including all material for publication in the print version of the article, except for the abstract, which should be no longer than 175 words. Additional material can be made available in the online version of the article. Such additions will be indexed in the print copy.

Applied Linguistics operates a double-blind peer review process. To facilitate this process, authors are requested to ensure that all submissions, whether first or revised versions, are anonymized. Authors' names and institutional affiliations should appear only on a detachable cover sheet. Submitted manuscripts will not normally be returned.

Forum pieces are usually reviewed by the journal Editors and are not sent for external review. Items for the Forum section are normally 2,000 words long. Contributions to the Forum section and offers to review book publications should be addressed to the Forum and Reviews Editor.

For more detailed guidelines, see our website http://www.oxfordjournals.org/applij/for_authors/index.html

PROOFS

Proofs will be sent to the author for correction, and should be returned to Oxford University Press by the deadline given.

OFFPRINTS

On publication of the relevant issue, if a completed offprint form has been received stating gratis offprints are requested, 25 offprints of an article, forum piece or book review will be sent to the authors free of charge. Orders from the UK will be subject to a 17.5% VAT charge. For orders from elsewhere in the EU you or your institution should account for VAT by way of a reverse charge. Please provide us with your or your institution's VAT number.

COPYRIGHT

Acceptance of an author's copyright material is on the understanding that it has been assigned to the Oxford University Press subject to the following conditions. Authors are free to use their articles in subsequent publications written or edited by themselves, provided that acknowledgement is made of *Applied Linguistics* as the place of original publication. Except for brief extracts the Oxford University Press will not give permission to a third party to reproduce material from an article unless two months have elapsed without response from the authors after the relevant application has been made to them. It is the responsibility of the author to obtain permission to reproduce extracts, figures, or tables from other works.

CONTENTS

ARTICLES

FORUM

REVIEWS

NOTES ON CONTRIBUTORS

APPLIED LINGUISTICS

Volume 31 Number 3 July 2010

CONTENTS

ARTICLES

FORUM

REVIEWS

Applied Linguistics: 31/3: 325–345 © Oxford University Press 2009
doi:10.1093/applin/amp028 Advance Access published on 30 June 2009

'Tails' of Linguistic Survival

IVOR TIMMIS

Leeds Metropolitan University, UK

Given the relatively short history of computerized corpora of spoken language, it is not surprising that few diachronic studies have been done on the grammatical features recently highlighted by the analysis of such corpora. This article, however, does take a diachronic perspective on one such feature: the syntactic feature of 'tails' (Dik 1978). The use of tails is analyzed in terms of form, frequency, and function in a 50,000 word corpus of informal conversations which took place in the North of England between 1937 and 1940. This analysis shows that tails were a systematic and quite frequent feature of spoken English at that time. It also shows that there are marked similarities in terms of form and function between tails in this small corpus and those in more widely based contemporary corpora. The article argues that the durability of tails may lie in the fact that the feature has both an important psycholinguistic function and important affective functions and concludes that this kind of diachronic research is of great potential value for spoken language research.

INTRODUCTION

Recent years have seen a multitude of corpus-based descriptive insights into the grammar of spoken English, perhaps best reflected in the publication of substantial sections on the grammar of conversation in the *Longman Grammar of Spoken and Written English* (Biber *et al*. 1999) and the *Cambridge Grammar of English* (Carter and McCarthy 2006). The general tenor of these insights has been to suggest that the written-based grammars which have traditionally held sway in linguistics have not paid due heed to 'features that occur widely in the conversation of native speakers of English, across speakers of different ages, sexes, dialect groups, and social classes, with a frequency and distribution that simply cannot be dismissed as aberration' (McCarthy and Carter 1995: 142). The starting point for this article, however, is the simple observation that a new description of a feature or a new perspective on a feature does not necessarily imply that the feature described is itself new or that it has recently acquired a new function. What is undoubtedly new, thanks to technological improvements in recording equipment and analytical software, is the quantity and quality of the evidence about conversational English available to the linguist. Thus far, however, there seems to have been little attempt to take a diachronic perspective on the features which this evidence has brought to light. Given that recent descriptive insights into conversational English have been so dependent on technological advances, this lack of a diachronic perspective is perfectly understandable: how can we hope to find linguistic evidence of the quality required to make worthwhile diachronic observations from eras when the recording equipment and analytical software which underpin current spoken corpus work were not available? The argument of this article, however, is that a small corpus of conversations which took place in Bolton, Lancashire, UK, between 1937 and 1940, though rudimentary by modern standards, *is* of sufficient quality to allow an interesting and valuable diachronic perspective on a particular feature of spoken grammar—'tails' (e.g. Dik 1978; Aijmer 1989; McCarthy and Carter 1995; McCarthy 1998), also known as 'right dislocation' (e.g. Dik 1981; Geluykens 1987; Aijmer 1989)—and that this perspective may offer insights into why some features of spoken language prove more durable than others. Given the unusual nature of the corpus, we need to look in some detail at its genesis and to consider its credentials in some depth before using it to examine the use of tails in Bolton in the period 1937–1940.

THE NATURE OF THE DATA

The corpus I am referring to—henceforward 'The Bolton Corpus'—consists of over 50,000 words of conversational English culled from the Mass Observation archives. While this is not the place for a detailed history of Mass Observation, some understanding of the movement is needed to explain how and why the conversations came to be recorded or, more accurately, *written down*. Mass Observation was founded in 1937 by Charles Madge, Humphrey Jennings, and Tom Harrisson, and was essentially concerned with making a detailed sociological and anthropological study of the working classes of Britain (Jeffrey 1999). As the name suggests, a distinctive feature of Mass Observation was that studies were to be carried out by large teams of observers who would infiltrate, in various ways, the communities they were studying. The anthropological aspect of the movement, which is an important factor to consider when we assess the reliability of the data, is probably best summed up by the following remark by Harrisson (1974: 5): '...it was slowly borne in upon me that while anthropologists were generously financed to go all over the world studying so-called primitive peoples, no one at that time was making comparable studies of ourselves.' In 1937, Harrisson, 'more or less by chance', chose Bolton, a textile town in the North of England, to be the focus of a particularly detailed and in-depth Mass Observation study of the working classes, took charge of the project himself, and assembled a team of observers to record diverse aspects of working class behavior in Bolton, or Worktown, as it was known in Mass Observation circles. The resulting archive contains masses of documents from the time and large numbers of reports and observations of various aspects of working class behavior and attitudes in Bolton. Among these documents and reports, one can find written records of conversations or, more precisely, conversational snippets, which took place at the time. Assembling the corpus, then, has largely been a painstaking process of using the archive catalogue to identify those parts of the

Worktown [Bolton] papers of the Mass-Observation archives[1] most likely to yield conversational data, and then scrolling through the microfilm to find the conversational snippets.

The spoken data in the archive material from which the corpus is drawn is of two basic types:

1 Conversations and fragments of conversations overheard and transcribed 'live' by observers, often operating incognito, in a variety of locations. The 'overheards', as they are known in the archive, which I have included in the Bolton Corpus mostly took place in pubs, in and around sports grounds (especially the bowling green), on streets, and in public buildings in the town centre. The observers even followed the people of Bolton on their annual holidays in Blackpool, a popular seaside resort not far from Bolton, and there is conversational data from similar venues there.

2 Oral comments elicited by observers in response to specific questions. These questions could be 'direct'—when it was obvious the observer was carrying out a survey of some kind—or 'indirect,' when the questions were infiltrated in the course of an apparently normal conversation.

As the intention is to capture natural conversation, the Bolton Corpus relies mostly on 'overheards', though 'indirects' and 'directs' have been included where it was clear to me that the observer was trying to capture not only what was said, but how it was said. In other words, I exercised quality control in selecting the spoken data to include in the corpus by excluding conversational snippets which seemed to have been 'cleaned up' and contained *none* of the features one would expect to find in conversational English e.g. contractions, ellipsis, hesitations, and repetitions, incomplete utterances, discourse markers, colloquial language, and so on. Despite this quality control, it is immediately obvious that the Bolton Corpus is rather unorthodox by modern standards:

1 The data were gathered in an opportunistic way: while it is clear that the focus is generally on the Bolton working class, we often lack detailed demographic information about the speakers. We should also note that, as the focus of the Mass-Observation study was mainly anthropological, the observers did not specifically try to capture systematically a range of spoken genres or contexts of interaction.

2 The observers could not make use of recording equipment. This means that the corpus consists mainly of short conversations, fragments of conversations, and isolated comments transcribed 'live'. The transcriptions are in a bastardized orthography which varies somewhat between different observers. I have kept examples from the Bolton Corpus in this article in their original transcriptions so that readers can judge for themselves the verisimilitude of the data, but I have provided a 'translation' into

Standard English where the example is in broad dialect. It could also be argued that the observers, most of whom were neither from Bolton nor working class, might be predisposed, consciously or subconsciously, to record what seemed to them exotic or quaint in the speech of Bolton people, particularly as they had no specific linguistic brief.

While acknowledging the limitations of the corpus, I am going to argue, for a number of reasons, that the corpus offers data of sufficient quality and quantity to provide unique and valuable insights into conversational language in Bolton at that time:

1 Attention to detail was central to the approach of Mass Observation and to Harrisson's anthropological approach in particular (Jeffrey 1999). Observers who, among other things, had to count the number of sweets in a sweet shop window and the number of people wearing hats on a Sunday, applied the same rigorous scrutiny to language and clearly tried to capture not just what was said, but exactly how it was said, as the corpus extract below, overheard in a Bolton pub, illustrates:

> *A: I think Farr's buggered now. The Arsenal are a good team.*
>
> *B: Aye, they get beaten when they play bad.*
>
> *A: Bassett's been with them a long time now.*
>
> *B: Look here, they geet him at the same time as the Wanderers geet Taylor, because he should have come here, Bassett, but Bolton thought he were too little so they let him go. Dick Lyn sent him here.*
>
> *A: Aye, that's all right but Taylor's been with Bolton good while, he has had a benefit.*
>
> *B: Aye. Art gooin to Blackpool on Monday?*
>
> *A: Aye. I'st be gooin. I'm taking the child and mother. Well, there's a few on us.*

The attempt to capture the dialect in this example is not only convincing to this researcher, born and brought up in Bolton, but reflects accurately grammatical features described in Shorrocks' (1999) *Grammar of the Dialect of the Bolton Area*. In the extract above, for example, we see the following dialect features which are described in Shorrocks (1999):

- 'bad' as an adverb form
- 'geet' as the past of 'get'
- 'were' as the third person past simple of 'to be'
- 'art' as the second person cliticized form of 'to be' with the pronoun 'thou'
- 'I'st' as the cliticized form of 'I shall'
- 'on' used where standard English would use 'of'

We also see an example of the tail structure which is the focus of this article:

- He should have come here, Bassett

2 There is evidence of impressive attention to detail here and of a sensitive ear for language which is consistent with the anthropological aspect of the Mass Observation study. Many of the conversations have about them, to use Phillips' (2000) phrase, the 'authenticity of the inconsequential' and capture features typical of spoken language in general. The faintly surreal dialogue below, for example, shows the following features of spoken language:

- Situational ellipsis
- Tails
- 'Them' as a demonstrative pronoun
- 'Like' as a discourse marker.

A: Good idea, these revolving doors.

B: Yes, keeps out the draughts.

A: I always think of persons going round and round when I come in.

B: I often think of that, walking round and round, like.

A: I've never been on one of them [moving staircases] yet. I wouldn't like to. I'd be nervous.

3 Labov (1972: 85–6) expressed the following wish about spoken data: 'We must somehow become witnesses to the everyday language which the informant will use as soon as the door is closed behind us: the style in which he argues with his wife, scolds his children, or passes the time of day with his friends'. Thanks in large measure to what would now be regarded as a cavalier disregard for research ethics, this is exactly what the Mass Observation team in Bolton achieved over 30 years earlier, as the earthy and vivid examples later in this article will illustrate.

4 A purely written record of speech is an unorthodox but not unprecedented source for historical corpora. The *Corpus of English Dialogues 1560–1760*, for example, compiled by Kytö and Culpeper (Rissanen 2000), uses witness depositions, trial proceedings and dialogues from drama and prose fiction among its texts. As Kytö and Walker (2003) argue, quality of spoken data is not so much a question of intrinsic value as a question of the value attached to the data for particular purposes and, as Biber *et al.* (1999) point out, for the purposes of grammatical analysis, an accurate written record of speech is quite adequate save for a few exceptional cases where prosodic information would be needed to disambiguate a particular utterance. In this case, I am going to argue that the Bolton Corpus, for the reasons above, offers data of sufficient quality to permit legitimate and valuable grammatical observations,

though it may well be less reliable for phonological purposes or extended discourse analysis.

TAILS FROM THE PAST AND THE PRESENT

For some time now, linguists (Melchers 1983; Quirk *et al.* 1985; Geluykens 1987; Ashby 1988; Aijmer 1989; Fretheim 1995; McCarthy and Carter 1997; McCarthy 1998; Lambrecht 2001; Durham 2007) have recognized a grammatical phenomenon which involves the placement of an extra element either before the canonical s-v-x- clause structure, as in example (a) from the Bolton Corpus, or after the canonical s-v-x- clause structure, as in example (b):

(a) **Most of these navvies**, *they* come in here and have a pint you see.
(b) *They* all want throwing out, **the government**.

As the examples show, the extra element (in bold) is co-referential with an element in the clause (in italics). This phenomenon is almost exclusive to spoken language, but is by no means exclusive to English (Lambrecht 2001). One term for this structure is 'dislocation', on the grounds that a noun phrase has been moved outside the conventional clause structure and replaced by a pronoun. Example (a), therefore, would represent 'left dislocation' and example (b) 'right dislocation'. The focus of this paper is on the structure in example (b) (and its main variants). A number of terms have been used for this structure including 'amplificatory tag statement' (Quirk *et al.* 1972), 'tag statement' (Melchers 1983), 'postponed theme' (Downing and Locke 1992), 'tail' (Geluykens 1987; McCarthy and Carter 1997), and 'noun phrase tag' (Biber *et al.* 1999). The appropriacy of the term 'right dislocation' has, however, been challenged. Lambrecht (2001) questions the appropriacy of 'right dislocation' on the grounds that no movement of the noun phrase takes place in reality, but chooses to continue to use 'right dislocation' for 'convenience'. Ruehlemann (2006) underlines that the prefix 'dis' carries a negative connotation, and, in similar vein, McCarthy and Carter (1997: 407) object to the term on the grounds that it implies that the structure is 'some kind of aberrant variation on a ''normal'' structure', and thus prefer the term 'tails' for 'right dislocation' and 'headers' (Carter and McCarthy 2006) for 'left dislocation'. This paper follows McCarthy and Carter (1997) in arguing that 'metaphors of abnormality' such as 'dislocation' will tend to support a view that spoken language is a defective form of written language and perhaps hinder attempts to account for spoken data in its own right which, as will be seen, is the aim of this paper. 'Tail', which, unlike 'tag', implies an integral connection with the body, is descriptive without being judgemental and is, therefore, the term which will be used henceforward in this article.

Tails are by no means the only recently described feature of spoken language to be found in the Bolton Corpus, but I have chosen to analyze tails in the

Bolton Corpus as this descriptive interest in the feature appears to be particularly recent and as I wanted to present a detailed analysis of a feature which could then be compared with current descriptions of tails in British English in terms of form, frequency, and function.

Frequency and distribution of tails

In the Bolton Corpus, there are 80 examples of tail structures in around 50,000 words, a frequency of 1 in 625 words [normalized: 1.6 per 1,000]. For comparison purposes, Table 1 shows the frequency of tails in the Bolton Corpus and more recent corpora.

The frequency ratings for tails in the contemporary corpora may look low by comparison with the Bolton Corpus, but this should not be taken as an indication that tails are an infrequent feature of contemporary spoken English, nor that they are necessarily less frequent now than they were then: Cullen and Kuo (2007) observe that the frequency rating for tails in the spoken component of the *Longman Corpus of Spoken and Written English* makes it twice as frequent as 'ought to' or the 'get passive', while Carter, Hughes, and McCarthy (1998) describe tails as a 'prominent' feature of the 5 million word CANCODE corpus.

We have already noted that the observers collected spoken data in an opportunistic way without seeking to represent specific spoken genres. This lack of generic representativeness sets limits on the value of the frequency ratings for tails in the Bolton Corpus. While it seems clear that tails were a common feature of speech in that context at that time, the types of conversation collected in the Bolton Corpus will undoubtedly have affected the frequency rating. In terms of contexts of use, the conversations in the Bolton Corpus

Table 1: The frequency of tails in 5 corpora

Corpus	Reference	Frequency (normalized per 10,000 words)
Bolton Corpus (50,000 words)	Author	16 per 10,000
The Longman Corpus of Spoken and Written English (4 million word spoken component)	Cullen and Kuo (2007)	2 per 10,000
London Lund Corpus (170,000 words of spoken extracts)	Aijmer (1989)	3 per 10,000
CANCODE mini-corpus[a] (30,000 words)	Carter and McCarthy (1995)	3.7 per 10,000
York corpus (1.5 million words)	Durham (2007)	2 per 10,000

[a] The Cambridge and Nottingham Corpus of Discourse English.

mostly took place in informal settings where one could assume a high degree of familiarity and shared knowledge between the participants. The most common settings for the spontaneous and informal conversations in the corpus are the pub, the bowling green, the rounders pitch, and town centre streets and public buildings. Some of the conversations are less spontaneous in that they were initiated by the observer when, for example, they went round the pubs asking local people for their opinions on works of modern art or when they stopped people in the street to ask them their opinions of the war situation or the weather. As we shall see more clearly when we come to consider the function of tails, these sub-sets of art evaluation and weather conversations, where the observers *invited* evaluative comments, are particularly likely to have boosted the frequency rating.

McCarthy and Carter (1997: 424) note that the following spoken genres seem to attract tails:

- informal casual conversations
- collaborative, multiparty talk
- comment/elaboration sequences
- narrative recounts where some evaluation is involved
- unplanned spoken commentaries on sports events

Informal, casual conversations and unplanned commentaries on sports events are particularly prominent in the Bolton Corpus though the commentaries on sports events are informal commentaries by spectators rather than broadcasts. Over 9,000 words of the 50,000 words in the Bolton Corpus are from sports-related conversations. The nature of the data may, as we have seen, have skewed the frequency rating for tails in the Bolton data, but it does seem clear that tails were a frequent enough feature to deserve serious analysis.

The structure of tails

In the Bolton Corpus, tails can be divided into two main categories in terms of structure:

1 Noun Phrase Tails i.e. the tail consists only of a noun phrase which is co-referential with the pronoun in the preceding clause e.g.

- They're a clever lot of people, **these Germans**.

This type of tail is recognized by all commentators on the structure and, indeed, is described as the 'canonical' variant by Durham (2007), though she uses the term 'right dislocation'.

2 Operator Tails i.e. the tail structure consists of a co-referential noun phrase *and* an operator e.g.

- Well, it's a funny population, **Bolton is**.

This type of tail is recognized by *inter alia* Melchers (1983), McCarthy and Carter (1997), Ruehlemann (2007) and Durham (2007). Durham (2007) uses the term ‘expanded right dislocation’, but, in view of the discussion of terminology above, ‘operator tail’ would seem to suit our purposes. Both these types of tail are reported by Durham (2007) to be common to all varieties of English.

Each of these categories can be usefully sub-divided. There were at least five examples in the Bolton Corpus in each of the sub-categories described below.

Noun phrase tails

(a) Full noun phrase tails e.g.

- It holds the record, **this pub**, for growing celery, hard to believe. It’s not a bad’un that, 9 feet 5.
- He’s a good singer **yon mon**, but yon pianist’ll knock bottom right out of 2 pianos if he plays like that all neet. [SE: He’s a good pianist, that man, but that pianist will knock the bottom right out of two pianos if he plays like that all night]

The full noun phrase tail structure is a very common form in the Bolton Corpus (23 out of 80 examples) and it is reported as the most common form of tails in their data by Carter and McCarthy (2006: 194): ‘Most commonly, a tail consists of a full noun phrase which clarifies or repeats the referent of a pronoun in the clause that comes before it . . .’ Carter, Hughes, and McCarthy (1998) note that the noun phrase can be complex and there is an example of this in the Bolton data:

- He’s a bloody shithouse **that fellow up yonder** and he wants a thump under the jaw.

There is also an example in the Bolton Corpus of a full noun phrase tail referring back to a full noun phrase which initiates the preceding clause:

- This feller must be well in the 33s, this right back

I have not come across this form in any other descriptions.

(b) (Demonstrative) Pronoun Tails i.e. the tail consists only of a pronoun e.g.

- It’s a serious picture **that**.
- It’s a pretty stiff one **this**.

Shorrocks (1999) reports this pattern as a feature of Bolton dialect, but the pattern is also noted as a structural possibility for tails by Carter and McCarthy (2006), Ruehlemann (2007), and Durham (2007). Aijmer (1989: 150) also reports that: ‘. . . the Tail need not contain a full description. It could be a demonstrative pronoun (that) or another concept with vague deictic reference (that sort of rubbish, the whole thing, this sort of lark).’ Current descriptions, Carter and McCarthy (2006), for example, note that pronouns other than demonstratives could be used alone in the tail position, but curiously there are no examples in the Bolton Corpus, though one might expect an accusative personal pronoun 'her' rather than a demonstrative in the following example:

- She's a good girl, **that**. She never grumbles whether thi' lose or not.

Operator tails

(a) Inverted operator tails i.e. the tail consists of an operator followed by a noun phrase e.g.

- 'ee's no bloody sluvvin **isn't yon mon**[2];'ees played afore.
- Eel watch **wilt ref**. [SE: He'll watch, will the referee]

Durham (2007) uses the term 'reverse right dislocation' for examples where the operator precedes the noun phrase, but, for our purposes, 'inverted operator tail' would seem to be the most appropriate term. Durham (2007) cites research (Wright 1905; Hedevind 1967; Melchers 1983; Petyt 1985; Shorrocks 1985) which suggests that this particular tail variant is restricted to Northern British dialects, particularly Yorkshire and Lancashire. In this respect, it is interesting that the 'inverted operator tail' is recognized by McCarthy and Carter (1997) as a variant of tails, but not by Ruehlemann (2007). Durham (2007) also reports that this tail variant was the one most frequently selected by speakers in the York spoken corpus (Tagliamonte 1996–1998) on which her research is based. In the Bolton Corpus, the operator most frequently selected in the 'inverted operator tail' is copula 'be', which is also the case in Melchers' (1983) and Durham's (2007) data. Another similarity with Durham's data is that personal pronouns do not seem to be selected in the tail in this variant. Even where one might plausibly expect a personal pronoun, it is avoided:

- She's a rum bugger, **is that**, but she's a good batter.
- E can skip it, **can yon one**. [SE: He can skip it (run fast), can that one]

(b) Simple operator tails i.e. the tail consists of a noun phrase followed by an operator e.g.

- That's not human. It's a proper bestiality, **that is**.
- You're a nice set of buggers, **you are**.
- He'll go crashing now ... He'll get a smash in the finish, **Hitler will**.
- I don't know. I think it's a jolly outrage, it's a shame, **I do**.

Other tail forms

There are also five examples of a structure in the Bolton Corpus where the post clause slot is occupied by an evaluative noun, sometimes qualified by an adjective, as in the example below:

- Some of these bloody pacifists want an operation to take out their bloody urine and inject some British blood in them, **the soft buggers**.

While this structure superficially resembles a 'tail', it is different in two important respects:

1. The NP in the post clause slot is co-referential with a full NP in the preceding clause.
2. The NP in the post clause slot is more than simply co-referential and adds extra propositional content.

This structure, as McCarthy (personal communication) puts it, 'seems to lie half-way between the standard co-referential NP tails and the kinds of anaphors we get in: 'John came back without the camera; the silly bugger had left it on the bus.' Indeed, it could be said to have more in common with apposition as the post clause NP is 'moveable'. The following, for example, would be possible, if unlikely *in speech*:

- Some of these bloody pacifists, **the soft buggers**, want an operation to take out their bloody urine and inject some British blood in them.
- Some of these bloody pacifists, **soft buggers that they are**, want an operation to take out their bloody urine and inject some British blood in them.

What this structure suggests, however, is that the post clause slot can carry more than afterthoughts or clarifications and can carry considerable evaluative force. This brings us to the discussion of the function of 'proper' tails.

The function of tails

Without access to the full discourse context of the tails in the corpus, or to the prosodic information, it is not possible to be categorical about the function of the tail in each example. What is clear, however, is that the great majority of tails in the Bolton Corpus are associated with some form of evaluation: no fewer than half the examples of tails in the Bolton Corpus, for example, co-occur with clauses in which there is an evaluative adjective—'good' is the most common with ten instances, and, in addition to other forms of evaluation, we also come across nouns which are clearly evaluative such as 'miracle', 'bestiality' and 'bugger' (three times in this function):

- It was a miracle, that Dunkirk.
- That's a proper bestiality, that is.
- You're a bonny bugger, you are.

It is interesting in this respect that in the Bolton Corpus tails are particularly frequent in the sub-corpus of conversations where the observers asked people in pubs to evaluate paintings (5,900 words) and the sub-corpus of conversations where an observer commented on the weather to passers-by (1,400 words). Table 2 shows the frequency in these sub-corpora compared with the corpus as a whole.

McCarthy and Carter (1997) and Aijmer (1989) note the tendency of tails to co-occur with evaluative comments, but in their data, as in the Bolton Corpus,

Table 2: The frequency of tails in the Bolton Corpus and sub-corpora

Corpus	Frequency (normalized)
Complete Bolton Corpus	16 per 10,000
Art evaluation sub-corpus	31 per 10,000
Weather conversations sub-corpus	46 per 10,000

it is not just a question of 'neutral evaluation'. Carter and McCarthy (1995: 151) argue that tails 'position the speaker in terms of his/her stance or attitude' while Aijmer (1989: 137) notes that tails can convey 'a spontaneous and emotional reaction on something (especially in the immediate context) or an emotionally coloured comment on a situation which is familiar to both the participants in the conversation'. This emotional and attitudinal overlay is very much present in the Bolton data and comes over in the use of 'colourful language' of various kinds in association with the tail structure. It is perhaps in this respect that the ability of the Bolton Corpus to penetrate 'behind closed doors' is most obvious and most useful as we are most likely to get some of this colourful language when speakers are completely off guard. For exemplification purposes, the colourful language falls conveniently into the categories below, though they are not mutually exclusive:

- Strong evaluative adjectives: rum, awful, shocking, numb [to mean stupid], stiff [to mean difficult]
- Strong evaluative nouns: nuisance, sluvvin [sloven], outrage, shame
- Swear words: bloody (7 times), bugger (4 times), pillan [pillock], shithouse, farting [as intensifier]
- Metaphorical and idiomatic language: taking the guts out of us; he's been on th'booze; bored bloody stiff; he can skip it

The three examples below reflect the emotionally coloured aspect of tails particularly well:

- They all want throwing out, the government, taking the guts out of us.
- They all let us down, the bloody Dutch and the Belgians and the French.
- He's a bloody shithouse that fellow up yonder and he wants a thump under the jaw.

There is a noticeable tendency in the Bolton data, as in Melchers' (1983) data, for evaluations of people to be in the third person. Out of 21 examples, only 3 are second person evaluations and they are sarcastic and/or derogatory. There are interesting echoes here of Strässler's (1982) observation that idioms, when used evaluatively, often refer to a third person: in the case of both idioms and tails it could be that the emotional colour or attitudinal overlay threatens face. Melchers' (1983) notes that inverted operator tails generally co-occur with positive evaluations; this is also the case in the Bolton Corpus, but not exclusively so. More data would be needed to confirm this, but it may be the case that negative evaluations with *inverted* operator tails are more likely when there is an evaluative noun in the preceding clause e.g.

- He's a nuisance is that man

Shorrocks (1999) and Carter and McCarthy (2006) comment on the emphatic potential of tails, which can be seen in the examples immediately below:

- I don't know. I think it's a jolly outrage, it's a shame, I do.
- He'll go crashing now . . . He'll get a smash in the finish, Hitler will.

This emphatic potential seems, however to be very closely related to the emotionally coloured and evaluative aspect of tails rather than to constitute a separate function.

Geluykens (1987: 122) argued that the main function of tails was as a repair mechanism to add afterthoughts: 'Tails are one specific instantiation of the repair mechanisms which are available to ''correct'' unplanned spoken discourse. Tails are a form of self-repair which is (mostly) self-initiated . . .' However, Aijmer (1989) argues that afterthoughts can be distinguished from tails on the basis of their intonation, noting that afterthoughts are typically marked by falling intonation whereas other types of tails are often marked by rising intonation. Similarly Fretheim (1995) argues that Norwegian tails are prosodically linked to the preceding clause in ways that afterthoughts are not. Lambrecht (1987: 234) is also unconvinced that tails (though he uses the term 'right dislocation') can be accounted for as afterthoughts: '. . . the speaker who uses [a tail] is fully aware that the mere mention of the pronoun is insufficient'. 'Afterthoughts' do not seem to be a convincing explanation of tails in the Bolton corpus, particularly in view of their frequency in the sub-corpus of art conversations when speakers had ample time to compose their utterances.

While it may be misleading to characterize tails as afterthoughts, it may be legitimate to speak of a retrospective aspect to tails. As we have seen, tails typically consist of a noun phrase in the post clause slot which is co-referential with a pronoun in the preceding clause. Tails have the potential, then, to amplify or clarify a pronoun in the preceding clause and 'postponed identification' or 'disambiguation' is indeed one of the functions ascribed to tails by Aijmer (1989). Ashby (1988: 220) considers this to be a major function of tails in his corpus of spoken French: 'While the term ''afterthought'' . . . seems inappropriate, a major function of the RDs [right dislocations] in my corpus does seem to be that of clarifying the identity of the referent about whom an assertion is being made'. It is possible that, as Aijmer (1989) suggests, the speaker decides mid-utterance that the pronoun needs further clarification and this is quite plausibly the case in the two examples below, particularly the first, where the observer's use of dashes implies that s/he heard the tail as a kind of afterthought:

- Well, I think it's what we want—unity—if we can get strength through it.
- It's awful, isn't it, Tuesday night?

In this function, tails have psycholinguistic value as they can form part of composite utterances which allow the speaker 'to cope with planning pressure, and at the same time to convey some fairly complex messages' (Biber *et al.* 1999: 1072).

Postponed identification or disambiguation are also *plausible* functions of many other examples of tails in the Bolton Corpus, but without access to the full discourse context it is difficult to make categorical judgements. Indeed, it would be difficult even with access to the full discourse context as it is essentially a question of trying to work out what is going on in the speaker's mind. Ashby (1988) and Fretheim (1995), however, both point out that the occurrence of pronouns in the tail position indicates strongly that disambiguation (or afterthought) cannot be the only function of tails as no further referential information is added. We can also note that Aijmer (1989: 150) describes tails as 'a grammaticalized device for creating an affective bond with the hearer' and considered this 'phatic' function to be far more frequent than the disambiguation function in her data. More importantly, we can note that affective and discourse informational functions are not mutually exclusive: '[Tails] are attentive to the online management of interaction. This is not, however, the same as saying they are after-thoughts . . . their interpersonal and affective aspects remain undiminished by relating them to the pressures on maintaining coherence in unplanned talk' (McCarthy and Carter 1997: 409).

Melchers (1983), Geluykens (1987), Ashby (1988), Fretheim (1995), and McCarthy and Carter (1997) all look at tails in terms of their position and function in discourse. In the case of English, for example, Melchers (1983: 59) notes that 'in left dislocation the topic is not necessarily part of the background of the preceding discourse, which seems to be a characteristic of right dislocation', while Geluykens (1987) argues that tails are used to identify referents which are 'inferable', neither totally new nor totally given in the preceding discourse. McCarthy and Carter (1997: 413), however, apply Burton's (1980) framework of discourse moves and argue that tails are 'a central component in the grammar of reciprocating moves', a move they define as one 'in which there is a general expression of mutuality and convergence by a speaker'. It is here that we come up against one of the limitations of the Bolton Corpus noted above: there simply aren't enough conversations containing tails which are of sufficient length to apply this kind of analysis.

THE WIDER RELEVANCE OF THE ANALYSIS

What clearly emerges from the analysis of tails in the Bolton Corpus is that they were a quite frequent feature of spoken English at that time, though the actual frequency rating in the Bolton Corpus has almost certainly been significantly exaggerated by the type of conversations collected. Tails were also a systematic feature of spoken English in Bolton at that time in the sense that they are relatively easily divided into a limited number of structural categories and that we can attribute two overarching functions to them:

1 Many of them co-occur with evaluative and/or emotionally coloured comment.
2 An information-structuring function can also be plausibly attributed to many of the examples.

A study of a particular feature of spoken language in Bolton between 1937 and 1940 may seem esoteric, but I am going to argue that it has wider implications arising from the analytical approach adopted. In the introduction, we referred to McCarthy and Carter's (1995) contention that grammarians have ignored 'features that occur widely in the conversation of native speakers of English, across speakers of different ages, sexes, dialect groups, and social classes, with a frequency and distribution that simply cannot be dismissed as aberration' (McCarthy and Carter 1995: 142). This study is consistent with and lends support to that view. While current corpora show that tails occur 'across speakers of different ages, sexes, dialect groups, and social classes', the Bolton Corpus shows that tails occur across speakers of different generations. Indeed, Durham (2007) notes examples of tails in Victorian literature, while Lambrecht (2001) shows that tails occur across a wide range of different languages. Such a feature cannot be dismissed as aberrational, nor as primarily a dialect feature (although, as we have seen, 'inverted operator tails' may be more common in Northern British dialects). Indeed, far from being aberrational, 'tails' seem to be a frequent feature of spoken English (and other languages). This kind of study can contribute, then, to attempts to identify the core communicative resources that fluent speakers have at their disposal.

If tails are not aberrational, then we need to account for them, and in its description of tails, this paper reflects a number of principles which have emerged from spoken language research in recent years:

1 The description was based on the study of real spoken data (the Bolton Corpus) and referred to a number of other corpus-based studies. This paper, then, follows the view expressed by McCarthy (1998: 173) that in the study of spoken language 'it is simply impossible to idealise the data away from who said it, to whom, at what point, with what apparent goals and purposes, in the context of what relationship, and under what circumstances'.

2 We noted above that tails allow speakers to cope with planning pressure by offering the possibility of postponed identification or disambiguation. In accounting for tails in terms of the circumstances of production, this paper takes the kind of functional approach to explaining features of spoken language advocated, for example, by Ruehlemann (2007). Biber *et al*. (1999: 43) stress the particular influence of processing constraints on grammatical choice in spoken language: '... The patterns of use associated with a grammatical feature are often strongly influenced by differing production and comprehension circumstances.'

3 In accounting for tails, we noted their interactive and affective function and their tendency to occur in conversations where interlocutors are familiar with one another. McCarthy and Carter (1997: 406) quote Hopper and Thompson (1993) to stress the range of factors which come to bear on grammar: '[grammar] is shaped by the entire range of cognitive, social and interactional factors involved in the use of language.' What is more, they emphasize, no single factor dominates. Similarly, Biber *et al*. (1999: 23) stress the influence of situational factors on grammatical choice: 'Speakers express their own personal attitudes, feelings and concerns, and they interact with one another to build a shared discourse jointly. In conversing, a speaker's use of grammatical features is strongly influenced by situational characteristics of this type.'

4 While the fragmentary nature of the Bolton data made it difficult to situate tails in full discourse context, we noted that Melchers (1983), Geluykens (1987), Ashby (1988), Fretheim (1995), and McCarthy and Carter (1997) all related tails to their position and function in discourse. A consistent thread in the literature on spoken language is that a sentence-based system of analysis is inadequate to provide a full descriptive account of spoken data (Channell 1994; Brazil 1995; McCarthy and Carter 1997; Hughes and McCarthy 1998; McCarthy 1998).

In short, much spoken language research has pointed to the need for a grammar which goes beyond traditional logical and ideational concerns and accounts for the here-and-now linear construction of speech and its interactive and affective dimensions. It is in that more general applied linguistic context that this particular piece of research has been carried out.

I would also argue that the diachronic perspective offered by this paper has wider implications. This paper has shown that there are marked similarities between tails in the Bolton Corpus and tails in descriptions based on more recent corpora. These similarities can be seen in structure and function and, to a less marked extent, in frequency and distribution. The similarities are perhaps most striking between tails in the Bolton Corpus and tails in the York corpus, but similarities with more widely based corpora, such as CANCODE or the *Longman Corpus of Spoken English*, are also evident. At first sight, the apparent consistency and durability of tails over a period of at least 70 years might be seen as surprising for two main reasons. First, as tails are almost exclusive to spoken language, one might expect them to be susceptible to language change. Indeed, Aitchison (2001) argues that language change is often most evident at the 'frayed edges of language' and it is to the 'frayed edges of language' that some linguists have consigned tails: '[tails] are seen as some kind of aberrant variation on a 'normal' structure and, indeed in some accounts (e.g., Melchers 1983) are described simply as non-standard, dialectal deviations which have been relegated to the peripheries of linguistic concern' (McCarthy and Carter 1997: 407). Secondly, if tails are predominantly a feature of Bolton dialect or other dialects, one can ask why they have not died out along with other dialect features (including most of those features of Bolton dialect highlighted earlier in this article), why they are still present in current more widely based corpora, and, indeed, why they are present in many other languages. How, then, do we account for their apparent durability and consistency?

Ruehlemann (2007) proposes the 'adaptedness hypothesis' in which features typical of spoken language are best understood in terms of how they enable speakers to cope with the circumstances and constraints under which conversation takes place. I am going to argue that tails are particularly well adapted to two of the circumstances and constraints of conversation: real-time processing and relation management, to use Ruehlemann's (2007) terms. Real-time processing, as Ruehlemann (2006) argues, allows speakers little time to plan ahead and requires them to edit what they say 'online'. Limitations of working memory further complicate the process of online editing. A feature such as tails which allows for a noun phrase to be placed outside the canonical clause structure seems to be well adapted to the spontaneous and unplanned discourse which is a characteristic of informal conversation: it allows the speaker to clarify, elaborate or reinforce online the subject initially chosen and in this sense tails can be said to have psycholinguistic value. It is interesting in this respect that 18 of the 80 tails in the Bolton Corpus are produced by spectators at sports events where spontaneous and unplanned comments abound. It is reasonable to suppose in the example below, uttered by a spectator at a Bolton Wanderers football match in the late 1930s, that the speaker is first struck by the player's age, then, in his enthusiasm to express the message, reaches for a non-specific noun phrase ('feller') to refer to the player, before realizing the need to clarify who he is talking about ('this right back'):

- This feller must be well in the 33s, this right back

Berg (1998) has argued for the primacy of processing factors among the many factors which can influence language structure and language change. Similarly, Aitchison (2003) has argued that psycholinguistic factors—memory limitations and processing procedures, for example—are the 'top layer of causation' of language change and stand in a hierarchical relationship with linguistic and sociolinguistic factors. If language change is ultimately explicable in terms of 'broad properties of the human mind' (Aitchison 2003), these broad properties of the mind may also be responsible for durability and stability in spoken language. Berg (1998: 284) argues that processing factors do not change much over the years: 'Given the time span over which psycholinguistic predictions could be expected to hold, processing principles can be assumed to be of a relatively constant and permanent nature'. As the conditions of speech production do not change over the years, it seems reasonable to suppose that features which are particularly well adapted to these conditions stand a greater chance of survival than those which are less well adapted. Berg (1998) argues that a linguistic feature may be stable because it is easy to process. In other words, psycholinguistic factors could be at once the motor of change and the anchor of stability.

We have also noted that tails have a clear evaluative function and are often a vehicle for emotionally coloured comments. These too are centrally important functions of spoken language. As Carter (2004: 117) argues: 'When speakers interact, they do more than transmit information . . . Speakers also often wish to give a more affective contour to what they or others are saying'. Carter (2004) goes on to argue (though he is discussing vocabulary at this point) that there are 'three essential expressive options' open to speakers: intimacy, intensity and evaluation. We have seen that tails are a feature of informal conversations and could be said to conventionally index such conversations. Their evaluative function is clear and we have also seen that they can express intensity in their role as vehicles for emotionally coloured comment. In Ruehlemann's (2007) terms, tails are well adapted to the factor of 'relational goal-orientation' which 'unfolds chiefly along **two dimensions**: as participant-relation, that is, the speaker's relation to other participants, and as proposition-relation, that is, the speaker's relation to what s/he is saying'. As conventionally recognized signals of informality and vehicles of emotionally coloured comment, tails are adapted to *both* these dimensions. In assessing the importance to conversation of the socioaffective functions which can be attributed to tails (and to other features of spoken language), we need to consider briefly the goals of conversation itself. It is interesting in this respect that Ruehlemann (2007) argues that 'the overriding goal in conversation is primarily relational rather than transactional', while Biber *et al.* (1999: 1041) argue that 'its primary function appears to be to establish and maintain social cohesion through the sharing of experience . . .'.

We referred above to the surprising consistency and durability of tails over a long period of years. I would argue that the capacity of tails to play multiple and central functions in relation to the constraints and goals of conversation is a significant factor in this consistency and durability. In this sense, it may be legitimate to speak of a 'linguistic survival of the fittest'. McMahon (1994: 340) underlines that evolutionary metaphors need to be applied to historical linguistics with great caution: '. . . interpretations of the term evolution as meaning progressive achievement or goal-directed activity are badly motivated and should not be borrowed into linguistics.

However, the Darwinian theory of biological evolution with its interplay of mutation, variation and natural selection, has clear parallels in historical linguistics, and may be used to provide enlightening accounts of linguistic change'. In this case, the metaphor is being applied to advance a hypothesis—it can be no more—to account for a particular case of language stability.

CONCLUSION

In this article, I have argued that tails are a surprisingly consistent and durable feature of spoken English and hypothesized that their durability can be attributed to their ability to meet important psycholinguistic and socio-affective needs. The longevity of tails lends weight to the argument that they have a perfectly proper place in descriptive grammars of spoken English and that they deserve consideration—no more, no less—in any assessment of the structures required to be an effective communicator in spoken English. In this sense, we can argue that there is a *prima facie* case for including tails in the English language teaching syllabus, but it can be no more than a *prima facie* case when there are so many other sociocultural and pedagogic factors to take into account.

The diachronic aspect of this research has, however, wider implications. There may be more data lurking in the Worktown papers to add to the Bolton Corpus, but it is a finite source and we have already acknowledged its limitations. As time goes on, more detailed and comprehensive diachronic studies of recently described spoken language features will be possible using corpora which are demographically and generically aligned. Such studies could give us important insights into the properties of the human mind and of human conversation which determine the nature of spoken language.

Documentary sources

Papers from the Mass Observation Archive, Part 3: The Worktown Collection 1937–40; Box 2: The Pub and the People; Box 3: Public Houses; Box 4: Sport; Box 42: Assorted short reports (1); Box 43: Assorted short reports (2); Box 48: Leisure activities, fairs and dance halls; Boxes 50A–50B: Churchill and Chamberlain: War Talk; crisis 1939–1940; Box 51: Reactions to news from Belgium and France; Box 52: Observations in Bolton in the early months of the war, 1939–1940; Box 56: Observations in Blackpool; Box 57: Observations in Blackpool cont'd; Box 58: Side shows and amusements.

NOTES

[1] Available on microfilm at the libraries of the University of Sussex and the University of Leeds and at Bolton Central Library.

[2] This example is also consistent with Carter, Hughes, and McCarthy's (1998) observations on negative concord in tails.

REFERENCES

Aijmer, K. 1989. 'Themes and tails: the discourse functions of dislocated elements,' *Nordic Journal of Linguistics* 12: 137–53.

Aitchison, J. 2001. *Language Change: Progress or Decay?* Cambridge University Press.

Aitchison, J. 2003. 'Psycholinguistic perspectives on language change' in D. Joseph and R. Janda (eds): *The Handbook of Historical Linguistics*. Blackwell, pp. 736–43.

Ashby, W. 1988. 'The syntax, pragmatics, and sociolinguistics of left- and right-dislocation in French,' *Lingua* 75: 203–29.

Berg, T. 1998. *Linguistic Structure and Change: An Explanation from Language Processing*. Clarendon Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and **E. Finegan.** 1999. *The Longman Grammar of Spoken and Written English*. Longman.

Brazil, D. 1995. *A Grammar of Speech*. Oxford University Press.

Burton, D. 1980. *Dialogue and Discourse*. Routledge.

Carter, R. 2004. *Language and Creativity: The Art of Common Talk*. Routledge

Carter, R. and **M. McCarthy.** 1995. 'Grammar and the spoken language,' *Applied Linguistics* 16/2: 141–58.

Carter, R., R. Hughes, and **M. McCarthy.** 1998. 'Telling tails: grammar, the spoken language and materials development' in B. Tomlinson (ed.): *Materials Development in Language Teaching*. Cambridge University Press, pp. 67–86.

Carter, R. and **M. McCarthy.** 2006. *The Cambridge Grammar of English*. Cambridge University Press.

Channell, J. 1994. *Vague Language*. Oxford University Press.

Cullen, R. and **M. Kuo.** 2007. 'Spoken grammar and ELT course materials: a missing link?' *TESOL Quarterly* 41/2: 361–86.

Dik, S. 1978. *Functional Grammar*. Holland.

Dik, S. 1981. *Functional Grammar*. Foris Publications.

Downing, A. and **P. Locke.** 1992. *A University Course in English Grammar*. Prentice Hall.

Durham, M. 2007. ' ''It's altered a lot has York'': Right dislocation in Northern England.' *York Papers in Linguistics* Issue 8, pp. 61–72.

Fretheim, T. 1995. 'Why Norwegian right-dislocated phrases are not afterthoughts,' *Nordic Journal of Linguistics* 18/1: 31–54.

Geluykens, R. 1987. 'Tails (right dislocation) as a repair mechanism in English conversation' in J. Nuyts and G. de Schutter (eds): *Getting One's Words into Line: On Word Order and Functional Grammar*. Foris Publications, pp. 119–29.

Harrisson, T. 1974. *The Pub and the People*. Mass Observation: Hutchinson.

Hedevind, B. 1967. *The Dialect of Dentdale*. Uppsala.

Hopper, P. and **S. Thompson.** 1993. 'Language universals, discourse pragmatics and semantics,' *Language Sciences* 15/4: 357–76.

Hughes, R. and **M. McCarthy.** 1998. 'From sentence to grammar: discourse grammar and English language teaching,' *TESOL Quarterly* 32/2: 263–87.

Jeffrey, T. 1999. 'Mass-Observation: A short history.' University of Sussex, *MOA Occasional Paper no. 10.*

Kytö, M. and **T. Walker.** 2003. 'The linguistic study of early modern English speech-related text: how 'bad' can 'bad' data be?' *Journal of English Linguistics* 31/3: 221–48.

Labov, W. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press.

Lambrecht, K. 1987. 'On the status of SVO sentences in French discourse' in R. Tomlin (ed.): *Coherence and Grounding in Discourse*. Benjamins, pp. 217–61.

Lambrecht, K. 2001. 'Dislocation' in M. Haspelmath, E. König, W. Oesterreicher and W. Raible (eds): *Language Typology and Language Universals: An International Handbook*. Vol. 2. Walter de Gruyter, pp. 1050–78.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge University Press.

McCarthy, M. and **R. Carter.** 1995. 'Spoken grammar: what is it and how should we teach it?' *English Language Teaching Journal* 49/3: 207–17.

McCarthy, M. and **R. Carter.** 1997. 'Grammar, tails and affect: constructing expressive choices in discourse,' *Text* 17/3: 405–29.

McMahon, A. 1994. *Language Change*. Cambridge University Press.

Melchers, G. 1983. 'It's a sweet thing is tea-cake. A study of tag statements' in S. Jacobson (ed.): *Papers from the Second Scandinavian Symposium on Syntactic Variation*. Acta Universitatis Stockholmiensis, pp. 57–66.

Petyt, K. 1985. *Dialect and Accent in Industrial West Yorkshire*. John Benjamins Publishing.

Phillips, T. 2000. *A Postcard History*. Thames and Hudson.

Quirk, R., S. Greenbaum, G. Leech, and **J. Svartvik.** 1972. *A Grammar of Contemporary English*. Longman.

Quirk, R., S. Greenbaum, G. Leech, and **J. Svartvik.** 1985. *A Comprehensive Grammar of the English Language*. Longman.

Rissanen, M. 2000. 'The world of English historical corpora: from Caedmon to the computer age,' *Journal of English Linguistics* 28/1: 7–30.

Ruehlemann, C. 2006. 'Coming to terms with spoken grammar,' *International Journal of Corpus Linguistics* 11/4: 385–409.

Ruehlemann, C. 2007. *Conversation in Context: A Corpus-Driven Approach*. Continuum.

Shorrocks, G. 1985. 'The syntax of the dependent pronoun in the dialect of Farnworth and district (Greater Manchester County, formerly Lancashire).' *Transactions of the Yorkshire Dialect Society*. 15(1980-85/Part LXXXIV): 40–9.

Shorrocks, G. 1999. *A Grammar of the Dialect of the Bolton Area, Part 2: Morphology and Syntax*. Peter Lang.

Strässler, J. 1982. *Idioms in English: A Pragmatic Analysis*. Gunther Narr Verlag.

Tagliamonte, S. 1996–1998**.** *Roots of Identity: Variation and Grammaticization in Contemporary British English***.** Economic and Social Sciences Research Council (ESRC) of Great Britain. Reference #R000221842**.**

Wright, J. 1905. *The English Dialect Grammar*. Clarendon Press.

Applied Linguistics: 31/3: 346–367 © Oxford University Press 2009
doi:10.1093/applin/amp037 Advance Access published on 21 September 2009

Speaking Correctly: Error Correction as a Language Socialization Practice in a Ukrainian Classroom

DEBRA A. FRIEDMAN

Michigan State University

This study uses a language socialization approach to explore the role of Ukrainian language instruction in the revitalization of Ukrainian as the national language. Based on 10 months ethnographic observation and videotaping of classroom interaction in two fifth-grade Ukrainian language and literature classrooms, it focuses on corrective feedback targeting children's use of Russian forms and considers how these practices are shaped by the imperatives of Ukrainian language revitalization and language ideologies that valorize 'pure language' as the sole legitimate variety of Ukrainian. The analysis reveals how corrective feedback is socializing children into speaking pure language and into dominant Ukrainian language ideologies that proscribe language mixing as a violation of the natural boundaries between languages, thus preserving a distinct Ukrainian language as an emblem of a distinct Ukrainian nation.

> Mastery of the native language is not an entitlement but the duty of a patriot.
>
> 'Language duties of a citizen' (posted in a Ukrainian classroom)

Nearly 20 years after independence, Ukraine continues to debate how to define itself in relation to the former dominant power, Russia. Ukrainian uneasiness about Russian influence is often expressed through concern about the integrity and even survival of the Ukrainian language. The language has long been an identity marker as a symbol of internal unity and external differentiation. But today, although Ukrainian is the state language, it is far from hegemonic in its titular nation. Not only is there a substantial ethnic Russian population,[1] but many ethnic Ukrainians speak Russian as their primary language (Arel 1996, 2002; Bilaniuk 2005; Pavlenko 2006; Bilaniuk and Melnyk 2008),[2] and Russian dominates in popular culture. In addition, many Ukrainian speakers do not speak Standard Ukrainian, but a Ukrainian/Russian hybrid called *suržyk*[3] that is widely viewed as a residue of the 'Russification' of Ukrainian life and culture (Flier 2000; Bilaniuk 2005).

As part of its nation-building project, the state is representing Ukraine as a nation of Ukrainian speakers (Arel 2002) and has embarked on a program to cleanse the language of perceived Russian influences and encourage more widespread use of literary (i.e., Standard) Ukrainian. A primary site for this effort is the nation's schools, all of which teach Ukrainian as an obligatory subject. Yet while Ukrainian language politics and language attitudes have received increasing scholarly attention in recent years (e.g. Arel 1996, 2002; Bilaniuk 2005; Bilaniuk and Melnyk 2008), I know of no research that has examined how Ukrainian language policies are being implemented at the classroom level or how these attitudes are being transmitted to the first post-Soviet generation of Ukrainians.

This article takes a language socialization approach to explore the role of Ukrainian language instruction in the revitalization of Ukrainian as the national language. Based on 10 months ethnographic observation and videotaping in two fifth-grade classrooms, it focuses on a recurrent feature of classroom interaction: corrective feedback targeting children's use of Russian forms. Taking the position that standards of linguistic correctness are socially constructed (Bourdieu 1980/1991; Silverstein 1996), I analyze these feedback practices as a manifestation of an ideology of 'pure language' (*čysta mova*) that originated in response to the historical position of Ukrainian as subordinate to Russian and the perceived need to establish it as a distinct language suitable for representing a distinct nation. I further argue that in addition to socializing children into the ways of speaking deemed to constitute Ukrainian language competence, these practices are socializing them into pure language ideologies that define what this competence consists of.

SOCIALIZATION INTO A LINGUISTIC COMMUNITY

This research is situated at the intersection between two fields of inquiry, language socialization and language ideology. It examines how Ukrainian schoolchildren are being socialized into a *linguistic community* (Bourdieu 1980/1991; Silverstein 1996, 1998, 2000), defined as 'groups of people by degree evidencing allegiance to norms of denotational . . . language usage' (Silverstein 1998: 402). What unites a linguistic community is not a set of language practices, but a set of language ideologies that define what counts as legitimate language. In the modern nation-state, this language is the national language(s) that has been standardized and legitimated through institutionalization in government, media, and education. This language subsequently becomes 'the theoretical norm against which all linguistic practices are objectively measured' (Bourdieu 1980/1991: 45), and language usage that deviates from standard norms is viewed as incorrect.

By drawing attention to verbal behaviors deemed to be problematic and responding to them in particular ways, corrective feedback routines constitute a central locus for socializing novices into a linguistic community. Recognition of the socializing function of corrective feedback has long had a place in language socialization research. In their pioneering studies of child language socialization in Samoa and among the Kaluli of Papua New Guinea, Ochs and Schieffelin (1984, 1995; Ochs 1984, 1988; Schieffelin 1990) revealed how feedback practices reflect underlying cultural beliefs and values and demonstrated how participation in corrective feedback routines socialized children into social roles and relationships and into local understandings regarding what constitutes appropriate language behavior. The role of corrective feedback in language socialization has also been noted in several studies conducted across a range of classroom contexts. In an early classroom application of the paradigm, Poole's (1992) analysis of interaction in two English as a Second Language classrooms identified a preference for feedback strategies identified by Ochs (1984, 1988) and Ochs and Schieffelin (1984, 1995) as typical of white middle class American caregivers. Based on these observations, Poole concluded that teaching practices are in large part culturally motivated and that language classroom interaction conveys implicit cultural messages. Duff's research in dual language immersion high schools in Hungary (1995, 1996) found that during student lectures in some English-medium classrooms other students self-selected to request clarification or correct presenters' language errors and even corrected teachers' language errors, practices unheard of in traditional, teacher-directed Hungarian classrooms (Duff 1995). Duff's analysis revealed how these practices were socializing both students and teachers into new ways of relating to knowledge and authority that mirrored democratization and educational reform that were then ongoing in post-Communist Hungary. In a study of the linguistic expression of respect in a village school in Thailand, Howard (2004) analyzed teachers' selective correction of children's inappropriate use of honorific particles as a strategy through which children come to associate honorific registers of Standard Thai with formal aspects of classroom discourse. She further noted how insistence on usage of these particles only during certain classroom activities served to structure and socialize children's attention and participation in the classroom. Finally, Jacobs-Huey (2007) examined negative feedback provided in response to use of terminology deemed to be unprofessional in an African-American cosmetology school. She analyzed these responses as evidence of the importance of language in the construction of expert identities within the community of professional African-American hair stylists and the socialization of novices into proficient use of professional language as a means of claiming expertise.

These studies have suggested that corrective feedback practices are not motivated solely by teachers' personal philosophies or notions of pedagogical efficacy, but embedded within larger social, political, and cultural systems of belief about norms of language use and expectations regarding the responsibility of novices in upholding these norms. In addition to its role in regulating language use, classroom corrective feedback contributes to a range of goals that reflect the values of the communities in which the classrooms are situated.

CONTEXT OF THE STUDY

Language and nation in Ukraine

The perception of a struggle between Ukrainian and Russian for linguistic and cultural dominance has long reverberated in Ukraine. Ukrainian and Russian evolved from the same parent language (East Slavic) and are syntactically similar, but differ phonologically. They also share a stock of words from East Slavic as well as Russian and international words that entered Ukrainian via Russian (Shevelov 1993; Bilaniuk 2005). Ukrainian also reflects influence from Polish, resulting from a long period of Polish rule over the territory that began to ebb in the mid-17th century. By the late 18th century the decline of the Polish state had led to the partition of Poland among Prussia, Russia, and Austria, and Ukrainian territories were divided between the Austro-Hungarian Empire, which controlled the western regions of Galicia, Bukovina, and Transcarpathia, and the Russian Empire, which controlled the rest. This division was to last, under various governments, until the end of the Second World War, when post-war agreements ceded control over western Ukrainian regions to the Soviet Union (Magocsi 1996).

This history had a profound effect on the development of the Ukrainian language. First, outside political domination meant that for centuries Ukrainian had few, if any, public functions, and by the early 20th century language shift to Russian was well under way among the upper classes and urban residents in Russian Ukraine (Shevelov 1989). In addition, the division of the territory into multiple political units complicated the process of creating a standardized language that would be accepted across all ethnic-Ukrainian territory. Finally, the process of language standardization became enmeshed with issues of national identity and political sovereignty.

In the mid-19th century the Ukrainian language became the focus of a nascent nationalist movement constructed around a common European ideology that viewed possession of a unique common language as an essential element of nationhood (e.g. Blommaert and Verschueren 1998; Irvine and Gal 2000). This belief inspired efforts to purify and standardize Ukrainian in order to establish it as a legitimate language distinct from Russian or Polish by eliminating forms deemed to be 'foreign' in favor of those grounded in the supposedly unique and authentic norms of village dialects (Wexler 1974). These activities greatly alarmed Russian imperial authorities, who regarded them as a threat to the inherent unity of the Ukrainian and Russian languages and thus the Ukrainian and Russian peoples. From the imperial Russian perspective, Ukrainian was a dialect of Russian, *malorossijskoe narečije* 'the Little Russian dialect,' and attempts to claim otherwise were viewed as separatism. From 1876 to 1905 public use of Ukrainian (in newspapers, theaters, etc.) was banned in Russian Ukraine (Wexler 1974; Magocsi 1996).

Following the Bolshevik Revolution in 1917, Ukrainian became an official language of the Ukrainian Soviet Socialist Republic and was for a time actively promoted under a policy of *ukrajinizacija* 'Ukrainization' (Wexler 1974; Shevelov 1989; Martin 2001). However, by the late 1920s seemingly apolitical activities such as reforming the lexicon to eliminate Russian-influenced forms had come to be seen as an expression of hostility towards Russia and a correspondingly positive orientation towards Poland, a crime known as *treasonous irredentism*. In a series of show trials beginning in 1929, more than one Ukrainian linguist disappeared into the gulag for *language sabotage*, that is, producing dictionaries or grammars accentuating differences between Ukrainian and Russian (Martin 2001). Meanwhile, although Ukrainian-medium schools and a Ukrainian language press continued to operate, perceptions of Ukrainian as a village language and the prestige of Russian furthered ongoing language shift (Shevelov 1989; Martin 2001). There was also a tendency towards convergence between the languages, both as a consequence of increased language contact as well as Soviet language policies. For example, when Ukrainian had two possible morphological or syntactic forms, reforms in the 1930s established those resembling Russian forms as the only permitted variants in Standard Ukrainian (Wexler 1974; Shevelov 1989).

The Ukrainian linguistic community

With independence in 1991 and its subsequent installation as the sole state language, Ukrainian has expanded into arenas previously dominated by Russian, such as higher education, television broadcasting, and government administration. Yet despite Ukrainian's increased status, many commentators have expressed concern about the lingering effects of Russification. The years since independence have seen a revival of tendencies towards linguistic differentiation and purification; many of the reforms of the 1930s have been reversed, and the legitimacy of forms thought to have originated in Russian has again come under question (Taylor 1998; Bilaniuk 2005).

But concerns about the purity of Ukrainian can also be found among the general population, where they are manifested in widespread negative attitudes towards the hybrid language known as *suržyk*. While linguists reserve the term for a 'hybrid in which the *entire* grammar of Ukrainian . . . contains Russian-influenced elements or distribution not otherwise represented in an identical function in Contemporary Standard Ukrainian' (Flier 2000: 114), Ukrainians may identify as *suržyk* any infiltration of Russian into Ukrainian speech (Arel 1996; Bilaniuk 2005). Critics of *suržyk* characterize it as an unnatural product of centuries of linguistic oppression and cite its existence as a threat to Ukrainian national consciousness. In response, a small industry has sprung up dedicated to its eradication, including style manuals, newspaper columns, and a program on Ukrainian state radio. The valorization of pure language, once the province of an intellectual elite intent on establishing Ukrainian claims to nationhood, has become 'naturalized' (Bourdieu 1977) in Ukraine as a dominant ideology, affecting how language is used and evaluated at the level of everyday language practices (Bilaniuk 2005).

THE STUDY

Data collection

Data were collected during the 2003–2004 academic year in fifth-grade Ukrainian language and literature classrooms at two schools in a small city in south-central Ukraine. In the late 1990s I had taught English at the local pedagogical university, and contacts there put me in touch with two schools regarded as having good Ukrainian language programs. One, a general education school, had been using Ukrainian as the medium of instruction since the late 1940s, making it one of the first in the city to do so. The other, a gymnasium[4] specializing in physics and mathematics, was originally a Russian-medium school, but switched to Ukrainian following independence. In October I began observing Ukrainian classes at both schools at least once per week and taking field notes. Upon obtaining written informed consent from teachers, children, and parents, I began videotaping classes using a digital video camera and shotgun microphone mounted on a tripod and stationed at the back of the classroom. In total, data comprise field notes from 88 lessons (66 h) and video recordings of classroom interaction from 42 lessons (31.5 h).

At the end of the school year I interviewed the Ukrainian teachers and principals at both schools. These interviews, lasting approximately 1 h, were conducted in Ukrainian and were audio-recorded. Interviews with teachers included questions about their assessments of students' proficiency in Ukrainian, their teaching philosophy, and what they saw as the primary goals of Ukrainian language instruction, as well as questions about specific activities that I had observed in class. Although teachers' busy schedules permitted only one formal interview, teachers sometimes chatted with me informally during breaks and shared their thoughts about the lesson. To get a better understanding of the curriculum and school cultures, I also collected textbooks and other materials and attended several school events.

Parents completed questionnaires regarding their occupations, native language(s), language of education, and language(s) used within the home. In addition, 20 parents consented to an interview. With one exception, I conducted these interviews in Ukrainian or Russian, depending on the interviewee's preference. The exception was a parent who was also a graduate student in English; this interview was in English. Interviews included questions regarding language use in the home and at work, feelings about having their child educated in Ukrainian, and what they wanted their child to learn about the Ukrainian language and culture. They were audio-recorded and lasted approximately 30–45 min.

Finally, my status as a native speaker of English made me a valuable commodity, and I was invited to speak with students in advanced-level English classes at the pedagogical university and both focal schools. One conversation with a group of undergraduates was audio-recorded with their consent;

in other cases I made subsequent notes on language-related issues that arose during these discussions.

Data analysis

Two Ukrainian assistants completed rough transcriptions of the recorded classroom data. Based on these transcriptions and a review of videotapes, I identified instances of corrective feedback, using the definition of *correction* delineated in Schegloff, Jefferson, and Sacks: 'The replacement of an ''error'' or ''mistake'' by what is ''correct''' (1977: 362) and prepared detailed transcripts of these segments (see the Appendix at *Applied Linguistics* online for transcription conventions). Analysis of the data is based on the original languages; however, for presentation purposes I have translated transcripts into English. When unsure about a passage I have checked with a native speaker consultant, a Ukrainian instructor at an American university.

Analysis of classroom data incorporates both microanalysis of corrective feedback sequences as well as macro-level analysis. The microanalysis considers (i) the nature of the error or *trouble source*, (ii) who initiates and who completes the correction, and (iii) the outcome of the correction (i.e. whether there is uptake). The macro-level analysis draws upon field notes, classroom texts, and interviews, as well as observations of language use in the local community and informal conversations with friends, neighbors, and other local residents in order to situate these practices within a larger context.

While I have made every attempt to incorporate an emic perspective, I also acknowledge the effects of my own position as a researcher, applied linguist, foreigner, and competent but non-native speaker of Ukrainian and Russian. I had lived and worked as an English teacher in Ukraine for three years prior to beginning this research (including one year in the city where the research was conducted); nevertheless, I was still an outsider in this community. Participants referred to me as 'our American guest,' and curiosity about me and my interest in Ukrainian (which many found puzzling) motivated many parents to consent to an interview and undoubtedly shaped how they responded to my questions. Finally, while my language skills were sufficient to allow me to analyze classroom interaction and conduct interviews in both Ukrainian and Russian, as a non-native speaker I have relied on multilingual research assistants and friends to help me understand nuances in the data; thus some of my interpretations have been filtered through theirs.

Setting

Although the majority of the city's residents are ethnic Ukrainians, industrialization, the presence of an air force base and several higher education institutes, and the city's reputation for a salubrious climate drew people from throughout the Soviet Union. Both Ukrainian and Russian are heard on its streets and are used interchangeably at public events, reflecting assumptions that everyone at least understands both languages. I was also informed that local etiquette requires answering a person in the language in which he or she addresses you, and the ability to switch easily between Ukrainian and Russian is a requirement for service jobs such as salesclerks.

At the time of this study, 33 out of 35 schools in the city used Ukrainian as the medium of instruction, a reversal of the situation before independence, when only two used Ukrainian. Although Ukrainian-medium schooling has met with resistance in some regions, it had been accepted among the parents I spoke to, who agreed that children would need Ukrainian proficiency in order to attend university or find a job. In addition, many parents, including some who identified themselves as Russian, stated categorically that children should know the national language of their country.[5] Most indicated that they also wanted their children to be proficient in Russian.

But while Ukrainian-Russian bilingualism was valued, I heard many complaints about language mixing. This problem was not seen as limited to Russian-speakers who had learned Ukrainian as a second language, but as afflicting native Ukrainian speakers as well. For example, in an audio-recorded conversation with pedagogical university undergraduates, a young woman commented:

> I think that here in Ukraine we have the problem, uh, a big problem of purity of speech. Because the majority of our people, even rather well educated ... mix Ukrainian words and Russian words in their speech. And they can speak neither pure Ukrainian nor pure Russian. And we often speak *suržyk* as we call it. [English in original]

When I asked why mixing was a problem, another student answered, 'Because the speech is not correct, it is not pure,' while a third added, 'It isn't so beautiful.'

The perception that *suržyk* is spoken by 'the majority of our people, even rather well educated' was echoed by some of my interviewees. This situation was not seen as an individual problem, but a social one, attributed to past Russian dominance. For example when I asked a child's mother about her native language, she identified it as Ukrainian, but added

> But I lived at the time when there was the Soviet Union, and our native Ukrainian language was rather polluted .... where we live in our territory ... here very many Russianisms have come about. And that language, which has been polluted by Russianisms has been preserved up to the present time, unfortunately. [Ukrainian in original]

Similarly, a school principal, after bemoaning children's tendency to speak what she characterized as *suržyk*, commented, 'I would love it if the children spoke pure Ukrainian. But that will take years. Because they have implanted the Russian language in us.' [Ukrainian in original]

Other native Ukrainian speakers negatively evaluated their own speech when measured against the standards of pure language. For instance, a parent described how he came to realize that his Ukrainian was actually *suržyk*:

> I started learning English and one day I woke up and thought why am I using Russian words, you know, *suržyk*, ah, everyone uses it, I still use some Russian words, but I woke up and I thought why do I speak so badly in Ukrainian. I'm learning a foreign language and I don't know my own language. [English in original]

Such comments illustrate the complex relationship between language ideologies and language practices. Many in this community occasionally mixed languages or used what they themselves characterized as 'Russianisms.' However, these same speakers labeled such practices as *nečysta* 'impure' and incorrect. While speaking pure language was viewed as an exception, it was nevertheless held up as the ideal to which everyone should aspire.

Participants

Although the focal schools differed in many ways, the corrective feedback observed in the two classrooms was strikingly similar. This analysis will focus on the class at one school, the gymnasium. One of the largest schools in the city, it had 1,066 students and 64 teachers and a reputation for academic excellence. Following the state-mandated curriculum, fifth graders had four 45-min Ukrainian language lessons and two Ukrainian literature lessons per week. They also studied Russian for 2 h per week.[6]

At the time of this study the Ukrainian teacher, Viktor Viktorovych[7] (hereinafter VV) was completing his 26th year as a teacher. He had been teaching at the school since 1985 and had taught the parents of several children in the class. VV was regarded as an excellent teacher of Ukrainian, and his classes were often observed by pedagogical university students.

A total of 24 children from this school participated in the study.[8] Their parents were educated professionals, such as engineers, economists, or computer programmers. Slightly over half the children (13/24) were of Ukrainian ethnicity; that is, both parents identified themselves as Ukrainian. One of the children was Russian, and one was Armenian. The remaining children were of 'mixed' ethnicity; that is, one parent self-identified as Ukrainian and the other as Russian (eight) or Polish (one). However, ethnic affiliation did not necessarily correlate with home language use. Only five children came from homes where exclusively Ukrainian was used, nine came from homes where exclusively Russian was used, and nine used both. The Armenian child spoke Armenian at home and, according to his mother, spoke Russian with neighbors and playmates. That is, nearly 80 per cent routinely used Russian outside of school. The influence of Russian could also be seen in children's language practices during breaks, when Russian was commonly used.

All parents claimed to know both Ukrainian and Russian and to have Ukrainian-language print material in their homes. I observed no connection between a child's home language and standing in the class; several students who received top grades came from Russian-speaking or bilingual homes. None of the Russian-speaking parents whom I interviewed felt that their children were disadvantaged by the difference between their home and school languages, noting that the children had been studying in a Ukrainian-medium school since the first grade and were therefore (in their view) fully competent in Ukrainian. VV agreed that Ukrainian-medium schooling had given the current group of fifth graders better command of the language than what he had observed in earlier generations of students. However, he also expressed concern about the dominance of Russian in many children's lives.

Language in the classroom

The classroom layout, with three parallel rows of student desks facing front, lent itself to the preferred lesson format, teacher-directed whole-class discussion. Children were expected to be active participants, and at the end of each lesson VV assigned grades based on the quantity and quality of each child's contributions. Classroom language use reflected assumptions that the children were Ukrainian–Russian bilinguals. With a few exceptions, the public discourse of the classroom was in Ukrainian, and the language curriculum emphasized metalinguistic analysis, stylistics, and spelling rather than instruction in grammar or pronunciation such as what one might find in a second language classroom. On the other hand, VV also drew upon children's knowledge of Russian. For example, he sometimes asked children to provide Ukrainian equivalents for Russian words, explaining that such exercises would help them distinguish between languages. He also occasionally quoted Russian poetry and invited children to join in his recitation.

Regardless of the children's backgrounds, Ukrainian was considered to be their *ridna mova* 'native language' by virtue of their status as Ukrainian citizens. As the language textbook declared, 'the Ukrainian language is the national language of the Ukrainian nation.... Therefore the Ukrainian language is the native one for each Ukrainian' (Peredrij *et al.* 2002: 4). In class VV routinely referred to *naša ridna mova* 'our native language' or *naša ukrajins'ka mova* 'our Ukrainian language.' He also spent time on activities designed to generate pride in the achievements of Ukraine and the beauty of Ukrainian, explaining to me that when children feel patriotic, they will want to study and use their national language. Children were surrounded by reminders of their obligation to learn Ukrainian, such as the document 'Language duties of a citizen' (quoted at the beginning of this article) that was posted on the wall and a poster in the hallway that directed them to 'love your nation, your land; study its customs, traditions; seek to learn and perfectly master the native Ukrainian language.'

Mastery of the language meant speaking it correctly. Speaking correctly was always relevant; *'slidkujte za sovjeju movoju,'* literally, 'look after your language,' was a frequently heard admonition, and VV once reminded the class, 'Although this is literature [class] we do not forget about the fact that we express our opinion in the literary Ukrainian language.' Corrective feedback targeting children's language use was pervasive, and in ten months of observation I noted only a handful of instances in which a hearable error was not corrected.

While any language error could trigger correction, most correction targeted Russian or Russian-influenced words. These targets can be broadly divided into three categories:

1 Russian words not in the Standard Ukrainian lexicon.
2 Words in the lexicons of both languages but pronounced following Russian phonological norms.
3 *Russianisms*, that is, words that follow Ukrainian phonology (and which may be used by some Ukrainian speakers) but are seen as (i) originating in Russian or (ii) evincing Russian patterns for word formation.

In other words, while use of Russian was acceptable in certain limited contexts, boundaries between languages were to be maintained.

'DOING CORRECTING'

Consistent with the teacher-centered orientation of classroom interaction, the most common type of correction was teacher-initiated teacher-correction, in which VV both indicated a trouble source and provided a replacement, usually in the same turn. Correction typically occurred immediately following the trouble source, often interrupting the turn in progress, a strategy that regularly resulted in uptake of the correction in the child's next turn. This format resembles what Jefferson (1987) calls *exposed correction*, in which the ongoing talk is briefly interrupted as the parties engage in the business of 'doing correcting.' But as Jefferson notes, this shift is collaboratively achieved; that is, it requires that both parties display an orientation to the fact that correcting is now being done (1987: 99).

The excerpt below illustrates the collaborative nature of corrective feedback routines. During a literature lesson the class was discussing a story about a boy's pet pigeon. As punishment for the boy's skipping school to play with his pet, his father takes the bird with him when he goes on a trip to another part of the country. As the excerpt begins, VV calls on Slava to summarize this portion of the story.

Excerpt 1
Business Trip (04/01/04)
Trouble source is in ***boldfaced italics***

1		VV	*Jak pokarav bat'ko (.) Stepanka?* **How did his father punish (.) Stepanko?**
2		Slava	*U:h (vin uže) buv* **U:h (he already) was**
3	TS→	Slava	*pojixav v* ***komandyrovku*** *i [vzjav-* **he went on a *business trip* and [he took-**
4	C→	VV	*[vidrjadžennja.* **[business trip.**
5	U→	Slava	*vidrjadžennja i vzjav z soboju holuba* **business trip and he took the pigeon with him**

As Slava explains how the story's protagonist (Stepanko) was punished, he uses *komandyrovka* to refer to the father's business trip (line 3). In line 4, although Slava's turn has not reached a point of possible completion, VV intervenes to initiate and complete a correction, replacing *komandyrovka* with *vidrjadžennja*. Slava could subsequently continue from the point at which he was interrupted, but he instead redoes the problematic portion of his turn to incorporate *vidrjadžennja* (line 5). He thus displays understanding of VV's prior turn as a correction and implicitly aligns with VV's stance that *komandyrovka* is problematic.

The problem with *komandyrovka* does not lie in its referential meaning; *komandyrovka* and *vidrjadžennja* refer to the same entity ('business trip'), and the Soviet-era *Dictionary of the Ukrainian Language* lists them as synonyms (Academy 1970–1980, vol. IV: 240). However, *komandyrovka* (Russian *komandirovka*), which is based on a loanword from German (*kommandieren* 'to order'), has two features marking it as a word that came into Ukrainian via Russian: (i) retention of the German infix *-ir-*, characteristic of Russian borrowings from German but rejected by Ukrainian purists (Wexler 1974: 65, 163) and (ii) the nominal suffix *-ka* signifying the result of an action, considered by some to be a Russianism. *Vidrjadžennja*, however, is a nominalization of the Ukrainian verb *vidrjadžaty* 'to dispatch, send forward' using the nominal suffix *-nnja*, a Ukrainian alternative to *-ka* (Wexler 1974: 176). That is, the form of *komandyrovka* links it with Russian and a now-discredited Soviet language policy decreeing that foreign borrowings were to take the same shape in Ukrainian as they did in Russian (Wexler 1974: 189). While not all speakers would consider *komandyrovka* to be incorrect, these participants orient to it as an error, interrupting the ongoing activity to replace it with another word. This correction sequence conveys an implicit message that 'Ukrainian' forms such as *vidrjadžennja* are preferred over 'Russianized' ones. This message was understood by at least one other child, who later chose the word *vidrjadžennja* when referring to the father's business trip.

Corrective feedback such as this occurred on average a half dozen times in every lesson. Children were well socialized into their role in these routines; I noted few instances in which children failed to take up a correction, either by redoing their turn or repeating the replacement word. This role positioned children as novices who had not yet mastered the ability to monitor their linguistic output, but who were nevertheless expected to recognize and replace an incorrect form once it had been called to their attention. Through participation in these routines, teacher and students collaboratively constructed and displayed understandings that the norms of Ukrainian language usage included avoidance of words that were Russian or Russianisms.

APPROPRIATING PRACTICES OF CORRECTION

Children's readiness to take up replacement forms does not, in itself, indicate that they understood the nature of their errors, and placement of teacher-correction immediately following the trouble source left little opportunity for children to display this understanding by initiating a correction. However, children occasionally demonstrated an ability to recognize a potential trouble source in their own or other's speech. The following excerpts contain two instances of child-initiated correction as evidence of children's sensitivity to the presence of Russian words as an error

During a literature lesson VV read aloud two poems, 'Winter' and 'I Love Spring,' and asked which season the author described best. As the first excerpt begins, Petja is explaining why he preferred the poem 'Winter.'

Excerpt 2a
Rhyme (02/26/04)
Trouble source is in ***boldfaced italics***

1	TS→	Petja	*Tut i:: uh nu u:h* ***rifma***	*[uh* ***xorošaja***
			Here a::nd uh well u:h the ***rhyme***	**[uh is** ***good***
2		VV		[((Looks to his right))
3		Petja	i [*vin jiji u:h*	
			and [he it u:h	
4		VV	[((Looks back at Petja))	
5	C→	VV	*Til'ky ne rifma, a ryma.*	
			Only not rhyme (Russian), but rhyme (Ukrainian)	
6	U→	Petja	*Ryma.*	
			Rhyme (Ukrainian)	

While Petja has difficulty articulating his thoughts in this demanding task, he eventually succeeds in stating his first point: 'the rhyme is good' (line 1).

As Petja nears the end of this utterance, VV looks away (line 2). However, upon returning attention to Petja (line 4), VV interrupts him to initiate and complete a correction. Using the format 'not X but Y,' he explicitly rejects *rifma* (Russian 'rhyme') and replaces it with *ryma* (Ukrainian 'rhyme') (line 5). Petja acknowledges the correction by repeating *ryma* (line 6).

Thus far this exchange has followed the usual pattern, with the student following the teacher's lead in affirming the problematic nature of using the Russian variant of a loanword. However, Petja's turn in line 1 contains another potential trouble source: *xorošaja* 'good.' This word is a cognate from East Slavic, but Petja's usage follows Russian norms in both morphology (the adjectival suffix *-aja* vs. Ukrainian *-a*) and phonology ([xʌróšaja] vs. Ukrainian [xoróša]). It is not clear whether VV has noticed this word, as he was not attending to Petja when it was uttered (see lines 1–2). In any case, his usual role as correction-initiator is pre-empted, as seen below.

Excerpt 2b
Rhyme (02/26/04)
Trouble source is in ***boldfaced italics***

6 Petja *Rym* [a.
Rhy[me (Ukrainian)
7 C→ Student [*I ne xorošaja a dobra*
[And not good (Russian) but good (Ukrainian)
8 U/TS→ Petja *U:h dobre: nu dobre* ***rifm****- uh*
U:h it nicely: well nicely *rhy*- uh
9 C→ VV *rymuje,*
rhymes,
10 U→ Petja *rymuje i:: tut (.) bil'še bil'še s:liv uh pro zymu u:h*
rhymes a::nd here (.) are more more w:ords uh about winter u:h

Even before Petja finishes repeating *ryma*, an unidentified male student self-selects to correct the second trouble source. Employing the same 'not X but Y' format, he rejects *xorošaja* and replaces it with *dobra*, another word meaning 'good' (line 7). By appropriating this format and linking his utterance to VV's prior talk through use of the coordinating conjunction *i* 'and,' the student formulates his utterance as a continuation of VV's turn in line 5, thus taking on the teacher's authoritative voice as he takes on his role of initiating correction. He thus moves beyond the limited role usually allotted to students in corrective feedback routines and takes responsibility for regulating the norms of classroom language usage.

VV does not acknowledge this correction or give any sign that he has heard it, but continues gazing at Petja with no change in facial expression. However, Petja *has* noticed. Although he has already completed his point about rhyme, Petja returns to the trouble source turn and begins to reformulate it from 'The rhyme is good' to 'It rhymes nicely,' incorporating the replacement word in its adverbial form *dobre* 'nicely' (line 8). However, Petja again encounters trouble, this time with the verb. He utters what appears to be the start of the word *rifmuet* (third person singular of the Russian *rifmovat'* 'to rhyme'), but cuts off before completion (*ryfm-*). Both this cut-off and the *uh* that follows signal the possibility of an upcoming correction initiation (Schegloff *et al.* 1977: 367). VV provides the form *rymuje* (third person singular of the Ukrainian *rymuvaty* 'to rhyme') in line 9, which Petja repeats before moving on to his long-delayed second point (line 10). Although Petja does not make this correction himself, by breaking off his utterance before completion he has displayed awareness that it is potentially problematic.

These incidents illustrate children's emerging ability to monitor linguistic output to avoid Russian forms. This ability requires both (i) linguistic knowledge to recognize distinctions between languages and (ii) social knowledge to realize that such distinctions are relevant to speaking correctly. Such instances demonstrate that as children participate in corrective feedback routines, they are not simply repeating the teacher's corrections, but appropriating the practices of correction that will enable them to take on more responsibility for monitoring their speech in line with pure language norms.

APPROPRIATING IDEOLOGIES OF CORRECTNESS

The implicit messages regarding the inappropriateness of language mixing conveyed in these corrective feedback routines occasionally surfaced in the form of explicit metalinguistic commentary. For example, during an exercise in which children were generating synonyms for *xurtovyna* 'snowstorm,' a child suggested *v'juha*, which appears to be the Russian *v'juga* 'blizzard' with [g] (voiced velar stop) altered to [h] (voiced pharyngeal fricative) in accordance with Standard Ukrainian phonological norms. In response, VV waved his hand, shook his head laterally and stated, '*V'juha* is a Russianism' before turning to call on another child. Through routine deployment as negative feedback, designating a form as 'Russian' or 'Russianism' became equivalent to labeling it 'incorrect.' Children occasionally used these terms in similar ways, as seen in the following.

This incident occurred during a language lesson as the class was reviewing homework in which they had provided antonyms for certain words. As Excerpt 3a begins, VV calls on Marko to suggest an antonym for *lahidnyj* 'gentle.'

Excerpt 3a
That's Russian (02/05/04)
Trouble source is in ***boldfaced italics***

1 VV ((Points to Marko))
2 VV *Bud' laska Mar [ko.*
Please Mar[ko.
3 TS→ Marko [***Grubyj.***
[***Rough.***
4 VV *Jakyj?*
What?
5 TS→ Marko ***Grubyj.***
Rough.
6 C→ VV *Hru [byj.*
Rou[gh.
[((Turns away from Marko))

In line 3, Marko offers *grubyj* 'rough,' a cognate from East Slavic distinguished by its pronunciation: /grúbyj/ with an initial /g/ in Standard Russian and /hrúbyj/ with an initial /h/ in Standard Ukrainian. Marko's pronunciation follows Standard Russian norms. In line 4, VV indicates a problematic hearing with '*Jakyj*?' 'What?' As Marko's turn was spoken in partial overlap with VV's prior turn, VV may not have heard it; alternatively, VV may be prompting Marko to self-correct. Marko responds to this repair initiation as indicating an uncertain hearing and repeats the word without alteration (line 5). VV then corrects him by saying the word with word-initial [h] as in Standard Ukrainian (line 6). However, he does not overtly indicate that Marko's pronunciation was problematic, nor does he provide an opportunity for Marko to take up the correction, as he closes the sequence by turning away even before he has completed his utterance (line 6). That is, neither party explicitly orients to the word as an error. However, another child has a different agenda.

Excerpt 3b
That's Russian (02/05/04)

7 VV *Šče jakyj.*
What else.
8 VV ((Points to Dar'ja))
9 C→ Dar'ja *Cja rosi- uh cja rosijs'ka. Suvoryj.*
That's Rus- uh that's Russian. Cruel.
10 VV ((Turns away from Dar'ja, walks to his right))
11 VV *Suvoryj. Ditky vam ne zdajet'sja ščo tut na žal'*
Cruel. Children doesn't it seem to you that here unfortunately

12 VV *nemaje toho sumnivnoho /e/ /y/ (.) jakyj nam tak potriben?*
we don't have that that alternating /e/ /y/ (.) that we need?
13 → VV *Hrubyj (.) do reči, ce ukrajins'ke slovo takož.*
Hrubyj (.) by the way, is a Ukrainian word also.

VV next calls on Dar'ja, who begins by stating, 'That's Russian' (line 9). As she does not specify a referent for the demonstrative *cja* 'that,' it is unclear whether it is Marko's pronunciation or the word itself that she is labeling as 'Russian.' In either case, Dar'ja problematizes her classmate's answer before offering a replacement, *suvoryj* 'cruel.' As children in this classroom were rewarded for identifying and correcting classmates' factual errors, Dar'ja may have sensed an opportunity to bid for recognition.

However, Dar'ja's claim is problematic in several respects. First, it is misleading (the word exists in Ukrainian) and unnecessary (VV has already corrected Marko's pronunciation). In addition, children were supposed to correct classmates only when VV invited them to do so. Dar'ja has not only self-selected, but she has targeted a word that VV has implicitly accepted, which could be interpreted as challenging his authority. In his following turn VV distances himself from Dar'ja both physically (walking away from her, line 10) and verbally. He first negatively assesses her suggestion of *survoryj*, noting that it does not meet the requirements of the exercise (lines 11–12).[9] He then expressly disagrees with her claim, declaring that *hrybyj* (with stress on the word-initial [h]) is also a Ukrainian word (line 13). Dar'ja's attempt at making a correction is thus unsuccessful. Nevertheless, this episode demonstrates her awareness of the importance of pure language norms in the classroom and how these norms could be invoked to claim superior knowledge.

SOCIALIZATION INTO SPEAKING CORRECTLY

The preceding provides a limited but representative sampling of the organization and targets of corrective feedback that occurred in this classroom. Children's family backgrounds were not reliable predictors of their propensity to use Russian words; Slava (Excerpt 1) came from a bilingual 'mixed' family (Russian mother, Ukrainian father), Petja (Excerpt 2) from a Ukrainian-speaking Ukrainian family, and Marko (Excerpt 3) from a Russian-speaking Ukrainian family. Over the course of my observations all of the children in the class were corrected for Russian-influenced errors, most on numerous occasions.

Of course, it is the business of language teachers to correct errors. But how is error defined? Labeling a form *incorrect* can be grounded in a number of criteria, such as incongruity with the structure of the language or communicative inadequacy. However, many of the forms corrected in this classroom were well fitted to Ukrainian norms. For example, Russianisms such as *komandyrovka* 'business trip' or *v'juha* 'blizzard' do not violate Ukrainian phonological rules, and in terms of form there is nothing to distinguish them from legitimate loanwords such as *ryma* 'rhyme' (which may have entered Ukrainian via Polish).[10] Nor can it be argued that Russian forms obscured the speaker's meaning. As Schegloff *et al.* (1977: 380) have noted, if the understanding of a turn is sufficient for an 'other' to make a correction, it is also sufficient for that 'other' to produce a sequentially appropriate next turn instead of a correction. The ability of VV or another student to supply a replacement word demonstrates that these forms presented no barrier to comprehension, nor would they prove troublesome outside the classroom, where both Ukrainian and Russian are understood.

Yet words may be referentially accurate, linguistically plausible, and perfectly comprehensible, but still be judged incorrect in terms of social meaning. These words were incorrect because they were associated with the 'wrong' language. While it is not unusual to discourage use of non-target languages in language classrooms, this practice rests on an assumption that languages are discrete entities with clearly defined borders, an assumption difficult to support when set against the language practices of this bilingual community. Indeed, when I mentioned a Russian form that one of the children had used in class, Ukrainian-speaking acquaintances would often admit that they used that form as well.

Moreover, the boundaries between Ukrainian and Russian are not as clear or impermeable as this strict compartmentalization of languages implies. In addition to cognates, many distinctions between Standard varieties (e.g. Russian [g] vs. Ukrainian [h]) become blurred or disappear at the dialectal level. The treatment of a word such as *komandyrovka* 'business trip' as an error despite its inclusion in the official (Soviet) Ukrainian dictionary further suggests that these corrections cannot be explained solely in terms of a 'Ukrainian only' philosophy; they also involved judgements regarding what should or should not be accepted *as* Ukrainian.[11]

The authority of pure Ukrainian seems to have been accepted in this classroom. Children displayed willingness to uphold pure language norms regardless of their own diverse ethnic or linguistic backgrounds; for example, although Dar'ja (Excerpt 3) came from a mixed (Russian mother, Ukrainian father) Russian-speaking family, she did not hesitate to disapprove of a 'Russian' word. And although it is possible to read children's use of Russian as resistance to these norms, participants themselves did not orient to it as such. While disobedient behavior could draw a strong rebuke, I never observed VV scold children for using Russian words, and he informed me that he did not penalize them for language mixing, feeling that they could not yet be held accountable for their language use. For their part, while children occasionally challenged VV on issues such as grades or the acceptability of an answer, they regularly took up corrections without protest. That is, all parties treated use of Russian forms as inadvertent errors rather than as deliberate acts of resistance. Within the classroom, pure language had become a dominant language ideology.

CONCLUSION

The impact of these corrective feedback practices extends beyond the walls of the classroom. While serving the pedagogical goal of teaching children to speak Ukrainian 'correctly,' these practices were also socializing children into a particular understanding of what 'speaking correctly' means. As children participated in corrective feedback routines, whether by taking up a teacher's replacement word, correcting themselves, or correcting classmates, they displayed allegiance to ideologically mediated standards of correctness that proscribe language mixing as a violation of the natural boundaries between languages, thereby reifying and naturalizing pure Ukrainian as the standard upon which all Ukrainian language practices can be evaluated. While at odds with community language practices, such standards were quite consonant with community language ideologies and were vocally supported by many community members, including the parents of some of these children. They also reflected and validated the valorization of pure language evoked through state-sponsored efforts to revitalize Ukrainian and establish it as a distinct language suitable for representing a distinct nation.

In bringing together two complementary research traditions—language socialization and language ideology—this study has underscored the historicized and ideological nature of corrective feedback routines in this classroom and illustrated how seemingly mundane classroom practices may be implicated in larger sociopolitical phenomena. It therefore contributes to an emerging body of language socialization research that has explored the impact of everyday socializing activities and their governing language ideologies on the complex processes of linguistic reproduction and change in multilingual societies (see Garrett and Baquedano-López 2002). In particular, these findings have implications for the role of schooling in language standardization and revitalization, areas that thus far have received little attention from language socialization researchers. As this study has shown, the discourse of the language classroom, a setting where appropriate ways of speaking are overtly displayed and promoted, is a potentially rich site through which standard language ideologies are reproduced, sustained, and transmitted.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

ACKNOWLEDGEMENTS

Versions of this paper were presented at AAAL 2007 in Costa Mesa, California and the 2007 Symposium on Language Socialization, Interaction, and Culture at UCLA. I am grateful to the Institute of International Education, which funded my fieldwork in Ukraine through a U.S. Student Fulbright Award, and to the Ukrainian teachers and students who welcomed me into their classrooms. I would also like to thank Elinor Ochs, Marjorie Goodwin, Paul Kroskrity, Olga

Yokoyama, Shawn Loewen, and the editors and anonymous reviewers at *Applied Linguistics* for insightful comments on previous drafts. All remaining errors are my own.

NOTES

1 In the 2001 All-Ukrainian Population Census, 17.5% of respondents identified their ethnicity as Russian (State Statistics Committee of Ukraine n.d.).

2 In a survey, when asked for 'language of preference' one-third of those who self-identified as ethnic Ukrainian designated Russian (Arel 2002: 238).

3 The *Dictionary of the Ukrainian Language* defines *suržyk* as 'elements of two or more languages, joined artificially, without adhering to the norms of the literary language, impure language' (Academy 1970–1980, vol. IX: 854).

4 A *gymnasium* is a selective public school that offers intensive instruction in certain subjects along with the regular curriculum.

5 I acknowledge, however, that those critical of current language education policy may have been reluctant to air their views to a stranger.

6 Russian is no longer an obligatory subject in Ukrainian schools, and many have dropped it from the curriculum. According to the principal, Russian at this school was retained at the request of parents.

7 All names are pseudonyms. Following Ukrainian practice, I refer to the teacher by his first name (Viktor) and patronymic (Viktorovych), derived from one's father's first name. This is a formal mode of address equivalent to *Mister* with a last name.

8 A table detailing students' ethnic and linguistic backgrounds can be found at *Applied Linguistics* online.

9 The exercise focused on words containing *e* or *y*, whose pronunciation varies depending on stress, and the stress pattern of the antonym was supposed to differ from that of the original word.

10 Participants demonstrated no concern about the Polish origin of many words in the Ukrainian lexicon, nor did classroom use of English borrowings such as *supermodel* 'supermodel' or *dyzajner* 'designer' generate any response. This apparent inconsistency underscores the selective nature of pure language ideologies (e.g. Annamalai 1989).

11 As an anonymous reviewer has pointed out, what counts as 'Standard Ukrainian' is far from resolved. While discussion of this issue is beyond the scope of this article, interested readers are referred to the account in Bilaniuk (2005).

REFERENCES

Academy of Sciences of the Ukrainian SSR. 1970–1980. *Slovnyk Ukrajins'koji Movy* [Dictionary of the Ukrainian Language]. Vydavnycvo Naukova Dumka.

Annamalai, E. 1989. 'The linguistic and social dimensions of purism' in B. H. Jernudd and M. J. Shapiro (eds): *The Politics of Language Purism*. Mouton de Gruyter, pp. 225–31.

Arel, D. 1996. 'A lurking cascade of assimilation in Kiev?,' *Post-Soviet Affairs* 12/1: 73–90

Arel, D. 2002. 'Interpreting ''nationality'' and ''language'' in the 2001 Ukrainian census,' *Post-Soviet Affairs* 18/3: 213–49.

Bilaniuk, L. 2005. *Contested Tongues: Language Politics and Cultural Correction in Ukraine.* Cornell University Press.

Bilaniuk, L. and **S. Melnyk.** 2008. 'A tense and shifting balance: bilingualism and education in Ukraine,' *The International Journal of Bilingual Education and Bilingualism* 11/3 & 4: 340–72.

Blommaert, J. and **Verschueren.** 1998. 'The role of language in European nationalist ideologies' in B. B. Schieffelin, K. A. Woolard, and P. V. Kroskrity (eds): *Language Ideologies: Practice and Theory*. Oxford University Press.

Bourdieu, P. 1977. *Outline of a Theory of Practice* (R. Nice trans.). Cambridge University Press.

Bourdieu, P. 1991. 'The production and reproduction of legitimate language' in *Language and Symbolic Power* (J. Thompson and M. Adamson, trans.). Harvard University Press (Original work published in 1980).

Duff, P. 1995. 'An ethnography of communication in immersion classrooms in Hungary,' *TESOL Quarterly* 29/3: 505–37.

Duff, P. 1996. 'Different languages, different practices. Socialization of discourse competence in dual-language school classrooms in Hungary' in K. Bailey and D. Nunan (eds): *Voices From the Language Classroom: Qualitative Research in Second Language Acquisition*. Cambridge University Press.

Flier, M. S. 2000. '*Surzhyk*: The rules of engagement' in Z. Gitelman (ed.): *Cultures and Nations of Central and Eastern Europe: Essays in Honor of Roman Szporluk*. Harvard University Press.

Garrett, P. B. and **P. Baquedano-López.** 2002. 'Language socialization: reproduction and continuity, transformation and change,' *Annual Review of Anthropology* 31: 339–61.

Howard, K. 2004. 'Socializing respect at school in Northern Thailand,' *Working Papers in Educational Linguistics* 20/1: 1–30.

Irvine, J. and **S. Gal.** 2000. 'Language ideology and linguistic differentiation' in P. V. Kroskrity (ed.): *Regimes of Language: Ideologies, Polities, and Identities*. School of American Research.

Jacobs-Huey, L. 2007. 'Learning through the breach: language socialization among African American cosmetologists,' *Ethnography* 8/2: 171–203.

Jefferson, G. 1987. 'Exposed and embedded corrections' in G. Button and J. R. E. Lee (eds): *Talk and Social Organization*. Multilingual Matters.

Magocsi, P. R. 1996. *A History of Ukraine*. University of Toronto Press.

Martin, T. 2001. *The Affirmative Action Empire: Nations and Nationalism in the Soviet Union, 1923–1939*. Cornell University Press.

Ochs, E. 1984. 'Clarification and culture' in D. Schiffrin (ed.): *Georgetown University Round Table in Languages and Linguistics*. Georgetown University Press.

Ochs, E. 1988. *Culture and Language Development*. Cambridge University Press.

Ochs, E. and **B. B. Schieffelin.** 1984. 'Language acquisition and socialization: three developmental stories and their implications' in R. Shweder and R. Le Vine (eds): *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge University Press.

Ochs, E. and **B. B. Schieffelin.** 1995. 'The impact of language socialization on grammatical development' in P. Fletcher and B. MacWhinney (eds): *The Handbook of Child Language*. Blackwell.

Pavlenko, A. 2006. 'Russian as a lingua franca,' *Annual Review of Applied Linguistics* 26: 78–99.

Peredrij, H. R., L. V. Skurativs'kyj, H. T. Šelexova and **J. L. Ostaf**. 2002. *Ridna Mova*: *Pidručnyk dlja 5 Klacu* [*Native Language: Textbook for the 5th Grade*]. Osvita.

Poole, D. 1992. 'Language socialization in the second language classroom,' *Language Learning* 42/4: 593–616.

Schegloff, E. A., G. Jefferson, and **H. Sacks.** 1977. 'The preference for self-correction in the organization of repair in conversation,' *Language* 53/2: 361–82.

Schieffelin, B. B. 1990. *The Give and Take of Everyday Life. Language Socialization of Kaluli Children*. Cambridge University Press.

Shevelov, G. 1989. *The Ukrainian Language in the First Half of the Twentieth Century (1900–1941). Its State and Status*. Harvard University Press.

Shevelov, G. 1993. 'Ukrainian' in B. Comrie and G. G. Corbett (eds): *The Slavonic Languages*. Routledge.

Silverstein, M. 1996. 'Monoglot ''standard'' in America: standardization and metaphors of linguistic hegemony' in D. Brenneis and R. Macaulay (eds): *The Matrix of Language: Contemporary Linguistic Anthropology*. Westview Press.

Silverstein, M. 1998. 'Contemporary transformations of local linguistic communities,' *Annual Review of Anthropology* 27: 401–26.

Silverstein, M. 2000. 'Whorfianism and the linguistic imagination of nationality' in P. V. Kroskrity (ed.): *Regimes of Language: Ideologies, Polities, and Identities*. School of American Research.

State Statistics Committee of Ukraine. n.d. All-Ukrainian population census 2001. Linguistic composition of population. Retrieved August 5, 2009 from http://www.ukrcensus.gov.ua/eng/results/general/language.

Taylor, J. 1998. 'Beyond politics: internal problems of the Ukrainian language,' *The Ukrainian Review* 45: 3–28.

Wexler, P. N. 1974. *Purism and Language: A Study in Modern Ukrainian and Belorussian Nationalism (1840–1967)*. Indiana University Press.

Applied Linguistics: 31/3: 368–390 © Oxford University Press 2009
doi:10.1093/applin/amp038 Advance Access published on 16 October 2009

Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics

JENIFER LARSON-HALL and RICHARD HERRINGTON

University of North Texas

In this article we introduce language acquisition researchers to two broad areas of applied statistics that can improve the way data are analyzed. First we argue that visual summaries of information are as vital as numerical ones, and suggest ways to improve them. Specifically, we recommend choosing boxplots over barplots and adding locally weighted smooth lines (Loess lines) to scatterplots. Second, we introduce the reader to robust statistics, a tool that can provide a way to use the power of parametric statistics without having to rely on the assumption of a normal distribution; robust statistics incorporate advances made in applied statistics in the last 40 years. Such types of analyses have only recently become feasible for the non-statistician practitioner as the methods are computer-intensive. We acquaint the reader with trimmed means and bootstrapping, procedures from the robust statistics arsenal which are used to make data more robust to deviations from normality. We show examples of how analyses can change when robust statistics are used. Robust statistics have been shown to be nearly as powerful and accurate as parametric statistics when data are normally distributed, and many times more powerful and accurate when data are non-normal.

INTRODUCTION

Statistics play an important role in analyzing data in all fields that employ empirical and quantitative methods, including the second language acquisition (SLA) field. This article is meant to address issues that are pertinent to the field of SLA, given our own constraints and parameters. For example, one statistical problem that we probably cannot avoid is the lack of truly random selection in experimental design, which Porte (2002) has noted. Given the populations we try to test and issues of validity versus reliability (do we use intact classrooms and get 'real' data, or use laboratory tests that can randomize better and get more 'reliable' data?) there is no simple way to always use true randomization in populations we test. However, there are other statistical issues in SLA that are amenable to improvement. For example, many SLA research designs use small sample sizes (generally less than 20 per group), meaning that the statistical power of a test of a normal distribution may be low (making it hard to reliably test whether data is normally distributed or not), yet these studies use parametric statistics which assume a normal distribution. Another problem with any size group is reliably identifying outliers.

In this article we will put forward two broad types of techniques which researchers can use to improve the quality of their statistical analyses. The first suggestion is to use graphic techniques that are the most helpful in understanding data distributions in order to assess statistical relationships and differences between groups. The second suggestion is that researchers learn about and begin to incorporate statistics into their statistical analyses that are robust (or in other words, insensitive to) violations of assumptions of a normal distribution.

GRAPHICS

Introduction

Because doing a statistical analysis is as much an art as a science (Westfall and Young 1993: 20), researchers need to provide as much information about their data as possible to their reading audience.[1] The best kinds of visual information can help readers verify the assumptions about the data and the numerical results that are presented in the text and provide intuitions about relationships or group differences. The American Psychological Association (APA) Task Force on Statistical Information (Wilkinson 1999) recommends always including visual data when reporting on statistics.

Tufte (2001) claims that improving the resolution of our graphics by providing as much information as possible may lead to improvements in the science we perform. At present, most published articles in the field of SLA, if they present graphics, show a barplot if the data are distributed into groups, and a scatterplot if the data involves relationships between variables. We suggest that these graphics be improved by using boxplots instead of barplots for group-difference data and adding Loess lines to scatterplots for relational data.

Boxplots instead of barplots

Barplots are popular in the SLA field. In the five years of papers published in *Applied Linguistics*, *Language Learning* and *Studies in Second Language Acquisition* from 2003 to 2007 that we examined, 110 studies contained group difference quantitative data that could have been represented with boxplots. However, of those 110 studies, only one used a boxplot, while 46 used barplots. An additional 12 used line graphs (the remainder did not provide graphics). A novice to the field would assume that barplots were the graphic of choice for SLA researchers, and continue to follow this tradition. However, barplots (and line graphs) are far less informative than boxplots, providing only one or two points of data (depending on whether error bars are used) compared with the five or more points that boxplots provide. While both types of plots may be somewhat impoverished by Tufte's (2001) standards, boxplots

Table 1: A comparison of the information used to create the boxplot versus the barplot for the 'Late' group in Figure 1

	Boxplot	Barplot
Mean	–	3.10
First quartile	2.3	–
Median (second quartile)	2.9	–
Third quartile	3.8	–
Minimum score	1.6	–
Maximum score	4.9	–
Outliers labeled	Yes	No

should always be preferred over barplots unless the data are strictly frequency data, such as the number of times that one teacher uses recasts out of the total number of instances of negative evidence.[2] In fact, one reviewer of this article lauded the recommendation to use boxplots over barplots and said, 'If we had a contest on which graphical method conveys the least amount of information and has the best potential to mislead, barplots would win easily'. Table 1 shows the information that is used to calculate both types of graphics that are shown in Figure 1. Table 1 clearly shows how impoverished the data used in the barplot is.

Figure 1 gives an example of a barplot and a boxplot of the same data, compared side by side.

Notice that the data look different in the two kinds of graphics. The boxplot provides far more information about the *distribution* of scores than the barplot. One of the advantages of the boxplot (invented by Tukey, 1977) is that it is helpful in interpreting the differences between sample groups without making any assumptions regarding the underlying probability distribution, but at the same time indicating the degree of dispersion, skewness, and outliers in the given data set. For example, in looking at the boxplot in Figure 1 (the graph on the right) we notice that the range of scores is wide for the non-native speakers (as indicated by the length of the whiskers on either side of the box for the 'Non', 'Late', and 'Early' labels), but quite narrow for the native speakers (NS). We can also note an outlier in the NS scores. Boxplots are robust to outliers but barplots may change considerably if only one data point is added or removed. Lastly, we could note that the data for the NS is *not* symmetric, since there is only a lower whisker but no upper whisker. This means the distribution is skewed. The other distributions in Figure 1 are slightly skewed as well, as their medians are not perfectly in the center of the boxes and/or the boxes are not perfectly centered on the whiskers.

Because many readers may not be familiar with boxplots, Figure 1 labels the parts of the boxplot (which is notched in this case, although it doesn't have to be). While a barplot shows the mean score, the line in the middle of the

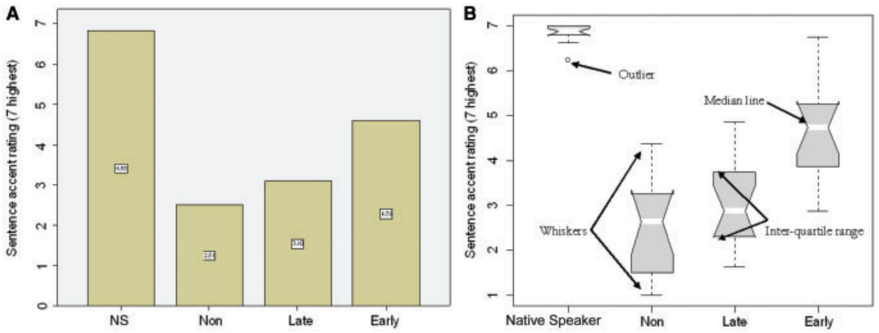

Figure 1: Comparison between a barplot (A) and a boxplot (B) of the same data

boxplot (here, in white) shows the median point. The length of the box contains all of the points that comprise the 25th to 75th percentile of scores (in other words, the first to third quartiles), and this is called the interquartile range (IQR). The ends of the box are called the hinges of the box. The whiskers of the boxplot extend out to the minimum and maximum scores of the distribution, unless these points are distant from the box. If the points extend more than 1.5 times the IQR above or below the box, they are indicated with a circle as outliers (there is one outlier in the NS group). The notches on the boxplot can be used to get a rough idea of the 'significance of differences between the values' (McGill *et al*. 1978). This is not exactly the same as the 95% confidence interval; the actual calculation in R is ±1.58 IQR/sqrt(n) (see R help for 'boxplot.stats' for more information). If the notches lie outside the hinges (outside the box part), as they do just slightly for the Non and Early groups, this would indicate low confidence in the estimate (McGill *et al*. 1978).

Readers who have been convinced that boxplots are useful will find that it is easy to switch from barplots to boxplots since practically any program which can provide a barplot (SPSS, SAS, S-PLUS, R) can also provide a boxplot. Directions for making boxplots in SPSS and R are included in the online Appendix A.

Loess lines on scatterplots

A move from barplots to boxplots will improve visual reporting with group difference data. A way to improve visual reporting of relationships between variables is to include a smoother line along with the traditional regression line on a scatterplot (Wilcox 2001). Smoothers provide a way to explore how well the assumption of a linear association between two variables holds up. If the smoother line and regression line match fairly well, confidence is gained in assuming that the data are linear enough to perform a correlation

(Everitt and Dunn 2001). There are many kinds of smoothers (Hastie and Tibshirani 1990), but the one that is used often for fitting non-parametric curves through data by authors such as Wilcox (2001) and Crawley (2007) is Cleveland's smoother, commonly called the Loess line (Wilcox 2001). This line is a locally weighted running-line smoother, and it calculates lines over small intervals of the data using weighted least squares. In layman's terms, it is like regression lines are being calculated for small chunks of the data at a time. Clearly, if the concatenation of locally produced regression lines matches the regression line calculated over the entire data set, the assumption of linearity throughout the data set is upheld. Figure 2 shows four sets of data that contain both regression lines and Loess lines (note that these graphs are meant for illustrative purposes only, not for making actual inferences about relationships of the variables labeled).

Although the smoother line can be used as a guide, it is impossible to set out infallible guidelines for visually determining whether the regression line is

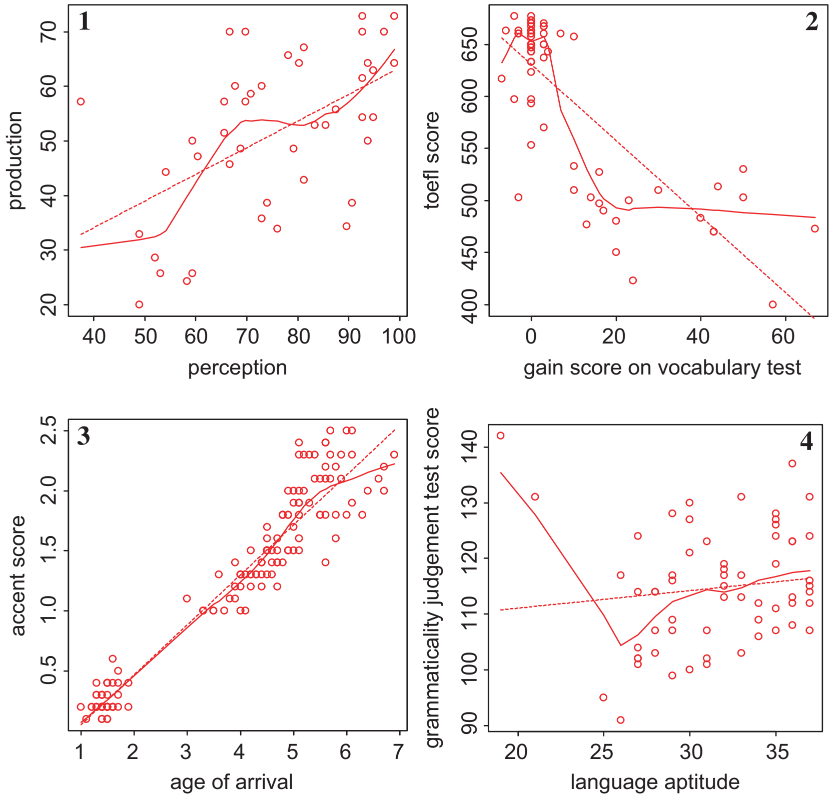

Figure 2: Four scatterplots with superimposed regression (dotted) and Loess lines (solid)

'close enough' to the Loess line to say that the data are linear (formal methods for testing curvature do exist however; see Wilcox 2005: 532–3). This is a matter of judgement that will improve with seeing more examples, which is why researchers who make claims about relationships between variables should provide scatterplots that contain both regression and Loess lines. Then, no matter what the author claims, readers will be able to make judgements for themselves on the appropriateness of assuming a linear relationship between the variables.

In Figure 2, we would say that the Loess lines in graphs 1 and 3 are 'close enough' to be considered linear. On the other hand, the Loess line in graph 2 shows a large deviation from a straight line, and it is likely the data should be analyzed as two different groups, as there seem to be two different patterns in the data. In graph 4, there appears to be a modest positive correlation between the variables, but the two outliers at the far left of the graph have skewed the regression line to be essentially flat. The smoother line shows a sharper angle in the non-outlier data.

Directions for creating a Loess line over a scatterplot in SPSS and R can be found in the online Appendix A. Other graphics that we don't discuss here, such as the relplot (which resembles the plot of ellipses shown later in this article in Figure 7; see Wilcox 2003 for more information) can help identify outliers in relationships between two variables. The kernel density estimator (g2plot using Wilcox's commands; see Wilcox 2003: 87 for an example) is an improvement on the histogram and can give a different perspective from boxplots. In addition, the shift function is a good graphic for comparing two groups (see Wilcox 2003: 276). A whole variety of exciting graphs that can be used with R can be viewed at addictedtor.free.fr/graphiques.

ROBUST STATISTICS

Introduction

In this section we explain to our reader why robust statistics are a desirable and useful tool to learn more about. What we call here robust statistics are not new; in fact, many of the robust alternatives to standard statistical estimates were proposed by scientists in the late 19th and early 20th centuries. However, the foundational works on robust statistics were published in the 1960s and early 1970s, with works such as Tukey (1960, 1962), Huber (1964) and Hampel (1968).[3] While work has continued vigorously on robust statistics since that time, practically speaking one needs statistical programs and adequate computational power in order to use robust statistics, and these requirements have only just come into view in the recent past[4] (we prefer the free R statistical program, see http://www.r-project.org; Maronna *et al.* 2006 assert that the most complete and user-friendly robust library is the one found in S-PLUS, which is also available in R; Rand Wilcox also has many robust functions that can be incorporated into R or S+ and are available at http://www-rcf.usc.edu/~rwilcox/in the allfun or Rallfun files).

The programs are available, the computers are fast enough, and researchers can now begin to take advantage of the improvements that incorporating robust statistics into their own work will provide. Appendix A, found online, will provide some code to understand how we ran all of the robust statistics that are used in this article.

We will introduce below the concepts of trimmed means and bootstrapping, which are useful procedures that can help readers understand how robust statistics differ from classical statistics. Before we do that, however, readers will want to know why the use of robust statistics is desirable. Conventional wisdom has often promoted the view that standard analysis of variance (ANOVA) techniques are robust to non-normality, and that small deviations from the idealized assumptions of statistical tests (such as a normal distribution) would result in only minimal error in conclusions that were reached. Such is the view still of almost any book on statistics or research methods that you could lay hands on in the social sciences, which may make readers somewhat skeptical of our claim. This view is fairly accurate only with respect to Type I error (Wilcox 2001) (rejecting the null hypothesis when in reality it is true, and there actually is no difference between groups). When it is assumed that there are no differences between groups in a group difference testing setting (for example, one might want to show that a group of advanced non-native speakers do not differ from a native speaker group), then the probability level corresponding to the critical cut-off score, used to reject the null hypothesis, is found to be close to the nominal level of 0.05. However, statistical simulation studies have found that standard methods are not robust when differences exist (Tukey 1960; Hampel 1973), which is more often the situation that researchers are hoping for (such as, for example, when two treatments are applied and the researcher is hoping that one will result in more language learning).

Tukey (1960) found that one of the most problematic distributions was one he called a 'contaminated normal' distribution, which visually is quite close to a normal distribution. The contaminated normal is slightly longer-tailed than normal distributions (Huber 1981; Wilcox 2001), as can be seen in Figure 3. The contaminated normal is formed mathematically by taking two normal populations with the same mean, but with one that has a larger standard deviation than the other, and mixing data from the population with the wider standard deviation into the population with the narrower standard deviation (Tukey 1960).

The problem with the longer tails of the contaminated normal is that the extra data points in the tails means that the amount of variability is increased, and this makes it more likely that differences which are in fact statistical[5] are found to be non-statistical (Tukey 1960; Huber 1981; Wilcox 2001). The reason this is important to SLA data is that real data sets in Applied Linguistics are probably not *exactly* normally distributed (Micceri 1989 claims

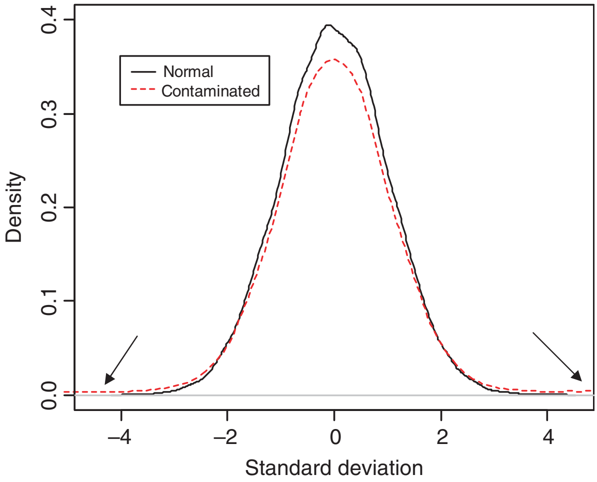

Figure 3: Density function of a normal distribution and a superimposed contaminated normal distribution

this for psychological data), and may demonstrate deviations from normality including heavier tails (as evidenced by outliers) or skewness. As readers can see in Figure 3, it would be quite difficult to tell the difference in a data set between data with an exactly normal distribution versus a distribution that is symmetric but heavy-tailed. Even small departures from normality (not to mention much larger ones such as obvious skewness) can have an effect on the statistical conclusions that can be drawn.

Wilcox (2001) notes that in a standard normal distribution the variance is 1, but in a contaminated normal like that in Figure 3 the variance has increased to 10.9. Such inflation of the variance means that the standard error will also be inflated, and since statistical tests divide by some measure of variability like variance or standard error, the resulting statistic will be smaller when the variance is larger (and less likely to be statistical).

An illustration from Wilcox (2003) can help clarify this point. Imagine we have 10 data points for 2 groups, shown in Table 2. For the sake of this article, let's say they represent scores on a test of how much vocabulary, out of a possible 25 points, was remembered after Group 1 received no treatment (the control group) and Group 2 received a special treatment (the treatment group). The mean score of the control group is 5.5 and the mean of the treatment group is 8.5. Is this difference statistical? To test the null hypothesis that there is no difference between the groups, apply an independent samples *t*-test. The *t*-test value is $t = -2.22$ and $p = 0.039$. The *p*-value is below the normal alpha level of $\alpha = 0.05$, and thus we may reject the null hypothesis, and conclude there is a statistical difference between groups. However, say that the score of the 10th participant in the treatment group is changed from 13 to 25. Now the

Table 2: Original scores for a fictional vocabulary retention experiment

Group 1: control	1	2	3	4	5	6	7	8	9	10
Group 2: treatment	4	5	6	7	8	9	10	11	12	13

average of the treatment group (Group 2) becomes 9.7. Logically, because the difference between sample means has increased, we would still want to conclude that there is a statistical difference between groups. However, because the score of 25 increases the variance (the distance from the mean) in the treatment group, this increases the denominator of the *t*-test equation,

$$t_{\mathrm{df}} = \frac{X_T - X_C}{\sqrt{\mathrm{var}_{\mathrm{pooled}}\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}},$$ [6]

leaving us with a smaller *t*-value ($t = -1.99$) and a *p*-value larger than our alpha ($p = 0.07$). In other words, with the one changed value we now conclude that the groups are statistically *not* different! This goes counter to our sense of group differences, but shows that more data in the tails of the distribution, and thus more variance, can affect *p*-values and statistical conclusions.

To summarize thus far, while small deviations from normality in the distribution are fairly robust to Type I errors (rejecting the null hypothesis when in reality it is true, and there actually is no difference between groups), we are much more likely to make a Type II error (accepting the null hypothesis when in reality it is not true and there actually is a difference between groups) with such deviations (Hampel *et al.* 1986). Making Type II errors means that we are losing power to find true differences between groups or relationships between variables. Power is a technical statistical term, but can be understood here in layman's terms to mean the strength to find a result.

We will give an example of the kind of problems that have been found with small departures from normality. Wilcox (1995) reported on the power of the Welch procedure that is used in *t*-tests when variances are unequal. The power of this test to find the true results when the distribution is normal is 0.93 (where 1.00 is perfect power), but drops to 0.28 when the distribution is a contaminated normal with a standard deviation of 10, and to 0.16 when the contaminated normal has a SD of 20 (Wilcox 1995: 69). On the other hand, a test procedure based on 20% trimmed means (a robust method described in more detail below) yields power of .89 with the normal distribution, and only lowers to 0.78 for a contaminated normal with $K = 10$, and 0.60 with a contaminated normal of $K = 20$ (ibid.). Statisticians agree that robust statistics are even more necessary when statistical models more complex than *t*-tests are used (Hampel *et al.* 1986).

Statistical modeling has shown that robust methods work much better than parametric methods when the underlying distribution is not normal, and they work nearly as well if the underlying distribution is in fact normal (Tukey 1960; Yuen and Dixon 1973; Huber 1981; Luh and Guo 2001; Wilcox 2001). As has been discussed above, the idea that statistical tests are robust to small deviations from normality should not be assumed. Additionally, rules of thumb, such as those which assert that if group sizes are 30 or more there is no reason to worry about meeting normality requirements (Pallant 2001; Weinberg and Abramowitz 2002), are also inaccurate. Westfall and Young (1993) performed a simulation study which found that with group sizes of $n = 160$, skewed distributions, even without outliers, could have very poor results. Using data from an actual study, Wilcox (2003: 123) found that even with $n = 105$, the *t*-test performed poorly and more than 300 subjects would have been necessary to get good results. Remember that poor results mean that although there may indeed be differences between groups or relationships between variables, traditional parametric statistics will not be able to detect that difference.

Huber (1981) states that robust methods are more similar to parametric methods than nonparametric or distribution-free methods, because they continue to use the same parametric models; the difference is that the parametric models 'are no longer supposed to be literally true, and . . . one is also trying to take this into account in a formal way' (Huber 1981: 6). Since robust methods can deal with non-normality, including skewness, and because it is nearly impossible for researchers to know with certainty that their distributions are normal ones, we know of no reason not to recommend that researchers learn more about robust methods and employ them in all cases.

Outliers

Many are familiar with Mark Twain's quote that there are 'lies, damn lies, and statistics' (see the August 2005 issue of *Statistical Science* for a sophisticated and sometimes tongue-in-cheek discussion of how such lying may be accomplished). One reason this aphorism may resonate with those who have used statistics in their own research is that the addition or subtraction of just one participant, or an incorrect data entry for one participant may result in a totally opposite conclusion to the one reached before the participant was added or subtracted or before the data entry was corrected. The kinds of non-robust estimators, such as the average, that are used in parametric statistics can be easily affected by just one extreme point.

Many researchers realize this, and perform their statistics with outliers removed, usually showing the reader a graph so that the outlier's 'outlyingness' can be perceived, and sometimes performing statistics with the outlier both included and removed. Removing outliers is definitely an important step to take to make the data fit the assumptions of normality that are imposed by classical parametric statistics. There are several problems with this ad hoc basis for removing outliers, however. The first problem is that throwing away data points that seem to be outliers results in the non-independence of the remaining data (Wilcox 1998), and independence of the data is one of the assumptions for all statistical tests. Huber affirms that 'classical normal theory is not applicable to cleaned samples' (1981: 4). The second problem is that the decision about what points to remove is personal and subjective. Robust methods provide objective and replicable ways of diagnosing outliers and then performing statistical inferences with these outliers removed (Hampel *et al.* 1986). The third problem is that what is often the problem in a distribution is not the obvious outlier but the 'outliers' to the normal distribution which reside in the heavier tails of the contaminated normal and are not easily dealt with. One way that has been devised to deal with the problem of outliers in robust statistics is by using trimmed means. Other, similar types of procedures (among them, M-estimates, L-estimates, and R-estimates) are more mathematically complicated but follow the same basic logic, so we introduce our reader to trimmed means as a general procedure which is widely employed in robust statistics.

TECHNIQUES USED IN ROBUST STATISTICS

Measures of location and trimmed means

The mean or average of the data is an example of a non-robust estimator. It can be highly influenced by just one outlier in the data. An alternative to the mean is the median score, which is quite robust to outliers. The problem with the median is that it effectively discards all data points except for one or two. We want the estimator we use to reflect what is typical of the data set without being distorted by outliers. The median is not distorted by outliers, but it also does not include much information from the data set.

The trimmed mean represents a compromise between the mean and the median, and between power and bias in the test statistics (Huber 1981). A trimmed mean captures the shape of the data without giving too much weight to outliers by trimming points off the ends of the data set. In theory, any amount could be trimmed, but Wilcox (2001, 2003) on the basis of simulation studies asserts that 20% is a good amount for general use.[7] The way to trim means is to first put the observations in numerical order. To trim by 20%, multiply .2 by the n. Thus, with a data set where $n = 10$, two points would be trimmed off both the lowest and highest end of the data (since $0.2n = 0.2(10) = 2$), resulting in a data set with six scores. This may seem unintuitive—if you have a small data set, you do not want to make it smaller by discarding data! However, robust statistics will in fact result in a more reliable description of the 'average' trend than if all of the data points had been left included.

In fact, one objection that might be raised to trimmed means is that they discard information. It is true that they do, but the idea is that they do not throw away as much information as the median, while being more resistant to outliers than the mean, yet still capturing the general trend of the data.

If the data set is not *exactly* normally distributed, and we would assume that most data sets in our field are not, the trimmed means will be a better reflection of what is typical in the data set.

Because the 20% trimmed means is mathematically quite easy to perform, we would like to note that researchers should not try to use this method without finding statistical programs which can evaluate data using complete arrays of robust techniques. For example, one cannot just plug the 20% trimmed mean into the equation for the sample variance in the same manner as for the untrimmed mean. Removing extreme data points from the set results in interdependence among the remaining points. We will need special equations now to calculate the variance if trimmed means or other robust estimators are used. Proper types of software which calculate trimmed means will ensure that these requirements are met, and can be found for free using the R statistical program and Wilcox's robust commands (recommended books for getting started with robust statistics in R are Crawley, 2007 and Wilcox, 2003).

Bootstrapping

Bootstrapping is another tool in the robust statistical toolbox that can help researchers make more accurate conclusions about their data. Bootstrapping is an approach to statistical inference that makes fewer assumptions about the underlying probability distribution that describes the data than the normal Gaussian distribution does (Efron and Tibshirani 1993). In this type of approach, as Westfall and Young (1993: 12) describe, 'the observed data are used repeatedly, in a computer-intensive simulation analysis, to provide inferences. In simple terms, resampling does with a computer what the experimenter would do in practice, if it were possible: he or she would repeat the experiment.'

It turns out that this process is exactly the same process that was used by the statistician Gosset as an empirical verification of his mathematical derivation of the null distribution of the Student's *t*-test (Student, 1908, as discussed in Wilcox, 2001). Gosset simulated the null distribution by sampling from a normal distribution, calculating the mean and standard deviation of each observation, and finding the resulting *t*-test statistic. Repeating the process over and over, critical values for *t* were then determined. Because Gosset did this without a computer, the process took over a year. Now resampling methods can do the same kinds of simulations in several seconds, except that in bootstrap resampling, the resampling is done from observed data and not from the hypothetical normal distribution.

Using this process bootstrapping generates a distribution (an empirically generated sampling distribution) that can be examined for the significance of the statistics in the same way that the critical value of a *t*-test, based on a normal distribution, can be examined for significance, just as Gosset did.[8] This approach assumes that the empirical distribution function is a reasonable estimate of the unknown, population distribution function. Using the data as an approximation to the population density function, data are re-sampled with replacement[9] from the observed sample to create an empirical sampling distribution for the test statistic under consideration. This resampling is done thousands of times without regard to the original groupings of data, resulting in proxy samples. In the resampling method, the hypothesis testing is done by noting that for the proxy samples, any statistical differences between groups should be due merely to chance. The percentage of statistical tests for the proxy samples which are as large or larger than the observed statistical difference determines the observed *p*-value for the data. Accordingly, when the observed *p*-value is less than 0.05 (or any other threshold we may care to set, but this is the generally accepted level in the field, although there are good arguments for setting it to 0.10, see Kline 2004), we reject the assumed null hypothesis of no difference in the population.

The number of bootstrap samples that should be performed should also be considered when doing a bootstrap. Although it is quite easy to ask for a very high number of samples, work by Wasserman and Bockenholt (1989) shows that in many cases, no more than 1000 bootstrap samples are required to obtain accurate confidence intervals for a location estimate.

An example of how this would work for a *t*-test is that the original *p*-value of the *t*-test done with the original data is compared with the *p*-values of the *t*-tests performed for all of the groups created by replacement sampling. The test statistic (the *t* value in the *t*-test) generated by the proxy samples are then compared with the test statistic generated by the original *t*-test; consequently, '[t]he resampling-based *p*-value is then the proportion of resampled data sets yielding a *t*-statistic as extreme as the original *t*-statistic' (Westfall and Young 1993: 13).[10] This *p*-value is used in the same way as the familiar *p*-value: if it is less than 0.05, the difference between groups is assumed to be statistical.

The reader should be able to see then, that the logic by which the *p*-value is generated in the resampling case is the same logic as that behind the 'classical' parametric tests that are used, but in the resampling case, the empirical cumulative distribution function (whatever that turns out to be, given the data) is used to make inferences about the likelihood of the *p*-value given the data instead of the normal distribution. Indeed, the middle 95% of the ordered means of the bootstrap sample will comprise the 95% confidence interval of the data. The great value of resampling is thus that researchers do not need to assume that the data are normally distributed. Although the bootstrap does not eliminate problems due to skewness (Wilcox 2003: 220), the combination of 20% trimming and the bootstrap does make a practical difference[11] (in some cases smaller confidence intervals for skewed data can be achieved by using a more refined bootstrap method referred to as the 'abc percentile' method, see Efron and Tibshirani 1993). Simulation studies run by Wilcox (reported on in Wilcox 2003: 220) show that with skewed, heavy-tailed distributions, bootstrap methods can reduce Type I error probabilities compared with Student's *t*, although they are still substantially higher than $\alpha = 0.05$. Although larger sample sizes are still desirable because they increase the precision of the confidence interval, the bootstrap is able to function well with symmetrically distributed moderately small samples, such as $n = 10$ (Westfall and Young 1993; Chernick 1999).

Combining bootstrapping with trimmed means has been shown, in a variety of papers by Wilcox and colleagues (Keselman *et al*. 2000, 2003; Wilcox 2001) to further reduce problems with skewed distributions. Problems with skewed distributions are not entirely erased in all cases but robust methods certainly provide a better way of dealing with skewed distributions than using traditional parametric methods.

Now we will illustrate the concept of the bootstrap by using data from a real experiment. The data come from an unpublished study of how accurately various groups of Japanese learners of English produced words beginning with /r/ and /l/ (data available upon request to first author; see Appendix B online for more details about this study). First we will show how the bootstrap operates on one group of data, just to illustrate the idea of bootstrapping. The scores of the group who lived in the US at an early age are given in Table 3, arranged in a numerically ascending order.

The distribution of scores is clearly non-normal and skewed, as shown on the histogram in Figure 4. The mean of the scores is 95.2, but there is one outlier whose score was much lower than this.

A bootstrap sample using 1000 randomly generated samples might include samples like those found in Table 4.

The bootstrapped sampling distribution (called a percentile or uncorrected bootstrap; this is different from the percentile-t bootstrap; see Wilcox 2003 for more information) contains a set of mean scores for the entire group calculated from the 1000 random samples sampled with replacement. In other words, each sample, such as Sample 1, is averaged to give a mean score. Because there were 1000 random samples, 1000 mean scores were thus generated. These new mean scores range from 87.86 to 99.14. As can be seen from the histogram for the bootstrapped sample (Figure 5), the array of mean scores now forms a distribution. This is the empirical distribution by which the mean score of the original data set will be judged.

Note that the bootstrap distribution in Figure 5 was done using all 14 scores in the original sample data. Using the 20% trimmed means would be a way to eliminate the skewing influence of the low score of 72. In the case of $n = 14$,

Table 3: Scores of 'early immersionists' on an accuracy of initial /r/ and /l/ measure

Participants	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
Accuracy	72	90	90	96	96	97	98	98	99	99	99	99	100	100

$0.2n = 2.8$ (when the result is a decimal then round down), so we would eliminate the lowest and highest two scores from the distribution.

Next we will compare the results of a parametric analysis and a robust analysis of the language accuracy task with all groups of participants included. There are three groups of Japanese users of English in this study: (i) the 'early immersionists' lived in the USA as children but returned to Japan by age 7; (ii) the 'late immersionists' lived in the USA as young adults; (iii) the 'non-immersionists' had never lived in an English-speaking country but were majoring in English at their university. Additionally, there was a group of native speakers of English who produced words beginning with /r/ and /l/. The measure being compared here is how accurately what the participants produced aligned with how native speakers of English perceived the initial sound to be (again, more details about the entire study can be found in the online Appendix B; also, specific code used to generate the robust analysis is found in online Appendix A).

Because there are four groups in all, a parametric analysis would use a one-way ANOVA. A one-way ANOVA returns a statistical main effect, $F_{3,55} = 5.27$, $p = 0.003$. Tukey post hoc tests among the groups found that the NS were statistically different from the non-immersionists ($p = 0.002$) but not the late immersionists ($p = 0.407$) or the early immersionists ($p = 0.834$). Using robust statistics, 20% means-trimmed bootstrapped multiple comparisons between the NS and the three non-native groups finds substantially different p-values

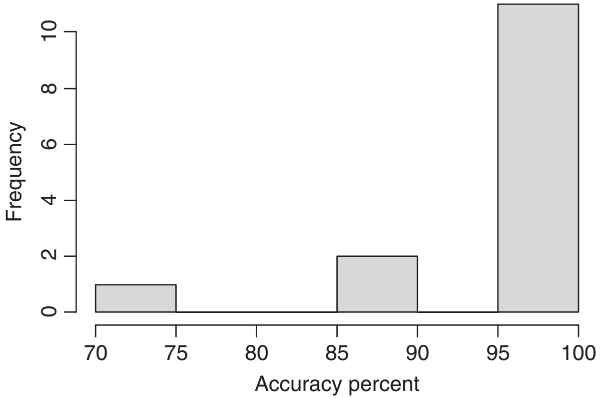

Figure 4: Histogram of original scores on the accuracy measure by early immersionists

Table 4: Two possible bootstrapped samples of the original accuracy measure

Sample 1	72	72	90	98	98	98	99	100	100	100	100	100	100	100
Sample 2	98	98	98	98	98	98	98	99	99	99	99	99	99	100

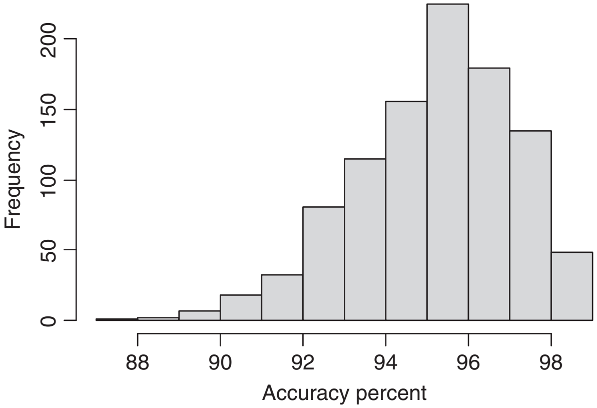

Figure 5: Histogram of bootstrapped mean scores on the accuracy measure by early immersionists

(non, $p = 0.0000$; late, $p = 0.01$; early, $p = 0.01$).[12] Because the sample sizes used were small, robust statistics provide a weight of evidence that with repeated testing, differences would be found between all of the non-native groups and the NS.

At this point, our eager reader may be wondering how to begin using robust statistics in his or her own work. The ideas presented in this article are merely a sampling of the wide variety of robust methods that are available, and for further information we recommend starting by looking at books by Crawley (2007) and Wilcox (2003). Information about the use of the bootstrap in a robust test of bivariate correlation can be found in Wasserman and Bockenholt (1989). Robust tests have been extended to virtually all of the ANOVA methods including repeated measures designs (see Wilcox 2003, 2005). Performing resampling methods is possible using many different statistical programs; MacKinnon *et al.* (2004) remark that resampling methods are available without further modification in AMOS (used in SPSS) and SAS. We note that the terminology of statistics changes very little when robust methods are used. In other words, when you use a robust *t*-test, you will still report the value of the test, a *p*-value associated with it, and confidence intervals and effect sizes as per parametric tests. The only difference is that you will report that you used the 20% trimmed means, or bootstrapping, and name the robust method that was applied, such as Yuen's (1974) method for comparing two independent groups (more information about names can be found in Wilcox, 2003 and 2005).

A NEW PERSPECTIVE ON DATA ANALYSIS

The example given above showed that statistical conclusions about differences between groups can change when robust techniques such as bootstrapping and trimmed means are used. This section will give further examples of how robust statistics can provide a new perspective on data analysis, using examples with language acquisition data. The first example uses data from a study made by the first author (Larson-Hall 2008) of the language abilities of 200 Japanese users of English, some of whom began studying English at a young age, and others who began their study in junior high (see Appendix B online for more information about this study). One of the research questions examined was whether the age that students began studying English affected their scores on an oral phonemic discrimination test and a grammaticality judgement test when total amount of input was factored out. Conventional analysis of covariance (ANCOVA) analysis found that there was no effect of group (earlier or later starters) for the grammaticality judgement test ($F_{1,197} = 1.69$, $p = 0.20$), but the effect of group was statistical for the phonemic discrimination test ($F_{1,197} = 6.55$, $p = 0.01$). The problem with conventional ANCOVA analysis is that it compares groups assuming a linear association, and if the data are not linear, the ANCOVA will generally not be statistical. On the other hand, a robust ANCOVA (we used the ancboot command, found in Wilcox, 2005, p. 529, which uses the 20% trimmed means and bootstrapping, and performs well with heteroscedasticity) does not require a linear association. The ancboot method of ANCOVA compares linear models along a running-interval smooth (similar to the Loess line), finding the tendency of the data instead of forcing it to be along a straight line. This analysis indicates when there are group differences at specific points along the *x*-axis. In the case of the Larson-Hall (2008) data, a robust ANCOVA found a statistical advantage for later starters on the GJT at 800 hours of input, but an advantage for earlier starters at 1833 and 2000 hours of input. For the phonemic discrimination test, a robust ANCOVA found a statistical advantage for the earlier starters at 1300, 1555, 1833 and 2000 hours. The results of the ANCOVA can be more clearly understood by looking at scatterplots with the two groups separated, as in Figure 6. The scatterplots are overlaid with the smooth lines, indicating the trend of the data in Figure 6 at specific hours of input on the *x*-axis.

What the robust statistics do is give a more nuanced picture of the combined influences of age and input on test scores, and in fact provide a way to integrate the results of previous studies which found no beneficial effects for a younger starting age with the results of this study which did find beneficial effects (previous studies were looking only at the very low end of hours of input, where advantages for earlier starters did not appear).

Another example comes from a reanalysis of raw data provided by DeKeyser (2000). DeKeyser gave 57 Hungarian immigrants to the USA a grammaticality judgement test and examined the correlation between their age of arrival (AOA) and their scores. Like many other studies examining the relationship between age of acquisition and ultimate language ability, his data showed a statistical and negative correlation between AOA and scores across the entire range of children and adults ($r = -0.63$). DeKeyser did not focus on this overall

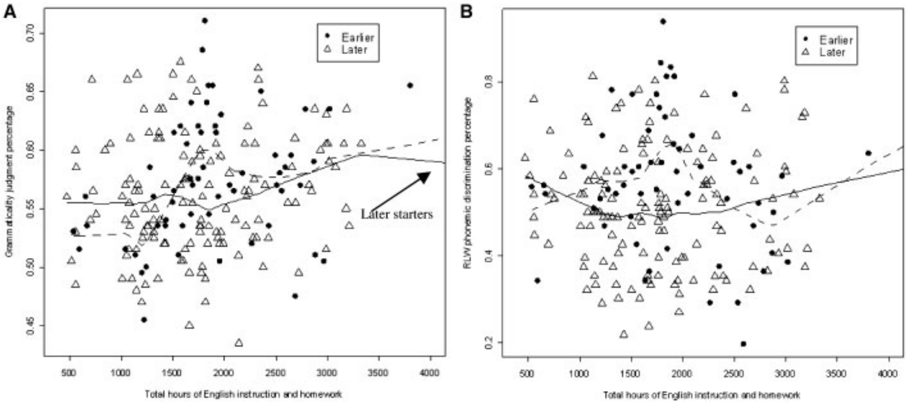

Figure 6: Scatterplots showing results for groups of earlier and later learners of English on a grammaticality judgement test (A) and phonemic discrimination test (B) as a function of hours of input. Smooth lines are calculated for both groups (dashed line for earlier starters and unbroken line for later starters)

score much, as he wanted to show there were different patterns between younger and older arrivals, and he somewhat arbitrarily chose a cut-off point of 15 to split the groups.

Robust estimates of location can result in different but more valid statistical results than classical parametric tests when the researcher is interested in making inference about the majority of the observations in a population. A robust correlation using the cor.plot command from the mvoutlier library in R with DeKeyser's data (using an algorithm for outlier detection; R code for this command is illustrated in online Appendix A) reveals that there is no statistical correlation across the part of the data in the sample that excludes outliers ($r = 0.03$, ns). Figure 7 shows ellipses containing the data used in each of the correlations (the classical correlation and the robust correlation). The figure shows that the robust correlation excludes most of the data from the youngest learners in order to find the data which best represents the overall trend. It can be seen that robust correlation could even provide a principled reason for splitting the data (although at a different point than the one DeKeyser used), and this example shows again that data can be seen in a new light when robust statistics are used.

These two examples serve to show that robust statistics can make a difference in the statistical analyses that are done in the field. We would like SLA researchers to become aware of some of the most important and enduring changes taking place in the field of modern statistics because we feel they can profitably be applied to improve the accuracy and reliability of our own studies.

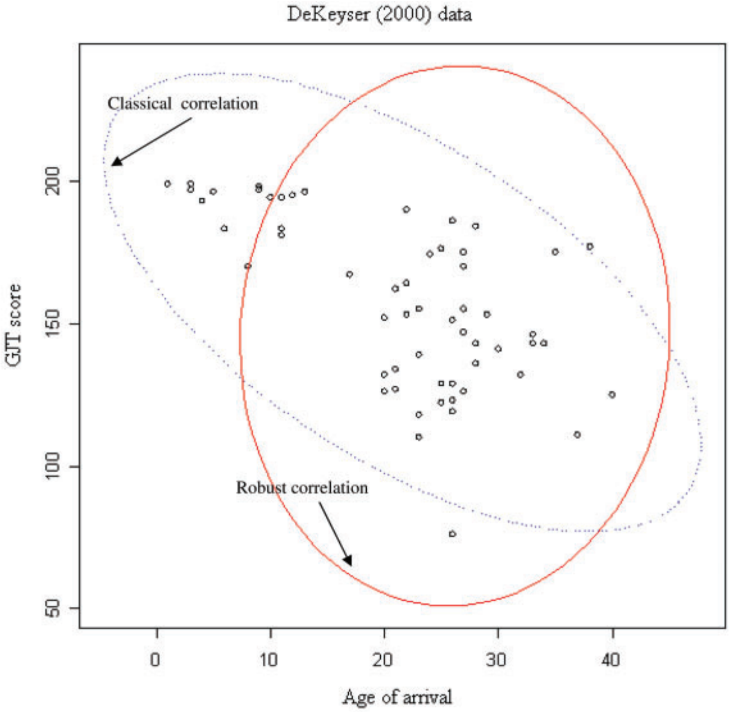

Figure 7: Comparison of data included in classical and robust correlation of the DeKeyser (2000) data

CONCLUSION

Quantitative articles which use statistical methods are the kind of studies most often published in the SLA field (Lazaraton 2000). As such, methods that improve the accuracy of statistical inference should be highly important to the SLA field. Our main purpose in this article is the introduction of effective graphical procedures and robust statistics to researchers in the language acquisition field. Small changes in the way researchers analyze and present data can make large differences in the way research is comprehended.

Work in modern statistics has shown that parametric tests which have been assumed to be robust to slight deviations from normality are in fact not. Robust statistics have been formulated to deal with real, applied data that does not necessarily conform to a normal distribution, and using robust statistics routinely will lead to more power to discover real differences and more accuracy in estimating the statistics involved. Robust statistics also provide objective and replicable ways of dealing with outliers, and deal better with small and non-normally distributed data sets than parametric statistics.

We have also suggested that researchers always include graphics in their research reports (along with raw data, if possible), and that these graphics should be as informative as possible. In practice, we suggested that researchers should use boxplots instead of barplots for group difference data, and add Loess (smooth) lines along with regression lines to scatterplots.

We want to emphasize that modern statistical methods do not solve every problem that may arise when conducting statistical analyses, but they do offer a much better way of approaching quantitative analysis. Although there may not always be one easy and best way to solve every problem using robust methods, years of research in statistics have shown that parametric statistics, which depend on the assumption of normality, are not nearly as accurate as robust statistics in almost every case (Tukey 1960; Yuen and Dixon 1973; Huber 1981; Luh and Guo 2001; Wilcox 2001). If we are to be responsible researchers we need to find out about the advances that have been taking place in the field of statistics for the last 40 years and incorporate these methods, which are much more practical for authentic data sets, into our analyses.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

NOTES

[1] Various statisticians have called for providing raw data (Fienberg *et al.* 1985; Westfall and Young 1993) but as far as we know, this is not required for any journals in the field of SLA. Raw data can help others verify that statistical procedures have been used correctly and that conclusions are based on solid statistical reasoning. An example will serve to illustrate our point. An article by Hirata (2004) erroneously concluded that groups in the study were statistically different. We know this because the author provided the raw data for her eight participants. In reporting on the differences between groups for the perception tasks, Hirata apparently used the significance value (the p-value) from Levene's test for equality of variances, not from the t-test. Hirata reported that the p-value for the difference between the experimental and control group was $p = 0.004$ in the post-test condition, while in fact the p-value was $p = 0.20$, meaning that the groups were *not* statistically different in the post-test.

[2] It should be noted that if the frequency of use of recasts were compared over 10 different teachers, a boxplot might be entirely appropriate to show the range of frequencies. In any case where an average could be computed a boxplot could be used.

[3] The term 'robust statistics' is applied to a whole range of techniques that are meant to make data more robust to violations of assumptions of classical techniques.

4 As recently as 1993, Westfall and Young stated that the major impediment to using robust statistics in applied work was that they were time-consuming with the computational power available. This is no longer the case.

5 The use of the term 'statistical (difference)' here is deliberate. Although some statisticians believe it is more accurate to say 'statistically significant difference' or 'significant difference', we have chosen to follow Kline's (2004) recommendation to return the use of the word 'significant' to its ordinary meaning of 'important' (which is does NOT necessarily mean when it modifies 'statistical') and simply call differences statistical.

6 Definitions for the variables in this equation: t_{df} = the t-value at the given degrees of freedom; X_T = mean of the treatment group; X_C = mean of the control group; $\text{var}_{\text{pooled}}$ = the pooled variance, which is equal to $(n_T - 1)$(standard deviation of the treatment group)$^2 + (n_C - 1)$(standard deviation of the control group)^2 all divided by $n_T + n_C - 2$; n_T = number in the treatment group; n_C = number in the control group.

7 It should be noted that this type of trimming is symmetric trimming, where an equal number of points is removed from both ends of the distribution prior to the computation of an estimate of the location of the distribution. More recently, methods for asymmetric trimming have been proposed (Keselman *et al.* 2007).

8 To remind readers, hypothesis testing using parametric statistics calculates a test statistic using various formulas but mostly involves using the average or average differences, within-group variances of the data set, and then returns a probability (or p-value) for the observed test statistic. This probability indicates the probability with which the same or even more extreme results would be found if the null hypothesis were true (Klein 2004: 63–4). The p-value is the probability of the data given the hypothesis, written ($p(D|H_o)$), not the probability of the hypothesis given the data, written ($p(H_o|D)$) (Nickerson 2000). Hypothesis-testing is a procedure that relies on the theoretical sampling distribution to determine whether the data are probable given the null hypothesis. The theoretical sampling distribution, assumed to be a Gaussian or normal curve, produces the p-values for the null hypothesis.

9 Resampling with replacement means that as each number from the original data set is randomly drawn, it is returned to the original set and may be chosen again in future draws. Each computer-generated sample is the same size as the original data set.

10 Westfall and Young created the *multtest* package in R which performs these types of bootstraps.

11 When resampling is done in combination with bootstrapping, samples are taken from the entire data set, not the trimmed set. Trimming is done on the bootstrap sample that is generated from all of the data.

12 Although by a rubric of $p < 0.05$ all of these values are statistical, because there are multiple comparisons, an adjustment is made that sets the cutoff p-value lower, to $p = 0.009$ in fact. By this cut-off value, the differences between the late and early groups with the NS are still not statistical, although of course at $p = 0.01$ they are much closer to that point to be statistical than in the non-robust version.

REFERENCES

Chernick, M. 1999. *Bootstrap Methods: A Practitioner's Guide*. Wiley-Interscience.

Crawley, M. J. 2007. *The R Book.* Wiley.

DeKeyser, R. M. 2000. 'The robustness of critical period effects in second language acquisition,' *Studies in Second Language Acquisition* 22: 499–533.

Efron, B. and **R. J. Tibshirani.** 1993. *An Introduction to the Bootstrap.* Chapman & Hall.

Everitt, B. and **G. Dunn.** 2001. *Applied Multivariate Data Analysis.* 2nd edn. Hodder Arnold.

Fienberg, S. E., M. E. Martin, and **M. L. Straf.** 1985. *Sharing Research Data.* National Academy Press.

Hampel, F. R. 1968. *Contributions to the Theory of Robust Estimation*. Unpublished PhD thesis, University of California, Berkeley.

Hampel, F. R. 1973. 'Robust estimation: A condensed partial survey,' *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* 27: 87–104.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and **W. A. Stahel.** 1986. *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

Hastie, T. J. and **R. J. Tibshirani.** 1990. *Generalized Additive Models.* Chapman and Hall.

Hirata, Y. 2004. 'Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts,' *Computer Assisted Language Learning* 17/3–4: 357–76.

Huber, P. J. 1964. 'A robust version of the probability ratio test,' *Annuals of Mathematical Statistics* 36/6: 1753–8.

Huber, P. J. 1981. *Robust Statistics.* John Wiley & Sons.

Keselman, H. J., J. Algina, R. Wilcox, and **R. K. Kowalchuk.** 2000. 'Testing repeated measures hypotheses when covariance matrices are heterogeneous: revisiting the robustness of the Welch-James test again,' *Educational and Psychological Measurement* 60: 925–38.

Keselman, H. J., R. R. Wilcox, and **L. M. Lix.** 2003. 'A generally robust approach to hypothesis testing in independent and correlated groups designs,' *Psychophysiology* 40: 586–96.

Keselman, H. J., R. Wilcox, L. M. Lix, J. Algina, and **K. Fradette.** 2007. 'Adaptive robust estimation and testing,' *British Journal of Mathematical and Statistical Psychology* 60: 267–93.

Klein, R. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* American Psychological Association.

Kline, R. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* American Psychological Association.

Larson-Hall, J. 2008. 'Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation,' *Second Language Research* 24/1: 35–63.

Lazaraton, A. 2000. 'Current trends in research methodology and statistics in applied linguistics,' *TESOL Quarterly* 34/1: 175–81.

Luh, W.-M. and **J.-H. Guo.** 2001. 'Using Johnson's transformation and robust estimators with heteroscedastic test statistics: an examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design,' *British Journal of Mathematical and Statistical Psychology* 54: 79–94.

MacKinnon, D. P., C. M. Lockwood, and **J. Williams.** 2004. 'Confidence limits for the indirect effect: Distribution of the product and resampling methods,' *Multivariate Behavioral Research* 39/1: 99–128.

Maronna, R. A., R. D. Martin, and **V. J. Yohai.** 2006. *Robust Statistics: Theory and Methods.* Wiley.

McGill, R., J. W. Tukey, and **W. A. Larsen.** 1978. 'Variations of box plots,' *The American Statistician* 32/1: 12–16.

Micceri, T. 1989. 'The unicorn, the normal curve, and other improbable creatures,' *Psychological Bulletin* 105/1: 156–66.

Nickerson, R. S. 2000. 'Null hypothesis significance testing: a review of an old and continuing controversy,' *Psychological Methods* 5/2: 241–301.

Pallant, J. 2001. *SPSS Survival Manual.* Open University Press.

Porte, G. K. 2002. *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research*. John Benjamins.

Student. 1908. 'The probable error of a mean,' *Biometrika* 6/1: 1–25.

Tufte, E. R. 2001. *The Visual Display of Quantitative Information.* 2nd edn. Graphics Press.

Tukey, J. W. 1960. 'A survey of sampling from contaminated distributions' in I. Olkin, S. G. Ghwyne, W. Hoeffding, W. G. Madow, and H. B. Mann (eds): *Contributions to Probability and Statistics: Essays in Honour of Harold Hotelling*. Stanford University Press, pp. 448–485.

Tukey, J. W. 1962. 'The future of data analysis,' *The Annals of Mathematical Statistics* 33: 1–67.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Addison-Wesley.

Wasserman, S. and **U. Bockenholt.** 1989. 'Bootstrapping: applications to psychophysiology,' *Psychophysiology* 26/2: 208–21.

Weinberg, S. L. and **S. K. Abramowitz.** 2002. *Data Analysis for the Behavioral Sciences Using SPSS.* Cambridge University Press.

Westfall, P. H. and **S. S. Young.** 1993. *Resampling Based Multiple Testing.* Wiley.

Wilcox, R. 1995. 'ANOVA: A paradigm for low power and misleading measures of effect size?,' *Review of Educational Research* 65/1: 51–77.

Wilcox, R. 1998. 'How many discoveries have been lost by ignoring modern statistical methods?,' *American Psychologist* 53/3: 300–14

Wilcox, R. 2001. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. Springer.

Wilcox, R. 2003. *Applying Contemporary Statistical Techniques.* Elsevier Science.

Wilcox, R. 2005. Introduction to robust estimation and hypothesis testing. 2nd edn. Elsevier Science.

Wilkinson, L. and **Task Force on Statistical Inference, A. P. A., Science Directorate, Washington, DC, US.** 1999. 'Statistical methods in psychological journals: guidelines and explanations,' *American Psychologist* 54/8: 594–604.

Yuen, K. K. 1974. 'The two-sample trimmed t for unequal population variances,' *Biometrika* 61: 165–70.

Yuen, K. K. and **W. J. Dixon.** 1973. 'The approximate behaviour and performance of the two-sample trimmed *t*.' *Biometrika* 60/2: 369–7.

Applied Linguistics: 31/3: 391–417 © Oxford University Press 2009
doi:10.1093/applin/amp040 Advance Access published on 24 November 2009

Toward Automated Multi-trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores

[1]YONG-WON LEE, [2]CLAUDIA GENTILE, and [3]ROBERT KANTOR

[1]Seoul National University, [2]Mathematica Policy Research, and [3]Educational Testing Service

The main purpose of the study was to investigate the distinctness and reliability of analytic (or multi-trait) rating dimensions and their relationships to holistic scores and *e-rater*® essay feature variables in the context of the TOEFL® computer-based test (TOEFL CBT) writing assessment. Data analyzed in the study were holistic and multi-trait essay scores provided by human raters and essay feature variable scores computed by *e-rater*® (version 2.0) for two TOEFL CBT writing prompts. It was found that (i) all of the six multi-trait scores were not only correlated among themselves but also correlated with the holistic score, (ii) high correlations obtained among holistic and multi-trait scores were largely attributable to the impact of essay length on both holistic and multi-trait scoring, and (iii) some strong associations were confirmed between several *e-rater* variables and multi-trait rating dimensions. Implications are discussed for improving the multi-trait scoring of essays, refining *e-rater* essay feature variables, and validating automated essay scores.

INTRODUCTION

Holistic (i.e. global or impressionistic) scoring has been widely used in many large-scale writing assessments including the computer-based Test of English as a Foreign Language™ (TOEFL®), Graduate Record Examination® (GRE®), and Graduate Management Admission Test® (GMAT®) (Williamson and Huot 1993). For holistic scoring rubrics, elaborate score descriptors are usually developed for several score levels, and the writing qualities of an essay are usually represented by a single, overall 'holistic' rating. One drawback of holistic scoring has to do with its inability to capture examinees' specific weaknesses and strengths in writing (Weigle 2002). This failure can be especially true for second language learners who are still developing their writing skills and who are thus likely to show uneven profiles across different aspects of writing. For examinees with such non-uniform patterns of proficiencies across different aspects of writing skills, analytic (or multi-trait) scoring rubrics can be useful in capturing their weaknesses and strengths (Raimes 1990; Hamp-Lyons 1991, 1995; Connor-Linton 1995; Sasaki and Hirose 1999; Bacha 2001).[1] For this reason, many educators believe that multi-trait scoring can be useful for generating diagnostic feedback to inform instruction and improve learning (Hamp-Lyons 1991, 1995; Roid 1994; Swartz *et al*. 1999).

Despite such advantages, multi-trait scoring has not been widely used for large-scale writing assessments for several important reasons. One has to do with the cost associated with human rating of essays (Veal and Hudson 1983; Huot 1990). Even when holistic scoring is used, the scoring of writing samples poses a cost challenge for testing programs, compared with machine-scored multiple-choice items. Because multi-trait scoring requires multiple ratings of each essay by human raters, the number of raters and time required for rater training and scoring is much greater for multi-trait than for holistic scoring. In addition, multi-trait ratings have often proven less useful than expected because rating dimensions are often highly correlated among themselves and with holistic scores, thus rendering them redundant from a psychometric point of view (Veal and Hudson 1983; Freedman 1984; Huot 1990; Bacha 2001).[2]

Recently, however, multi-trait scoring has received renewed attention in writing assessment, particularly in the context of automated essay scoring and evaluation. There are several automated essay scoring (AES) systems that are currently in use for large-scale assessment programs (interested readers, see Kukich 2000, Shermis and Burstein 2003, and Dikli 2006), although the focus of the current investigation is on a recent version of *e-rater*® (electronic rater) that was developed by Educational Testing Service® (ETS®). One exciting implication of such technology is that the large rating cost traditionally associated with multi-trait scoring can be reduced significantly if valid multi-trait scores can be computed automatically. Besides, if computer-generated holistic and multi-trait scores can also be traced back to more micro-level essay text features, these features can be used to provide performance feedback to learners.

Another interesting area of application for such technology has to do with online writing practice services, such as *Criterion*^{SM} and *My Access!*®, that are becoming popular nowadays. E-*rater*, for instance, has been embedded in an Internet-based online writing practice service, *Criterion*^{SM}, to score essays written and submitted by students or prospective test-takers (http://www.ets.org/criterion).[3] Recent versions of *Criterion* include two main automated evaluation components: (i) *e-rater* and (ii) *Critique*^{TM} Writing Analysis Tools (*Critique* henceforth). E-*rater* provides an instant holistic score for an essay, while *Critique* flags the parts of essays that are suspected of containing various grammar, usage, mechanics, and style (GUMS) errors in the essay. These GUMs errors are aggregated and transformed into four accuracy ratio variables, which are then used as scoring variables together with eight other automated text variables in recent versions of *e-rater* (version 2.0 and above). A brief description of 12 *e-rater* scoring variables and various types of errors detected by *Critique* can be found in Attali and Burstein (2006) and Quinlan, Higgins, and Wolff (2009), respectively.[4]

Despite the connection between *Critique* errors and four of the text feature variables of *e-rater*, there seems to be an apparent need for strengthening links further between the automated score (provided by *e-rater*) and diagnostic feedback (provided by *Critique* and *Criterion*) in the current system. If such links are more strongly established, the feedback can become more useful to learners and the validity of automated scores can be further established. In this study, we take special note of the potential values of automated multi-trait scores as linking pins to further strengthen the connections between the automated holistic score and the automated diagnostic feedback. Moreover, given the dependence of the computational relationship between the GUMS features and *e-rater* scoring variables, it is also critically important to examine and enhance the construct-relevance and practical usefulness of the computable text features employed by *e-rater* and *Critique*.

With these as a backdrop, we attempt in this study to move beyond statistical emulation of human holistic scores and explore the use of automated scoring (via *e-rater*) for generating multiple trait scores and performance feedback for the learners, in addition to the holistic score. In this new AES framework, it is envisioned that the descriptive performance feedback will be closely aligned with the automated multi-trait scores, which can also be linked to the automated holistic (or composite) score. Exploring the generation of useful automated trait scores in *e-rater*, however, requires the availability of valid human-assigned trait scores to use as criteria or targets for *e-rater*. These human multi-trait scores, once they are obtained, can be used to examine if various micro text feature variables used in *e-rater* can be clustered or re-organized in such a meaningful way that they form a basis for computing automated trait scores (i.e. scores on organization, vocabulary, language use, mechanics, etc.) and composite scores.

The main purposes of the study described here, therefore, were (i) to investigate, in the context of the TOEFL computer-based test (CBT) writing assessment, whether distinct (separable) and reliable (dependable) multi-trait rating dimensions can be identified for human rating and (ii) examine the relationships of the human-assigned, multi-trait scores not only to human-assigned, holistic scores but also to *e-rater* essay feature variables. More specifically, an attempt is made in this study to evaluate a multi-trait scoring rubric developed for the TOEFL CBT Writing section and examine the nature of *e-rater* automated essay feature variables in relation to multi-trait rating dimensions. This will help us to examine not only the usefulness of multi-trait scoring in generating performance feedback but also the possibility of refining or reorganizing the *e-rater* essay feature variables for automated trait scoring.

Multi-trait scoring and diagnostic feedback in writing assessment

In analytic (or multi-trait) scoring, writing samples are rated on several important aspects of writing quality, rather than being assigned a single overall rating (Weigle 2002). From the perspectives of score users, one important reason for favoring the multi-trait scoring method is its usefulness in capturing ESL learners' weaknesses and strengths in writing and generating diagnostic feedback to guide instruction and learning (Hamp-Lyons 1995). Another important reason for pursuing multi-trait scoring has something to do with raters' decision-making processes. In investigating the reactions of raters to ESL students' essays, Santos (1988) and Cumming *et al.* (2002) found that raters were able to judge two aspects of students' writing independently: (i) rhetoric and ideas (or content) and (ii) language. This distinction, consistently evident in the raters' thinking processes while evaluating the essays, suggests that analytic features or multiple traits (rather than a single holistic scale) are inherent aspects of skilled assessors' approach to essay evaluation.

One of the best-known multi-trait rubrics in ESL is one developed by Jacobs and her colleagues (Jacobs *et al.* 1981). In their rubric, essays are rated on five different rating dimensions of writing quality, each having a different weight: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). Two additional examples of multi-trait scales are the Test in English for Educational Purposes (TEEP; Weir 1990) and the Michigan Writing Assessment Scoring Guide (Hamp-Lyons 1991). The TEEP framework consists of seven 4-point scales that cover four aspects of communicative effectiveness (relevance and adequacy of content, compositional organization, cohesion, and adequacy of vocabulary for purpose) and three accuracy dimensions (grammar, mechanical accuracy/punctuation, and mechanical accuracy/spelling). In contrast, the Michigan Writing Assessment framework contains three 6-point scales: ideas and arguments, rhetorical features, and language control. (See Weigle 2002 for more details about these three multi-trait rubrics.)

Recently, Gentile and her colleagues (2002) used in their research study a six-trait, multi-trait rating rubric developed by a panel of ESL writing experts for TOEFL CBT writing assessments. The six separate rating scales for the rubric cover five major multi-trait rating dimensions including development, organization, vocabulary, sentence variety/construction, grammar/usage accuracy, and mechanics. In a sense, this framework is similar to the Jacobs *et al.* (1981) five-dimensional rating scheme. One noteworthy difference, however, is that the language use dimension is further divided into two sub-dimensions of 'sentence variety/construction' and 'grammar/usage accuracy' in the Gentile *et al.* (2002) analytic framework.

Some researchers argue that the intent of holistic scoring is to focus raters' attention on the strengths of writing, not on its deficiencies. Jarvis and others (2003) have pointed out that the ESL learners can compensate for potential deficiencies in their writing by capitalizing on a few of their strengths. Another important point is that different raters can also assign the same holistic score by using somewhat different rating criteria (or weighting the same criteria somewhat differently). All of these factors can potentially complicate the interpretation of holistic scores. In this respect, multi-trait scoring rubrics are generally known to provide more useful diagnostic feedback about examinees' writing skills (Jacobs *et al.* 1981; Hamp-Lyons 1991; Bacha 2001; Kondo-Brown 2002; Weigle 2002). This can be particularly true for second language learners who may have uneven profiles of performance across different aspects of writing (Weigle 2002).

Holistic and multi-trait rating dimensions and *e-rater* essay feature variables

Since the goal of this project was to explore the possibility of using the automated essay feature variables computed and employed by *e-rater* for the purposes of computing multi-trait essay scores and generating performance feedback for ESL learners, we examined conceptual relationships among the rating criteria in the human holistic scoring rubric for TOEFL CBT Writing, the rating dimensions for Gentile *et al.*'s (2002) multi-trait scoring rubric used in this study, and the automated text feature variables used in *e-rater*. The conceptual links between these rating criteria/dimensions and the *e-rater* scoring variables are illustrated in Figure 1.

Since the six rating dimensions for the multi-trait rubric were determined partly based on ESL writing experts' thorough analysis of the holistic scoring rubric for TOEFL CBT, it was not difficult to confirm some conceptual links between the rating criteria in the holistic rubric and the rating dimensions in the multi-trait rubric. For instance, notice in Figure 1 that the verbal descriptors related to task fulfillment (e.g. 'effectively address the writing task'),

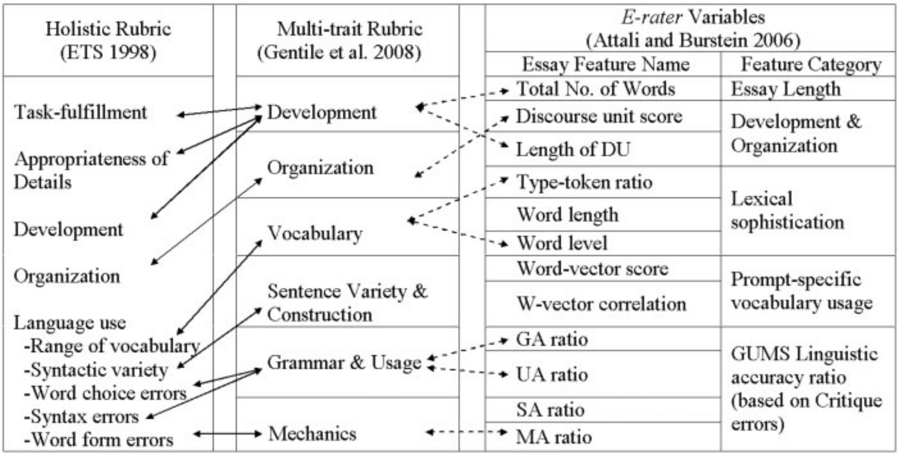

Figure 1: Conceptual links between rating dimensions/criteria in holistic and multi-trait rubrics and 12 e-rater essay feature variables

development (e.g. 'well-developed'), and appropriateness of supporting details (e.g. 'inappropriate or insufficient details') used in the holistic rating rubric for TOEFL CBT Writing can roughly be linked to the development dimension in the TOEFL multi-trait rubric. In a similar fashion, the statement about organization (e.g. 'well-organized') in the holistic rubric can be linked to the organization dimension in the multi-trait rubric. (Please see, ETS 1998 and Lee *et al.* 2008, for the verbal descriptors of the holistic and the multi-trait rubrics for TOEFL CBT Writing, respectively.)

Notice also the verbal descriptors in the holistic rubric related to the two major aspects of language use: variety and accuracy. The variety aspect of language use was represented by such phrases as 'range of vocabulary' and 'syntactic variety' in the holistic rubric, which in turn can be linked to the vocabulary and the sentence variety/construction dimensions in the multi-trait rubric, respectively. In contrast, the accuracy aspect of language use in the holistic rubric was indicated by such descriptors as 'errors in sentence structure and usage', 'word choice errors', and 'word form errors.' The multi-trait rubric, however, used a separate, overarching dimension of 'mechanics' to capture various mechanics-related errors in the essay. All other non-mechanics errors were used by the analytic raters when they scored the essay on the grammar/usage accuracy dimension.

Also displayed in Figure 1 are the conceptual relationships between the six multi-trait rating dimensions for the multi-trait rubric and the 12 essay feature variables used in *e-rater* version 2.0. The conceptual links shown in Figure 1 were tentatively created based on the inspection of the definitions of the *e-rater* variables and have yet to be empirically examined in this study. The 'length of discourse unit' and 'discourse unit score' variables, for instance, are claimed to tap into the surface levels of developmental and organizational qualities of essays (Attali and Burstein 2006). However, it was not easy to link the two prompt-specific vocabulary usage variables (word vector score and correlation) only to one particular multi-trait rating dimension, because these two variables were designed to evaluate content word usage in a particular essay in reference to the holistic score points. Moreover, there does not seem to be any *e-rater* variables that can be linked directly to the sentence variety/construction dimension in the multi-trait rubric.

Research questions

One important concern raised by researchers in conjunction with multi-trait scoring of essays is that some raters can unconsciously fall back on holistic methods while doing multi-trait scoring (Bacha 2001; Weigle 2002). For this reason, it is important to provide enough scoring guidelines to the raters for each of the rating dimensions. Moreover, in investigating the distinctiveness of multi-trait rating dimensions, it is also important to find and use appropriate, advanced statistical methods that would allow for more in-depth and rigorous analysis of the multi-trait scores than simple comparison of correlations.

More specifically, the current program of research was conducted with the following four questions in mind:

(i) How reliable are the multi-trait ratings assigned by human raters?
(ii) What are the relationships among the various multi-trait scores obtained for this study?
(iii) What are the relationships between the holistic essay score and the multi-trait scores?
(iv) What essay feature variables for a recent version of *e-rater* are most closely related to each of the six different multi-trait rating dimensions used in this study?

METHODS

Data source

Data analyzed included the scores assigned to 930 essays written for two TOEFL CBT writing prompts: one holistic essay score (an average of two raters' scores), six multi-trait essay scores (an average of two raters' scores each), and twelve *e-rater* (version 2.0) essay feature variable scores (assigned by the computing program). One prompt asked examinees to state their opinion regarding the importance of students studying 'history/literature versus science/mathematics' (or the other way around), whereas the other prompt asked examinees to discuss the advantages and disadvantages of 'practicing sports.' The writing section of TOEFL CBT consists of a single essay prompt that is selected for each examinee from a pool of prompts. In this study, half of the examinees ($n = 465$) took one prompt, and another half took the other prompt.

Of the 930 test-takers, 491 were males, 426 were females, and 13 examinees were of unidentified gender. At the time of the testing, the examinees also took the whole battery of TOEFL CBT including the writing section, and their computer-based TOEFL scores ranged from 0 to 300 (maximum possible score 300), with a mean of 198 and a standard deviation of 60. The participants were from 58 diverse language backgrounds, with the seven largest native language groups being Japanese (24%), Korean (13%), Spanish (7%), Chinese (6%), Arabic (4%), German (3%), and French (2%).

For each of these two prompts, a sample of 465 essays was selected from a larger pool of essays and used for multi-trait scoring as well as for textual analysis by a recent version of *e-rater* (version 2.0). To create this sample, two separate, smaller data sets were combined and used for multi-trait rating for each prompt: (i) a sample of 265 essays systematically selected to cover

Table 1: Means and standard deviations for the holistic and multi-trait essay scores used in the study

Essay scores	Score range	Prompt 1		Prompt 2	
		Mean	SD	Mean	SD
Holistic	1–6	3.6	1.4	3.6	1.4
Multi-trait					
DEV	1–6	3.6	1.2	3.6	1.1
ORG	1–6	3.8	1.2	3.9	1.1
VOC	1–5	3.1	1.2	3.0	1.2
SVC	1–5	3.1	1.3	3.1	1.3
GU	1–5	2.8	1.2	2.8	1.2
MEC	1–5	3.1	1.2	3.3	1.2

Note: $n = 465$. DEV=development, ORG=organization, VOC = vocabulary, SVC = sentence variety/construction, GU = grammar/usage, and MEC = mechanics.

a full score range and (ii) a sample of 200 essays randomly selected from the larger pool of essays.[5]

Table 1 shows the means and standard deviations of the holistic and multi-trait scores used in this study. The holistic scores were obtained from the operational TOEFL CBT data, and were based on two independent readings and holistic ratings of the essay response on a 1–6 scale. In most of the analyses, the average of the two independent ratings was used, which ranged from 1 to 6 with possible scores in intervals of 0.5. The holistic score used in this study was the average of the first two ratings before adjudication (see the *Computer-based TOEFL Score User Guide*, ETS 1998, for more details about the rating rubric). In contrast, multi-trait scores were obtained for each essay by re-rating the essays using a slightly modified version of Gentile *et al.*'s (2002) multi-trait rating rubric. The six different multi-trait rating dimensions included development (DEV), organization (ORG), vocabulary (VOC), sentence variety/construction (SVC), grammar/usage (GU), and mechanics (MEC). The development and organization scores were on scales of 1–6, whereas the rest of the multi-trait scores were on scales of 1–5. Each of these six multi-trait scores used in this study was the average of two independent ratings for each essay, with possible scores in intervals of 0.5.

Table 2 shows the means and standard deviations of the 12 *e-rater* essay feature variables used in this study. These 12 essay feature variable scores were computed by a recent version of *e-rater* (version 2.0) for each of the 930 essays used in this study. Please note that mathematically transformed values were also computed for the four accuracy ratio variables and are reported in parentheses in Table 2. Since some of these ratio variables often turn out to have extremely small variances, these variables are usually mathematically converted to more statistically stable values (by way of

Table 2: Means and standard deviations for the twelve e-rater variables used in the study (n = 465)

Variable name	Prompt 1		Prompt 2	
	M	SD	*M*	SD
1. Discourse unit score	−3.40	2.42	−3.34	2.24
2. Length of discourse unit	42.50	24.20	39.80	18.90
3. Type/token ratio	0.36	0.11	0.35	0.10
4. Word length	4.70	0.33	4.60	0.29
5. Vocabulary level	52.70	6.60	56.00	5.90
6. Word–vector score	4.82	1.18	4.30	1.29
7. Word–vector correlation	0.19	0.06	0.19	0.07
8. Grammatical accuracy ratio (log)	0.99	0.01	0.99	0.01
	(4.56)	(0.80)	(4.52)	(0.81)
9. Usage accuracy ratio (log)	1.00	0.00	1.00	0.000
	(4.91)	(0.69)	(4.95)	(0.67)
10. Mechanical accuracy ratio (log)	0.96	0.04	0.96	0.04
	(2.58)	(1.29)	(2.29)	(1.07)
11. Stylistic accuracy ratio (log)	0.88	0.11	0.86	0.10
	(3.34)	(0.83)	(3.43)	(0.83)
12. Total number of words	207.70	103.50	214.50	105.50

logarithmic transformation). Such log-transformed values are used in more recent versions of *e-rater*, instead of the original ratio variables.

Rating procedures

Multi-trait rating scales developed for the study by Gentile *et al.* (2002) were modified and used for this study. In the study by Gentile *et al.* (2002), a panel of three ESL writing experts identified six rating dimensions as central to effective essay writing, based on a careful examination and analysis of the holistic scoring rubrics for TOEFL CBT and TWE® (the Test of Written English™) (ETS 1998), results of the pilot study, and examinee essay samples, as explained in the previous section. Each of these six dimensions is also described in detail in Lee *et al.* (2008). The same rating rubric and designs were used for the two prompts.

A total of 15 raters were recruited from two different pools of ESL teaching practitioners: (i) participants in the English language assessment summer institute on item writing held at ETS in the summer of 2003 and (ii) trained online essay raters for TOEFL CBT. Two separate, full-day training sessions were conducted, one for development and organization dimensions and the other for the remaining four dimensions (vocabulary, sentence variety/construction, grammar/usage, and mechanics). All of the essays were double-rated by two independent raters on each of the multi-trait rating dimensions. For actual scoring of essays, online scoring kits were prepared so that raters could rate the essays on a computer at a place they chose.

Data analysis

Several statistical methods were used to analyze the holistic, multi-trait, and *e-rater* text feature variable scores obtained for the TOEFL essays used in this study. These analyses included: (i) reliability, (ii) correlation, and (iii) multidimensional scaling analyses. More detailed descriptions of each of these analyses follow:

Reliability analyses

Reliability analyses were conducted to investigate the reliability of the holistic and multi-trait scores obtained for the TOEFL essays. The computer programs GENOVA (Crick and Brennan 1984) and mGENOVA (Brennan 1999) were used to compute the reliability coefficients for the holistic and multi-trait scores and the composite of the multi-trait scores. The obtained reliability coefficients for the two prompts were plotted together and compared.

Correlation analyses

Correlation matrices for the holistic and multi-trait scores were obtained to examine the relationships among these scores for the two prompts. Two different types of correlations were computed: (i) Pearson product–moment (zero-order) correlations among holistic and multi-trait scores and (ii) partial correlations computed after partialling out the effect of essay length on both the holistic and multi-trait scores.[6] In addition, both zero-order and partial correlations were also obtained between the human-assigned essay scores (i.e. holistic and multi-trait scores) and the *e-rater* text feature scores to examine the relationships between the holistic and multi-trait rating dimensions and the automated text features used in *e-rater*. The partial correlations were examined to see how much unique contribution each of the regular *e-rater* variables could make in predicting each of the multi-trait scores, independently of essay length.

Multidimensional scaling (MDS) analyses

MDS analyses were conducted to obtain a graphical representation of the structural relationships among the holistic and multi-trait scoring dimensions. Simply speaking, MDS analysis is a series of related multivariate, statistical techniques used in data visualization for exploring similarities or dissimilarities among items, entities, or objects (in our case among rating dimensions) (Borg and Groenen 1997).

In this study, the MDS analysis was conducted in the following three steps: (i) estimate intercorrelations among seven scoring dimensions (including holistic and multi-trait rating dimensions), (ii) compute distance measures by using the obtained inter-correlations among the scoring dimensions as similarity (or dissimilarity) values, and (iii) finally assign a location of each dimension in a one, two, and three-dimensional space suitable for graphing. A separate MDS analysis was conducted for each of the prompts using the computer program SPSS Version 12 (SPSS 2003). A more detailed, technical description of the MDS analysis procedure used for the study can be found in Lee *et al.* (2008).

RESULTS

Score reliability of multi-trait and holistic scores

Figure 2 displays the score reliability coefficients for each of the six multi-trait scores and for the composite score estimated for a double-rating scheme for Prompts 1 and 2.[7] Similar results were obtained for both prompts. The reliability indices ranged from 0.81 to 0.91 across the six dimensions for the two prompts. Higher score reliability estimates were obtained for the vocabulary, development, and sentence variety/construction dimensions (0.85–0.91) than for the organization, grammar/usage, and mechanics dimensions (0.81–0.86). Overall, acceptable levels of score reliabilities (>0.80) were achieved for both prompts across all of the six rating dimensions.

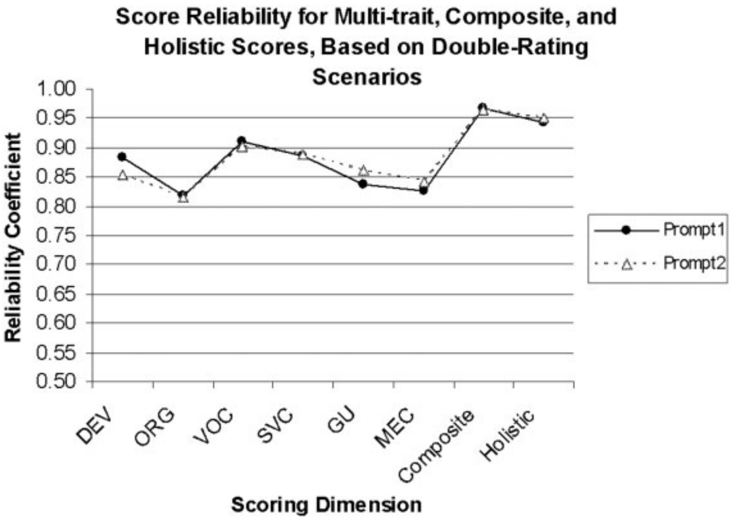

Figure 2: Reliability of multi-trait and composite scores based on double ratings for each of the two TOEFL CBT prompts

In addition, the score reliability coefficients for the holistic score are also shown in Figure 2. The reliabilities for the holistic score for a double-rating scheme were very high for the two prompts (0.94, 0.95), although they were slightly lower than those for the composite of the multi-trait scores for a double rating scheme (0.97, 0.96).

Correlation among multi-trait and holistic scores

Table 3 shows the zero-order and partial correlations among the holistic score, the six multi-trait scores, and the essay length variable (TNW—total number of words) for Prompts 1 and 2, respectively. In each of these two panels, the elements below the diagonal represent the zero-order correlations, while those above the diagonal (italicized) represent the partial correlations.

Table 3: Zero-order and partial correlations between holistic and multi-trait scores

	Holistic	Multi-trait					
		DEV	ORG	VOC	SVC	GU	MEC
Panel a: Prompt 1							
Holistic	**1.00**	*0.35*	*0.35*	*0.50*	*0.49*	*0.47*	*0.40*
Multi-trait							
DEV	0.88	**1.00**	*0.39*	*0.35*	*0.29*	*0.29*	*0.24*
ORG	0.85	0.84	**1.00**	*0.25*	*0.22*	*0.25*	*0.24*
VOC	0.90	0.85	0.81	**1.00**	*0.50*	*0.49*	*0.34*
SVC	0.87	0.81	0.78	0.87	**1.00**	*0.67*	*0.39*
GU	0.83	0.77	0.74	0.83	0.88	**1.00**	*0.45*
MEC	0.72	0.66	0.66	0.70	0.72	0.73	**1.00**
TNW	0.89	0.88	0.80	0.84	0.79	0.75	0.60
Panel b: Prompt 2							
Holistic	**1.00**	*0.35*	*0.24*	*0.44*	*0.48*	*0.55*	*0.28*
Multi-trait							
DEV	0.88	**1.00**	*0.25*	*0.25*	*0.26*	*0.29*	*0.27*
ORG	0.83	0.81	**1.00**	*0.19*	*0.18*	*0.15*$^{+}$	*0.20*
VOC	0.88	0.82	0.78	**1.00**	*0.38*	*0.45*	*0.26*
SVC	0.87	0.81	0.77	0.83	**1.00**	*0.69*	*0.26*
GU	0.85	0.77	0.71	0.81	0.89	**1.00**	*0.37*
MEC	0.75	0.73	0.71	0.72	0.72	0.73	**1.00**
TNW	0.90	0.88	0.80	0.82	0.80	0.74	0.67

Note: $n = 465$. Bold-faced numbers indicate the diagonal. Elements below the diagonal are original zero-order correlations. Elements above the diagonal (italicized) are partial correlations. All of the correlation coefficients were statistically significant at the 0.05 level (two-tailed) and all except one ($^{+}$) also are significant at the 0.01 level.

In terms of the zero-order correlations, we found that the six multi-trait scores were correlated significantly among themselves. The highest correlation was observed for the language use sub-dimension pair of sentence variety/construction and grammar/usage for both prompts (0.88, 0.89). The lowest correlations were obtained for the mechanics and development pair for the first prompt (0.66) and for the mechanics and organization pair for the second prompt (0.71).

Each of the six multi-trait scores was also correlated significantly with the holistic score for both prompts. Of the six multi-trait scores, the vocabulary, development, and sentence construction scores were most strongly correlated with the holistic scores (0.87–0.90). The organization and grammar/usage dimensions were also highly correlated with the holistic scores (0.83–0.85). The lowest correlation was observed for the mechanics dimension (0.72, 0.75).

One intriguing result was that both multi-trait and holistic scores were significantly correlated with the essay length variable (measured by TNW in an essay). Above all, the holistic score was more highly correlated with essay length (0.89–0.90) than any of the six multi-trait scores (0.60–0.88). When only the six multi-trait scores were compared, the first four dimensions (development, organization, vocabulary, and sentence variety/construction) were more sensitive to essay length than were the last two accuracy-related dimensions (grammar/usage and mechanics). More specifically, the development score was most strongly correlated with the essay length variable (0.88), while the mechanics score was most weakly correlated (0.60–0.67).

Further partial correlation analysis revealed that, after the impact of essay length was controlled, correlations among multi-trait scores were much lower but still statistically significant (0.15–0.69). All six multi-trait scores were also still correlated with the holistic score at a weak but statistically significant level (0.24–0.55). Interestingly enough, however, it was the vocabulary, sentence variety/construction, and grammar/usage scores that were more highly correlated with the holistic score (0.44–0.55). The correlations between the holistic score and three multi-trait scores of development, organization, and mechanics were lower (0.24–0.40).

Multidimensional scaling analyses of holistic and multi-trait scores

To examine further the empirical relationships among the holistic and multi-trait scores through a graphical representation of these scores, multidimensional scaling analysis was conducted separately for each of the two prompts. Figure 3 shows the plots of the holistic and multi-trait scores in a two-dimensional space obtained from the multidimensional scaling analysis. Since the results of one-, two-, and three-dimensional solutions showed similar relations among the holistic and multi-trait scores, only the plots for the two-dimensional solution (which are somewhat intuitively easier to interpret) are provided here.

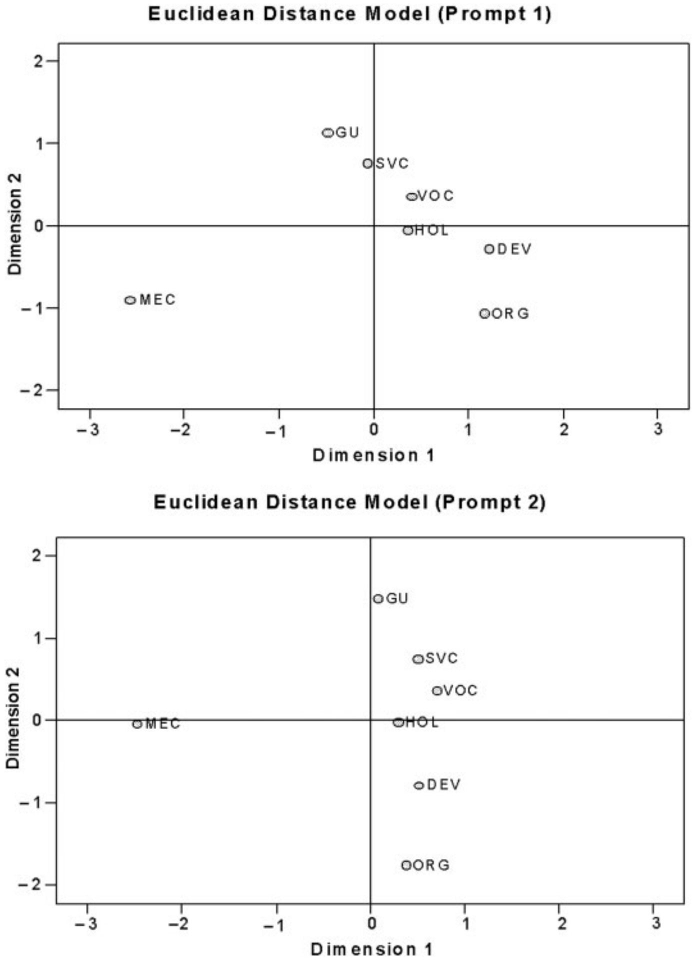

Figure 3: Representation of holistic and multi-trait scores in the two-dimensional space based on multi-dimensional scaling analysis for two writing prompts

Similar results were obtained for both prompts, as shown in Figure 3. First, the mechanics score (MEC) seemed to be somewhat distinct from the remaining six scores for both prompts. The first dimension represented by the *X*-axis (abscissa) seemed to be playing an important role in separating the mechanics score from the remaining six scores. The mechanics score is located horizontally on the far left (negative) side of Dimension 1, whereas the remaining six scores were scattered around the mid-point on the first dimension, leaning more toward the positive side.

Second, the remaining six scores were differentiated vertically on the second dimension represented by the *Y*-axis (ordinate). For both prompts, the holistic score (HOL) was located near the mid-point on the second dimension, dividing the remaining five multi-trait rating dimensions (except mechanics) into two theoretically meaningful clusters (e.g. content/rhetoric versus language). Located above the holistic score (on the positive side) are the three language-related dimensions of vocabulary (VOC), sentence variety/construction (SVC), and grammar/usage (GU), whereas the development (DEV) and organization (ORG) dimensions were located below the holistic score (on the negative side).

Correlation analyses of multi-trait scores and *e-rater* essay feature variable scores

Table 4 displays the averaged zero-order correlations and Table 5 shows the averaged partial correlations between the six multi-trait scores and the 12 *e-rater* essay feature variables across the two prompts. As expected, the magnitude of the partial correlations between the human-assigned essay scores and the eleven *e-rater* essay feature variables were much smaller overall than that of the zero-order correlations between these variables. Since the results for Prompts 1 and 2 are similar, only the averaged zero-order and partial correlations across the two prompts are reported here.[8] More detailed descriptions of the correlations follow.

Table 4: Averaged zero-order correlations between holistic/multi-trait scores and e-rater essay feature scores across the two prompts

e-rater variables	Holistic	Multi-trait						TNW
		DEV	ORG	VOC	SVC	GU	MEC	
D-Unit score	0.69	0.68	0.70	0.63	0.63	0.58	0.57	0.68
Length of DU	0.26	0.24	0.20	0.27	0.25	0.22	0.16	0.34
Type/token R	0.36	0.39	0.45	0.27	0.31	0.29	0.37	0.43
Word length	0.13	0.11	0.14	0.14	0.13	0.12	0.10	–0.03
Vocabulary level	0.60	0.58	0.52	0.60	0.56	0.52	0.40	0.62
Word–vector score	0.53	0.45	0.43	0.54	0.52	0.49	0.35	0.46
Word–vector correlation	0.77	0.71	0.67	0.71	0.70	0.68	0.57	0.78
GA ratio	0.39	0.35	0.34	0.38	0.38	0.37	0.37	0.31
UA ratio	0.11	0.09	0.12	0.11	0.10	0.12	0.18	0.05
SA ratio	0.62	0.58	0.52	0.65	0.60	0.55	0.46	0.60
MA ratio	0.50	0.44	0.46	0.45	0.46	0.46	0.59	0.39
Total number of words	0.90	0.88	0.80	0.83	0.80	0.74	0.64	1.00

Note: $n = 930$. All correlations were significant at the 0.01 level (two-tailed).

Table 5: Partial correlations between holistic/multi-trait scores and e-rater essay feature scores for the two prompts

e-rater variables	Holistic	Multi-trait					
		DEV	ORG	VOC	SVC	GU	MEC
D-Unit score	0.11	0.14	0.23	0.00*	0.06*	0.05*	0.11
Length of DU	−0.14	−0.15	−0.18	−0.04*	−0.07*	−0.08*	−0.11
Type/token R	−0.24	−0.09*	0.05*	−0.37	−0.23	−0.18	0.01*
Word length	0.38	0.29	0.30	0.32	0.26	0.21	0.15
Vocabulary level	0.07*	0.05*	−0.03*	0.15	0.06*	0.07*	−0.07*
Word–vector score	0.30	0.08	0.09	0.31	0.28	0.23	0.06*
Word–vector correlation	0.20	0.02*	0.04*	0.10	0.15	0.19	0.09*
GA ratio	0.20	0.11	0.10	0.17	0.18	0.17	0.17
UA ratio	0.18	0.11	0.16	0.14	0.13	0.15	0.22
SA ratio	0.14	0.06*	−0.02*	0.28	0.17	0.12	0.04*
MA ratio	0.30	0.15	0.18	0.16	0.19	0.22	0.43
Total number of words	NA	NA	NA	NA	NA	NA	NA

Note: All correlations were significant at the 0.05 level, except the asterisked (*) ones.

The essay length (TNW) turned out to be the automated essay feature variable that had the strongest zero-order correlation with all of the six multi-trait scores (0.64–0.88) and the holistic score (0.90). Among them, the holistic and development scores were most sensitive to the essay length variable in particular. Nevertheless, it should be noted that the strength of the relationship between the essay length and the grammar (0.74) and mechanics scores (0.64) seemed somewhat weaker than those between the essay length and the rest of the multi-trait and holistic scores (HOL, ORG, VOC, SVC).

Of the two variables related to discourse and organization that are used in *e-rater*, the discourse unit score had consistently moderate zero-order correlations (0.57–0.70) with all of the seven essay scores (including the single holistic and six multi-trait scores), and it was correlated more strongly with the first three human supplied scores (holistic, DEV, ORG) than with the remaining four (VOC, SVC, GU, MEC). This also turned out to have the strongest, partial correlation with the organization score assigned by human raters (0.23). In contrast, the average length of discourse units in the essay turned out to have relatively lower correlations (0.16–0.27), and even negative partial correlations (–0.18 to –0.04), with all seven essay scores.

Among the three *e-rater* lexical complexity variables, the vocabulary-level feature variable based on word-frequency levels had consistently moderate correlations with the seven scores for both prompts (0.40–0.60). In relation to this feature, one encouraging finding was that this *e-rater* vocabulary level variable was most highly correlated with the vocabulary (VOC) and holistic scores (HOL) assigned by human raters (0.60). Even when the partial correlations were examined, the vocabulary level feature variable had the highest correlation with the vocabulary score (0.15).

Of the two prompt-specific vocabulary usage variables, the word–vector (cosine) correlation variable had high correlations with the seven essay scores (0.57–0.77) than the word–vector score (0.35–0.54). The first variable (word–vector correlation) was most strongly correlated with the holistic score and most weakly with the mechanics score, while the second variable (word–vector score) was most strongly correlated with the vocabulary (0.54) and holistic scores (0.53).

Following are several other statistically significant results: First, among the four linguistic accuracy variables, the stylistic accuracy ratio variable was consistently most highly correlated with the seven essay scores (0.46–0.65) in terms of zero-order correlations, and it was most highly correlated with the vocabulary score. Second, in relation to the linguistic accuracy variables, the most consistent finding was that a close link was confirmed between the mechanical accuracy ratio computed by *e-rater* and the mechanical accuracy rating (MEC) given by human raters. The mechanical accuracy ratio was correlated with all of the seven essay scores (0.44–0.59).

SUMMARY AND DISCUSSION

The main purposes of the present study were to investigate whether distinct (separable) and reliable (dependable) multi-trait rating dimensions could be identified in the context of the TOEFL® CBT writing assessment and to examine the relationships of the multi-trait scores to the holistic score and to the *e-rater*® essay feature variables. High score reliability was achieved for all of these six multi-trait rating dimensions. It was found that (i) all of the six multi-trait scores were correlated among themselves; (ii) these multi-trait scores were also correlated with the holistic score; (iii) high correlations obtained among holistic and multi-trait scores were largely attributable to the impact of essay length on both multi-trait and holistic scoring; and (iv) some strong associations were confirmed between several *e-rater* variables and multi-trait rating dimensions. These findings are discussed next in more detail.

Reliability of multi-trait scores

Results of the study show that overall the trained human raters were able to apply the multi-trait scoring rubric consistently across the rating dimensions. All of the essays analyzed were double-rated by two raters on each of the multi-trait rating dimensions and the average of the two rater's scores was used in data analysis to ensure the reliability of the multi-trait scores examined in this study. For this reason, acceptable levels of score reliability (>0.80) were obtained for all of the six multi-trait rating dimensions overall.

Relationships among multi-trait scores

Examinations of zero-order correlations among the multi-trait rating dimensions revealed that all of the six multi-trait rating dimensions were correlated among themselves and also with essay length at moderate to high levels. Among the six multi-trait rating dimensions compared, the mechanics dimension seemed to be most distinct from the rest of the multi-trait dimensions. For both prompts, the lowest correlations among the six rating dimensions involved the mechanics dimension (i.e. the mechanics and development pair for Prompt 1 and the mechanics and organization pair for Prompt 2). Another related pattern deserving mentioning here was that the first four dimensions (development, organization, vocabulary, and sentence variety/construction) were more sensitive to essay length than were the last two accuracy-related dimensions (grammar/usage and mechanics). The development score was most strongly correlated with essay length, while the mechanics score was most weakly correlated. (Please also note that the holistic score was actually more strongly correlated with essay length than any of the six multi-trait scores.)

Relationships between holistic and multi-trait scores

Close examinations of correlations and MDS analysis results revealed that all of the seven rating dimensions (including the holistic scoring dimension) seemed to be measuring related but somewhat distinct aspects of essay quality. First, each of the six multi-trait scores was found to be correlated with the holistic score. The development, vocabulary, and sentence variety/construction scores were most strongly correlated with the holistic score. The lowest correlations were observed for mechanics. These results are consistent with previous research findings on the relationships between multi-trait and holistic ratings assigned to ESL learner's essays (Bacha 2001) and those between lexical diversity and holistic scores (Engber 1995; Laufer and Nation 1995).

Second, results of MDS analyses not only confirmed the patterns of the relationships among the rating dimensions from the zero-order correlations but also provided more in-depth insights into the nature of the holistic rating and the relationships between the multi-trait and holistic rating dimensions. An inspection of two-dimensional plots of the seven scoring dimensions showed that the mechanics dimension was most distinct from the remaining six rating dimensions. The holistic rating dimension was also found to be very useful in grouping the remaining five multi-trait rating dimensions (except mechanics) into two theoretically distinct clusters of dimensions in the plots. This suggests that the holistic score does reflect both the content-related and language-related qualities of the essays, as defined in the TOEFL CBT scoring rubric, and that the content/rhetoric dimensions are separable to some extent from the language-related dimensions, as pointed out by some ESL writing researchers (Santos 1988; Cumming *et al*. 2002).

The main implication of these findings is that, by virtue of its distinctiveness from other scores, a separate mechanics score seems clearly justified in any effort to provide a set of multi-trait scores. The creation of super-ordinate rating dimensions of content and language for profile scoring is also an additional area of research deserving further investigation. From these results, however, the justification for other multi-trait scores seems to be somewhat more equivocal.

Role of essay length in holistic and multi-trait scores

A strong empirical relationship, not only between the essay length and holistic score but also between essay length and each of the six multi-trait scores used, was confirmed in this study. This means that essay length co-occurs with other highly valued aspects of essay quality captured through holistic and multi-trait scoring rubrics. Such essay-length-related findings were not completely unexpected, given previous research findings on the strong relationships between essay length and holistic scores (Carson *et al*. 1985; Reid 1986; Ferris 1994; Frase *et al*. 1999; Grant and Ginther 2000; Jarvis 2002; Jarvis *et al*. 2003) and between lexical diversity measures and holistic scores (Engber 1995; Laufer and Nation 1995).

To better understand the empirical, essay-length independent relationships between the holistic and multi-trait essay scores, partial correlations were also computed in this study after removing the effect of essay length from the original correlations. Although the obtained partial correlations were significantly lower than the original correlations, all of the six multi-trait scores remained correlated not only among themselves at a significant level but also with the holistic scores.

One interesting pattern emerged from the partial correlations among the holistic and multi-trait scores, however. The development score was no longer correlated most strongly with the holistic score. Instead, the three dimensions related to the knowledge of language components (vocabulary, sentence variety/construction, and grammar/usage) turned out to be more highly correlated with the holistic scores than the development and organization scores. This indicates that the three language-related dimensions have greater, essay-length-independent, explanatory power for the holistic scores than the development and organization dimensions.

The main implication of these findings is that, if essay length could be controlled or constrained (e.g. imposing a strict fixed essay length requirement under un-timed testing conditions), multi-trait ratings might have greater distinctiveness and therefore greater utility.[9] This is an issue that could be researched.

Relationships between multi-trait scores and *e-rater* essay features

A total of 12 essay feature variables used in *e-rater* 2.0 were analyzed in this study. Among these 12 variables, essay length (measured by the total number of words) turned out to be the strongest predictor of each of the six multi-trait scores as well as the holistic score. Since the role of essay length for the holistic and multi-trait scores was already discussed in the previous section, the remainder of the discussion is focused on some of the remaining 11 variables.

Of the two development/organization features, the discourse unit score seemed to be working as desired in tapping the surface level of organizational quality of essays, but the average length of discourse units did not seem to be working as intended to capture the developmental aspect of essay quality. Please note that the discourse unit score is defined as the difference between the actual and optimal number of discourse units in the essay, which can be related to the organizational aspect of essay quality.

Among the three lexical complexity variables, both the vocabulary level and average word length variables seemed to be able to capture what human raters value in terms of the lexical variety/sophistication aspect of essay quality. However, the type–token ratio was shown to be sensitive to essay length and, interestingly, correlated negatively with human judgement of lexical sophistication for ESL essays when essay length is controlled for. In relation to this, one interesting research area deserving further investigation is the development and use of more sophisticated type-token ratio measures that are not dependent on text length for ESL learners' essays (Jarvis 2002).

Lastly, the most clear-cut finding from the *e-rater* variable analysis was that it was possible to establish a link between the mechanical accuracy ratio computed by the automated scoring engine and the mechanics score assigned by human raters. This pattern was observed both for the zero-order and partial correlations. The mechanical accuracy ratio was most strongly correlated with the mechanics score for both prompts. This suggests that, in terms of mechanical accuracy, the mechanical accuracy ratio may reflect the same qualities that human raters attend to in examinee's essays.

However, we were not able to confirm a similar link between the grammatical accuracy ratio variable and the grammar/usage scores assigned by human raters or between the usage accuracy ratio and the grammar/usage score. Further investigation seems necessary to identify the potential causes of such a weak correlation between the e-rater and human-assigned grammar/usage accuracy scores. In contrast, the stylistic accuracy ratio turned out to be most strongly correlated with the vocabulary score, whether the zero-order or partial correlations were used. In a sense, the highest correlation between the stylistic accuracy ratio and the vocabulary score is somewhat expected, given that one major type of errors that contributes to the stylistic accuracy ratio is excessively repeated words across sentences and passage in the essay.

Overall, we saw reasonably strong associations between several *e-rater* variables and multi-trait rating dimensions in some areas of essay quality, such as organization, vocabulary, and mechanics. This means that, for these variables, both *e-rater* and human raters are focusing on similar or related aspects of examinees' essays. This provides some evidence supporting the validity of not only the automated text features but also the automated holistic scores computed based on these features in *e-rater* and *Critique*. To a certain degree, it seems also justifiable to use some of these existing *e-rater* variables to compute automated trait scores representing different aspects of essay quality.

Despite these encouraging results, we also noticed some conceptual mismatch between the six multi-trait scores and the 12 *e-rater* essay feature variables. For instance, there is clearly no *e-rater* variable that captures directly the sentence variety/construction aspect of essay quality. Further research is necessary to create the *e-rater* essay feature variables to capture a full range of essay quality features valued in ESL writing. These may include not only sentence variety but also other essay features, such as depth of development, coherence, and appropriateness of lexical choice.

IMPLICATIONS FOR THE BROADER FIELDS OF SECOND LANGUAGE WRITING

The findings of the current study have important implications for the validation strategies for automated essay scores in general and also the broader fields of second language writing and instruction and computer-based writing assessment and instruction. The issues are discussed briefly next.

Validation strategies for automated essay scores

Yang *et al.* (2002) classify validation approaches for automated scores into three major types: (i) approaches focusing on the relationships between scores generated by the computer and human scorers, (ii) approaches focusing on the relationships between test scores and external measures, and (iii) approaches focusing on the scoring processes. In terms of the first and second types of validity evidence, previous research studies have demonstrated that a high score agreement rate could be achieved between human raters and automated scoring systems (Kukich 2000; Attali and Burstein 2006; Ben-Simon and Bennett 2007) and that the automated and human scores exhibited reasonably similar relations with various independent indicators of writing skills, although these relations tended to be somewhat weaker for automated scores (Powers *et al.* 2001a; Lee 2006).

The third type of approaches involves more descriptive and qualitative analysis of the patterns and nature of the disagreements between the automated scoring systems and human scorers (Yang *et al.* 2002). In the context of AES, this means that writing or content experts should be invited to examine not only the nature of the automated essay features employed in AES but also the way they are combined to produce automated scores. In particular, it is critical to judge the theoretical and practical relevance of the automated essay feature variables to the target construct of writing, identify irrelevant features as well as missing ones, and evaluate the appropriateness of the weights assigned to the selected set of features for the AES systems (Ben-Simon and Bennett 2007).

The current study represents one of the first serious attempts to use such a more content-oriented approach (along with other approaches) in the examination of the validity of the automated essay scores for ESL learners. In this study, through the examination of the conceptual and empirical relationships between the *e-rater* essay feature variables and the six multi-trait rating dimensions, we were able to identify not only some meaningful construct-relevant relationships between them but also the relevant essay feature variables that are lacking or need to be further refined in the current *e-rater*/*Criterion* systems. Continued research on this line will further strengthen the validity and accuracy of automated essay scoring and feedback.

New generation of TOEFL: TOEFL *i*BT

The findings of the study could also have important implications for the newly launched Internet-based TOEFL (TOEFL *i*BT). The TOEFL *i*BT writing section is made up of two writing tasks (i.e. one independent and one integrated writing task). One important feature of TOEFL *i*BT is to promote 'enhanced scoring', that is, to provide performance feedback in addition to the total and section scores to TOEFL test-takers (ETS 2007). Here, the performance feedback includes descriptive information about the test taker's proficiency level and areas of strengths and weaknesses. Automated multi-trait scoring technology being explored in this study can potentially contribute to making the writing performance feedback more tailored to individual examinees with different profiles of strengths and weaknesses. Especially, since the TOEFL CBT writing tasks are very similar to TOEFL *i*BT independent writing tasks, it is our hope that insights gained from this study will prove valuable in investigating the feasibility of automated multi-trait scoring for TOEFL *i*BT independent writing tasks.

Moving beyond controversies on AES

Despite its efficiency, objectivity, and consistency in scoring, and many other positive benefits for writing assessment instruction, AES has generated a series of controversies and heated debates in the communities of writing experts and second language professionals. While some criticisms have been raised against AES (Herrington and Moran 2001; Ericsson and Haswell 2006; Ziegler 2006; Phillips 2007), some counter-arguments have also been put forward by writing experts in defence of AES subsequently (Haswell 2006;

Cumming 2007). There is another group of writing experts (Williamson 2004; Anson 2006; Haswell 2006) who rather strongly urge the writing community to actively investigate the potential values of AES for teaching, learning, and assessment, and also its potential harmful effects than arduously criticize or defend it. In this more empirically based perspective, it is important to carefully examine how the automated scores and feedback are utilized by the learners, identify the potential causes of inaccurate diagnosis and detection in the AES system, and find a way to enhance or augment the system.

Along the same line, the techniques and procedures that underlie AES methods, particularly the scoring variables and the weights used in the AES systems, probably need to be communicated (or disclosed) to the writing experts, language teachers, and the public for evaluation and feedback in the future, although the proprietary issues of the technology constrain the developers not to do so at the moment (Powers *et al*. 2001b; Phillips 2007). Dikli (2006) even suggests envisioning AES systems as a free public utility rather than proprietary, vendor-created, and owned software. Such progressive, forward-looking AES implementation models, if adopted, will be able to facilitate open, constructive discussions among the developers, writing experts/teachers, and students and further advance the AES technology in the desired direction of making it best serve the learners and teachers. Our sincere hope is that the current study contributes to sparking such open discussions on AES in the communities of writing experts, applied linguists, and language teachers.

ACKNOWLEDGEMENTS

This article is based on a longer version of a TOEFL/ETS research report. Some portions of the earlier versions of the report were presented at the 27th Language Testing Research Colloquium (LTRC) held in Ottawa, Canada and at the 14th World Congress of Applied Linguistics (AILA) held in Madison, Wisconsin in July 2005. Needless to say, the responsibility for any errors that remain is solely ours, and any opinions expressed in this article are those of the authors and not necessarily of ETS. The copyright in this abridgement is solely held by Oxford University Press, for permission to reuse any text from this article please contact journals.permissions@oxfordjournals.org

NOTES

[1] The terms ‘analytic scoring’ and ‘multi-trait scoring’ are used synonymously in this report.

[2] This also may mean that only a small increase in score reliability can be achieved for the composite of multi-trait scores, compared with the holistic score, as one anonymous reviewer has rightly pointed out.

[3] See Dikli (2006) for more information on My Access!®, which is another example of online writing practice service that provides instant scoring and feedback to students.

4 In *Critique*, flagging words and phrases suspected of containing errors is largely done based on low probability bigrams or trigrams when evaluated against a large language corpus of well-formed text produced by native English speakers (Leacock and Chodorow 2003). The errors detected by *Critique* are classified into four major categories: grammar, usage, mechanics, and style (GUMS) (Quinlan *et al.* 2009). Included in the 'grammar' errors are fragments, run-on sentences, subject-verb agreement and possessive errors, etc. The 'usage' (U) category includes errors related to articles, confused words, and wrong word forms, while the 'mechanics' (M) errors include spelling, duplicate word, hyphenation, and other punctuation errors. Lastly, among the 'style' (S) errors are repeated words, inappropriate words, excessive use of coordinating conjunctions, two long or short sentences, and sentences in the passive voice. Research has been underway to improve *Critique*'s capabilities to detect more ESL-relevant errors, such as article, preposition, and lexical choice errors. For more information, please see Han *et al.* (2006) and Tetreault and Chodorow (2008).

5 Since the lowest score point of 1 is rarely used by raters for TOEFL CBT essays, it is often difficult to represent this score category in a random sample. For this reason, a systematic (or 'stratified') sampling scheme was used for the first sample to cover a full range of essay scores, including the lowest score point of 1. The stratified sample consisted of 50 essays for each of the score categories from 2 to 6 and 15 essays for the score category of 1 to represent the entire holistic score range.

6 Both TNW (total number of words) and TNW-squared values were used as covariates in computing the partial correlation to control the linear and quadratic effects of essay length on the correlation.

7 The type of score reliability coefficients reported here is the reliability coefficients for absolute score interpretation (often called *dependability indices*). Since the rating main effect was small for each of the dimensions for the two prompts, both the dependability indices and generalizability coefficients were very close (see Brennan 2001 for more information on these coefficients). For this reason, only the dependability indices, which are rather conservative estimates of reliability, are reported here.

8 Of course, there were slight differences across the two prompts in terms of magnitude of correlations. However, the size of correlation for the same pair of variables was generally similar across the prompts. More importantly, the overall patterns of relations were almost the same.

9 In relation to this, one reviewer pointed out that it would also be important to develop research agenda devised to examine why essay length functions so well as a proxy for essay quality (rather than try to remove its influence). Since what were captured through the development and organization scores in the multi-trait rubric appear closely and necessarily linked with essay length, controlling essay length would make it difficult, if not impossible, to examine the characteristics of such essay quality features at the high end of the proficiency scale. In a sense, both essay length and development can be understood as manifestations of fluency. When the cost of scoring systems is an important

consideration, an argument can be made for using length simply as a proxy for development and organization in the essay. However, as soon as the concern shifts away from scores on writing assessments to instruction, then models are required that allow us to better understand how a scoring dimension like development manifests itself in written work across proficiency levels.

REFERENCES

Anson, C. M. 2006. 'Can't touch this: reflections on the servitude of computers as Readers' in P. F. Ericsson and R. Haswell (eds): *Machine Scoring of Students Essays: Truths and Consequences*. Utah State University, pp. 38–56.

Attali, Y. and **J. Burstein.** 2006. 'Automated essay scoring with *e-rater*® V.2,' *Journal of Technology, Learning, and Assessment* 4/3: available at http://www.jtla.org. Accessed 16 April 2007.

Bacha, N. 2001. 'Writing evaluation: what can analytic versus holistic essay scoring tell us?' *System* 29: 371–83.

Ben-Simon, A. and **R. E. Bennett.** 2007. 'Toward more substantively meaningful automated essay scoring.' *Journal of Technology, Learning, and Assessment* 6/1: available at http://www.jtla.org. Accessed 6 January 2008.

Brennan, R. L. 1999. *Manual for mGENOVA Version 2.0.* The University of Iowa.

Borg, I. and **P. Groenen.** 1997. *Modern Multidimensional Scaling: Theory and Applications.* Springer.

Carson, S., B. Bridgeman, R. Camp and **J. Waanders.** 1985. *Relationship of Admission Test Scores to Writing Performance of Native and Non-native Speakers of English* (TOEFL Research Report No. 19). ETS.

Connor-Linton, J. 1995. 'Looking behind the curtain: what do L2 composition ratings really mean?' *TESOL Quarterly* 29: 762–5.

Crick, G. E. and **R. L. Brennan.** 1983. *Manual for GENOVA: A Generalized Analysis of Variance System* (ACT Tech. Bulletin No. 43). American College Testing Program.

Cumming, A. 2007. 'Book review: Machine scoring of students' essays: Truth and consequences,' *Assessing Writing* 12: 80–2.

Cumming, A., R. Kantor, and **D. Powers.** 2002. 'Decision-making while rating ESL/EFL writing tasks: A descriptive framework,' *Modern Language Journal* 86/1: 67–96.

Dikli, S. 2006. 'An overview of automated scoring of essays,' *Journal of Technology, Learning, and Assessment* 5/1: available at http://www.jtla.org. Accessed 30 March 2007.

Educational Testing Service. 1998. *Computer-based TOEFL® Score User Guide.*

Educational Testing Service. 2007. *TOEFL® iBT Tips: How to Prepare for the TOEFL iBT.*

Engber, C. A. 1995. 'The relationship of lexical proficiency to the quality of ESL compositions,' *Journal of Second Language Writing* 4: 139–55.

Ericsson, P. F. and **R. Haswell.** 2006. *Machine Scoring of Students Essays: Truths and Consequences.* Utah State University.

Ferris, D. 1994. 'Lexical and syntactic features of ESL writing by students at different levels of L2 proficiencies,' *TESOL Quarterly* 28: 414–20.

Frase, L., J. Faletti, A. Ginther and **L. Grant.** 1999. *Computer Analysis of the TOEFL Test of Written English* (TOEFL Research Report No. 64). ETS.

Freedman, S. W. 1984. 'The registers of student and professional expository writing. Influences on teacher responses' in R. Beach and S. Bridwell (eds): *New Directions in Composition Research*. Guilford Press, pp. 334–47.

Gentile, C., A. Riazantseva, and **F. Cline.** 2002. *A Comparison of Handwritten and Word-processed TOEFL Essays: Final Report*. Internal document. ETS.

Grant, L. and **A. Ginther.** 2000. 'Using computer-tagged linguistic features to describe L2 writing differences,' *Journal of Second Language Writing* 9: 123–145.

Hamp-Lyons, L. 1991. 'Scoring procedures for ESL contexts' in L Hamp-Lyons (ed.): *Assessing Second Language Writing in Academic Contexts*. Ablex, pp. 241–76.

Hamp-Lyons, L. 1995. 'Rating nonnative writing: the trouble with holistic scoring,' *TESOL Quarterly* 29: 759–62.

Han, N.-R., M. Chodorow, and **C. Leacock.** 2006. 'Detecting errors in English article usage by nonnative speakers,' *Natural Language Engineering,* 12/2: 115–29.

Haswell, R. 2006. 'Automatons and automated scoring: drudges, black boxes, and Dei Ex Machina' in P. F. Ericsson and R. Haswell (eds): *Machine Scoring of Students Essays: Truths and Consequences*. Utah State University, pp. 57–78.

Herrington, A. and **C. Moran.** 2001. 'What happens when machines read our students' writing?' *College English* 63/4: 480–99.

Huot, B. 1990. 'The literature of direct writing assessment: major concerns and prevailing trends,' *Review of Educational Research* 60: 237–63.

Jacobs, H. L., S. A. Zingraf, D. R. Wormuth, V. F. Hartfiel, and **J. B. Hughey.** 1981. *Testing ESL Composition: A Practical Approach*. Newbury House.

Jarvis, S. 2002. 'Short texts, best-fitting curves, and new measures of lexical diversity,' *Language Testing* 19: 57–84.

Jarvis, S., L. Grant, D. Bikowski, and **D. Ferris.** 2003. 'Exploring multiple profiles of highly rated learner composition,' *Journal of Second Language Writing* 12: 377–403.

Kondo-Brown, K. 2002. 'A FACETS analysis of rater bias in measuring Japanese second language writing performance,' *Language Testing* 19/1: 3–31.

Kukich, K. 2000, September/October. 'Beyond automated essay scoring,' *IEEE Intelligent Systems* 15/5: 22–7.

Leacock, C. and **M. Chodorow.** 2003. 'Automated grammatical error detection' in M. D. Shermis and J. C. Burstein (eds): *Automated Essay Scoring: A Cross-disciplinary Perspective*. Lawrence Erlbaum, pp. 195–207.

Lee, Y.-W. C. 2006. *Variability and Validity of Automated Essay Scores for TOEFL iBT: Generic, Hybrid, and Prompt-specific Models*. Internal document. ETS.

Lee, Y.-W., C. Gentile, and **R. Kantor.** 2008. *Analytical Scoring of TOEFL CBT Essays: Scores by Humans and e-rater®* (TOEFL Research Report No. 81; ETS RR-08-01). ETS.

Laufer, B. and **P. Nation.** 1995. 'A vocabulary size test of controlled productive ability,' *Language Testing* 16: 33–51.

Phillips, S. M. 2007. *Automated Essay Scoring: A Literature Review*. Society for the Advancement of Excellence in Education.

Powers, D., J. Burstein, M. Chodorow, M. E. Fowles, and **K. Kukich.** 2001a. *Comparing the Validity of Automated and Human Essay Scoring* (GRE Board Professional Report No. 98-08aR). ETS.

Powers, D., J. Burstein, M. Chodorow, M. E. Fowles, and **K. Kukich.** 2001b. *Stumping E-rater: Challenging the Validity of Automated Essay Scoring* (GRE Board Professional Report No. 98-08bP). ETS.

Quinlan, T., D. Higgins, and **S. Wolff.** 2009. *Evaluating the Construct-coverage of the e-rater® Scoring Engine*. (ETS Research Report No. RR 09-01). ETS.

Raimes, A. 1990. 'The TOEFL test of written English: causes for concern,' *TESOL Quarterly* 24: 427–42.

Reid, J. 1986. 'Using the Writer's Workbench in composition teaching and testing' in C. Stansfield (ed.): *Technology and Language Testing*. TESOL, pp. 167–88.

Roid, G. H. 1994. 'Patterns of writing skills derived from cluster analysis of direct writing assessments,' *Applied Measurement in Education* 7/2: 159–70.

Santos, T. 1988. ''Professors' reactions to the academic writing of non-native speaking students,' *TESOL Quarterly* 22: 69–90.

Sasaki, M. and **K. Hirose.** 1999. 'Development of an analytic rating scale for Japanese L1 writing,' *Language Testing* 16/4: 457–78.

Shermis, M. D. and **J. C. Burstein.** 2003. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Lawrence Erlbaum.

SPSS. 2003. *SPSS Base 12.0 User's Guide*. Author.

Swartz, C. W., S. R. Hooper, J. W. Montgomery, M. B. Wakely, R. E. L. de Kruif, M. Reed, T. T. Brown, M. D. Levine, and **K. P. White.** 1999. 'Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods,' *Educational and Psychological Measurement* 59/3: 492–506.

Tetreault, J. and **M. Chodorow.** 2008. Native judgments of non-native usage: Experiments in preposition error detection. *COLING*

Workshop on Human Judgments in Computational Linguistics. Manchester, UK.

Veal, L. R. and **S. A. Hudson.** 1983. 'Direct and indirect measures for large-scale evaluation of writing,' *Research in the Teaching of English* 17: 285–96.

Weigle, S. C. 2002. *Assessing Writing*. Cambridge University Press.

Weir, C. J. 1990. *Communicative Language Testing*. Prentice Hall Regents.

Williamson, M. M. 2004. 'Validity of automated scoring: prologue for a continuing discussion of machine scoring of student writing,' *Journal of Writing Assessment* 1/2: 85–104.

Williamson, M. M. and **B. A. Huot.** 1993. *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Hampton Press.

Yang, Y., C. W. Buckendahl, P. J. Jusziewicz, and **D. S. Bhola.** 2002. 'A review of strategies for validating computer-automated scoring,' *Applied Measurement in Education* 15: 391–412.

Ziegler, W. W. 2006. 'Computerized writing assessment: community college faculty find reasons to say "not yet"' in P. F. Ericsson and R. Haswell (eds): *Machine Scoring of Students Essays: Truths and Consequences*. Utah State University, pp. 138–46.

Applied Linguistics: 31/3: 418–442 © Oxford University Press 2009
doi:10.1093/applin/amp041 Advance Access published on 20 November 2009

The Effects of Content and Language Integrated Learning in European Education: Key Findings from the Andalusian Bilingual Sections Evaluation Project

FRANCISCO LORENZO, SONIA CASAL and PAT MOORE

Universidad Pablo de Olavide, Seville

Content and Language Integrated Learning (CLIL) represents an increasingly popular pedagogic approach that has evolved in response to the recognised need for plurilingual competence in Europe. In this article, we present key findings from one of the first large-scale, multidimensional CLIL evaluation projects. We begin by outlining the emergence of European CLIL and by comparing it with other, non-European bilingual education initiatives and then we narrow the scope to Southern Spain, where the research was conducted. We outline the Andalusian Bilingual Sections programme, one of the cornerstones of the government's *Plurilingualism Promotion Plan* (2005), within which the research was conducted. In presenting results, we focus on specific areas that we believe make significant contributions to some of the key concerns in contemporary CLIL research including the linguistic competence of CLIL learners, the question of starting age, the distribution and functionalities of L2 use in CLIL classrooms, and the ways in which CLIL appears to be impacting on the educational system in general.

INTRODUCTION

The idea of teaching subject matters through more than one language is not new; indeed the very foundations of formal education in Europe were multilingual (Lewis 1976; Adams 2003; Braunmüller and Ferraresi 2003). For a variety of reasons, however, pragmatic as well as political, as general education spread to the masses it became increasingly monolingual. In the process bilingual education became a prerogative of the elite (de Mejía 2002). Recently, however, there has been a shift in attitudes towards the notion. The *1+2 principle*, encapsulated in the European Commission's *White Paper on Education and Training* (1995), idealises European citizens as having at least partial competences in two languages other than their first and argues that this goal needs to be incorporated into national curricula. As a consequence, most European states are currently implementing bilingual-type programmes in national education. The abundance of new initiatives suggests that this represents more than just a quantitative increase of second language provision in schools. The change now is pervasive and the foundations appear to be set for European multilingualism—the social phenomenon of multiple languages in social groups, and European plurilingualism—an ample language repertoire amongst a majority of individuals, which should enable students not only to *savoir* but also to *savoir faire* and *savoir être* in a reconfigured continental environment.

The acronym CLIL, standing for Content and Language Integrated Learning, has been adopted to describe this new European trend. CLIL serves as an umbrella term embracing all scenarios and whatever combination of regional, heritage, minority, immigrant and/or foreign languages they involve; providing for a highly diversified language curriculum. The origins of CLIL can be traced to the German-Franco programmes at the geographical core of Europe which have slowly spread out until now they are to be found in all but a few of the furthest reaches of the continent: Iceland on the far northwest, Portugal on the far southwest, Greece on the far southeast and Latvia on the far northeast (Eurydice 2006). Nonetheless, this extensive presence stands in contrast to the lack of a coherent conceptual framework which may be applied in all contexts. As Dalton-Puffer notes:

> Content and language integrated learning has happened at two curiously distant levels of action: on the level of local grassroots activity on the one hand and on the level of EU policy on the other leaving the intermediate level of national educational policies largely unaccounted for (2008: 139).

The study here presented comes from another of Europe's frontiers, Andalusia—the region which extends across the whole of Southern Spain and which, with some 8 million. inhabitants, may be compared with other European nations. In 2005 the Andalusian government launched the *Plan de Fomento del Plurilingüismo* (the Plurilingualism Promotion Plan; henceforth the Plan).[1] The Plan represents a concerted effort to adhere to European policy and is built around five programmes incorporating seventy-four distinct strategies to be implemented over the period 2005–9. Its ultimate aim is to engender a radical shift from social monolingualism to multilingualism through education, under the European ethos that 'Europe will be multilingual or Europe will not be'. In Andalusia, it should be pointed out, possibilities for extra-mural exposure to and use of educational L2s are scarce and this reinforces the need for multilingualism through schooling.

The overall scope of the initiative clearly distinguishes it from other similar ventures: the entire educational network, primary and secondary, some four thousand schools, is to incorporate up to two new foreign languages as media of instruction, and half of the network is imparting up to 40 per cent of the curriculum in more than one language, taught by teachers recruited on the basis of their language profiles. All in all, both in numbers and extent, the venture resembles other national initiatives in language change through education such as the shift to bilingual teaching through Chinese and English in Hong Kong (Johnson 1997) or the language reversal move in Singapore (Pakir 1993) among others designed to promote multilingualism through schooling (for other examples, see Ager 2001 or Tollefson 2002).

As a route map for multilingual education the Plan gained institutional recognition through a European Language Label Award, to the satisfaction of local language planners who interpreted this concession as confirmation that the region, which has enjoyed significant subsidies from Europe, had invested wisely. More importantly, the Plan incorporates provision for monitoring and evaluation, which has shed light on a number of different aspects of CLIL implementation in formal settings. We believe that these are pertinent not only as a local example but to CLIL initiatives across the continent. Findings regarding language behaviour and competences in content-based settings; the discourse functions employed by content teachers as opposed to language teachers and native assistants in the programme; the impromptu incorporation of language across the curriculum and the effects on the education system are deemed particularly relevant.

METHODOLOGY AND RESEARCH QUESTIONS

Background: objectives and dimensions

This study is framed within the larger context of CLIL research, a brief review of which here follows. At the outset there was concern regarding the potential effects on content learning yet a series of studies focusing in particular on Mathematics (Jäppinen 2005; Seikkula-Leino 2007; Van de Craen *et al*. 2007) and the Social Sciences (Lamsfuß-Schenk 2002; Stohler 2006; Vollmer 2008) found that CLIL learners were at least matching, and at times even exceeding, monolingual peers. In general, these researchers have concluded that CLIL may hold the potential for positive cognitive gains. In tandem, both cross-section and longitudinal studies into CLIL learners' linguistic competences have suggested that not only do they demonstrate increased L2 proficiencies (Admiraal *et al*. 2006; Rodgers 2006; Ackerl 2007; Mewald 2007; Serra 2007) but that their L1 also appears to benefit from the bilingual experience (Nikolov and Mihaljević Djigunović 2006; Merisuo-Storm 2007). A parallel line of research has looked at bilingual education within the wider social context (see Housen on the European Schools network in Brussels, Italy and the UK (2002); Baetens Beardsmore on a selection of key bilingual case studies across Europe (1993) or Zydatiβ on the Berlin schools network (2007)). Results should also be interpreted alongside data coming from research on North American immersion (Johnson and Swain 1997; Arnau and Artigal 1998; Wesche 2002) and content-based teaching (Mohan 1986; Stryker and Leaver 1997; Snow 1998).

Within this context, this study takes a novel approach to CLIL research as it encompasses both linguistic analysis and the implementation of language planning of supranational language policies. Four key metaconcerns served as the cornerstones for the evaluation project here reported, and might help shape future evaluation projects. The four, further broken down into component corollaries, are:

1. Competence development
    - (i) Linguistic Competences in accordance with the levels of the Common European Framework of Reference (henceforth CEFR) (2001)
    - (ii) Conceptual Competences relating to the successful integration of content and language
    - (iii) Procedural Competences as demonstrated by the use of communicative, cognitive and meta-cognitive strategies
    - (iv) Attitudinal Competences combining both intercultural awareness and motivational factors

2. Curricular organisation
    - (i) The Model of Bilingual Education favoured—CLIL encompasses a wide range of potential models: single or dual, semi or complete immersion, translanguaging, modular thematic blocks and language showers
    - (ii) The Characteristics of the Bilingual Sections—incorporating the content subjects involved, the L2s and L3s most frequently chosen and the composition of the groups: what proportion of the school body is involved; how the groups are formed and whether they represent any particular social classes
    - (iii) The Coordination of Language and Content Integration—both the actors: administrators; language specialists, who may be teachers (L1 as well as L2s and L3s) or native-speaker/expert-user classroom assistants, and content specialists; and the methodologies and materials employed (both for teaching and testing)

3. Classroom praxis
    - (i) L2 use—incorporating both frequency and functions
    - (ii) Typology of Classroom Activities—including considerations relating to the pedagogic approach inherent therein and the classroom interaction patterns implied
    - (iii) Linguistic Approaches—sociolinguistic, discursive, functional, lexico-semantic, structural, etc.
    - (iv) Linguistic Range—academic and sociocultural themes and topics, metalanguage
    - (v) Skill and Competence Development—range, distribution and implementation
    - (vi) Materials—the mix of commercial and adapted materials involved, the use of authentic source materials, the development of material banks

(vii) The Design of Didactic Units—aligning conceptual and linguistic factors, thematic relevance, textual considerations, awareness-raising, etc.

(viii) Assessment Techniques—the objective/subjective mix, use of portfolios, self and collective evaluation, etc.

4. Levels of satisfaction

   (i) Perceptions of usefulness and success of diverse aspects of the bilingual programme including the early introduction of an L2 in primary education, the increase in L2 provision via content-integrated learning and the scope of the programme from the perspective of numbers involved.

Participants

Participant selection was organised in line with three major variables: urban/rural; primary/secondary education; and L2 (English, French and German). In the academic year 2007–2008, when the fact-finding component of this research was conducted, there were 403 schools across the region running bilingual sections. A two-stage sampling was employed in data-gathering. In the first stage, a sample of sixty-one institutions was randomly chosen across the eight provinces of the area of the study ensuring that each particular zone was evenly represented through a stratified sample approach. In the second stage, fourth year primary (aged 9–10) and second year secondary (aged 13–14) students were identified as target respondents. This population was selected because, at the time in question, taking all three L2s into account, these were the learners who had had the longest possible experience of the bilingual programme within the Andalusian project.

Control groups were evaluated alongside bilingual sections. This was facilitated by the school organization system itself, as all the participating secondary and most of the primary schools involve parallel bilingual and mainstream (monolingual) peer group streams. A few of the (smaller) primary schools using English as an L2 had implemented institution-wide programmes, however, and so the total number of English bilingual section learners outnumbered that of the control groups. As French and German projects, which were set up experimentally prior to the publication of the Plan, involve whole schools rather than bilingual sections within otherwise monolingual institutions, it was only possible to include Control groups for English L2.

It should be pointed out at this stage that one of the ways in which the Andalusian project differs from many of its European counterparts is that admission to bilingual sections is open to everyone—there is no pre-testing or screening. When mooted, the idea of testing for admission to bilingual sections was roundly rejected on the grounds of potential elitism. In practical terms, petitions frequently outnumber places and random selection systems are employed. The formation of bilingual section groups is monitored and approved in the Community School Council, a joint parent-teacher-student

Table 1: Number of questionnaires analysed

	Learners				Teachers	Coordinators	Parents
	English	French	German	Total	Total	Total	Total
Primary	389	221	83	693	155	32	531
Secondary	373	201	62	636	243	29	441
TOTAL	762	422	145	1329	398	61	972

Table 2: Breakdown of linguistic evaluation: student numbers

	Linguistic evaluation				
	English		French	German	Total
	Control	Bilingual			Bilingual
Primary	145	380	221	83	684
Secondary	303	374	202	60	636
TOTAL	448	754	423	143	1320

body. Bilingual sections are, therefore, essentially egalitarian (although the possibility of corollaries between social class and parental choice cannot be ignored).

The organisation of classroom groups is also worth noting. In order to minimise the possibility of in-school schisms, legislation was enacted which obliged schools to preserve original classroom groups for everything but the three content subjects taught in the L2 (the choice of subjects varies, depending on teacher profiles). During bilingual section classes the students regroup into temporary bilingual and monolingual streams. This means that learners experience a wider variety of classmates.

Tables 1 and 2 show the final sample size and distribution (for more detailed information on the sample selection and data-gathering, see Casal and Moore 2009).

Instruments

In line with the objectives and scope of the evaluation, the desired degree of accuracy, the proposed timescale and the need to optimise financial resources, a variety of data collection methods were used:

— A set of categorical questionnaires was elaborated and administered to the teaching body, CLIL learners and their parents. These focused on metaconcerns 2, 3 and 4 above: curricular organization, classroom praxis and levels of satisfaction.

- At each institution, the Bilingual programme coordinators were recorded in a structured interview designed to facilitate a SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis. A SWOT analysis is applied in the assessment of complex strategic situations through the analysis of internal (strengths and weaknesses) and external (opportunities and threats) factors. It can serve as an interpretative filter to reduce the information to a selection of key issues relating to project implementation.
- A series of diagnostic tests was employed to assess language competences amongst bilingual and control learners. These tests were skills-based and conjointly designed by native speakers of the three L2s, each of whom is also an external examiner for an ALTE (Association of Language Testers of Europe) member organisation from her particular country. In essence the tests for the three different L2s were adaptations of a single model, elaborated in accordance with descriptors from the CEFR at A1 (primary) and A2 (secondary) levels in combination with national curricula. The tests incorporated a variety of text types (letters, articles, signs, etc.) with diverse functional goals (describing, classifying, informing, giving instructions, etc.) and featuring typical content-related skills such as numeracy and orientation (map-reading) in accordance with the developmental levels of the learners.

Data collection and analysis

The nine-member research team comprised linguists, native-speaker assessors, interviewers and a statistician. After initial piloting and assessor benchmarking, data collection was conducted over a three-month period in the Spring of 2008. In the first instance, teacher, language assistant and parent questionnaires and parental letters of consent for audio-visual recording were sent to all participating schools. A paired team of one assessor and one interviewer then visited each of the schools. Together, they supervised the linguistic assessment and the learner questionnaires in classroom time. At primary level each activity took thirty minutes; at secondary level the linguistic assessment took one hour and the completion of questionnaires thirty minutes. Then, while the interviewer conducted the SWOT analysis with the coordinator, the assessor interviewed a random sub-sample of learners, in pairs, in order to evaluate their speaking skills. The team also collected the previously completed questionnaires. The Plan stipulates that institutions containing bilingual sections must participate in monitoring projects and questionnaire return was high.

The statistical analyses presented in this article are based on the main descriptive figures provided by the questionnaires, alongside the results obtained on the tests. Descriptive statistics were used throughout the data analysis in a number of different ways. First, descriptive statistics were important in data cleaning, ensuring the number of valid cases for each variable and assuring that the 'N' differs slightly between variables. Secondly, descriptive analysis provides a panoramic view of the situation under study. Given the fact that questionnaire data were mainly categorical, frequency analysis was more appropriate for variable types, as this avoided the loss of information which might have resulted from collapsing it into categories. *T*-tests were used to compare means arising from the diagnostic test results.

Taking the above into consideration, we believe that this research meets the four absolute prerequisites for reliability which Cummins stipulated for research focusing on the linguistic assessment of content/immersion learners (1999: 27):

1. Studies must compare students in bilingual programmes to a control group of similar students.
2. The design must ensure that initial differences between treatment and control groups are controlled statistically or through random assignment.
3. Results must be based on standardised test scores.
4. Differences between the scores of treatment and control groups must be determined by means of appropriate statistical tests.

Research questions

This article focuses specifically on those findings which appear to offer significant contributions to key discussions within the contemporary field of European CLIL research. The results here presented address four of the core research questions:

1. *Linguistic outcomes and competence levels*: How do the language competences of CLIL students compare with those of their mainstream peers? If the CLIL learners do show increased gains, to what extent do these differences appear to be the result of language learning based on academic content processing?
2. *Acquisitional routes and individual differences in CLIL programmes*: How do entry points in CLIL programmes affect acquisition? Does CLIL affect conative factors? If so, how?
3. *L2 use in CLIL classrooms*: How can the CLIL language environment be characterised on the basis of different instructional actors' and practitioners' use of the L2 (content teachers, language teachers and native-speaker language assistants)?
4. *CLIL educational effects beyond the L2*: Is CLIL having any visible effect aside from that observed in L2 learning? To what extent is the integrative nature of CLIL impacting on L1 language education? How does a language component integrated in school subjects involve language sensitive organizational patterns in the wider school context?

In short, results from the evaluation project are narrowed down to questions pertaining to the overall results of the linguistic evaluation; the learning process that may be envisaged from competence results; differences in language use among the teaching body in bilingual programmes and the ways in which CLIL impacts upon the educational process. The following four sections tally with the four research questions.

RESULTS AND DISCUSSION

Linguistic outcomes and competence levels

When the results of the linguistic evaluation had been compiled, it emerged that the CLIL learners were clearly outperforming their mainstream peers. Global average scores were 62.1 per cent for the bilingual groups in comparison with 38 per cent for the control groups. Figure 1 presents the results of the linguistic evaluation component incorporating both primary and secondary samples and all three languages. Given the disparity in numbers (see above), it also includes the results only for English L2 (see below for a breakdown of the results across the three L2s). It should here be pointed out that the evaluation procedure comprised four equally weighted tests, corresponding to the four basic skills, and results are therefore presented as marks out of 100. A mark of 50 was interpreted as A1 (primary)/A2 (secondary); 75 implied A1+/A2+ and full marks signalled that the learner was at the next level (A2/B1). (For example, 20.2% of the English L2 secondary CLIL contingent received 25/25 in the spoken component.) For full numerical, statistically confirmed, results see Tables A1 and A2 in the Appendix.

These results demonstrate a clear competence differential between bilingual and control groups, confirmed as significant in the statistical analysis.

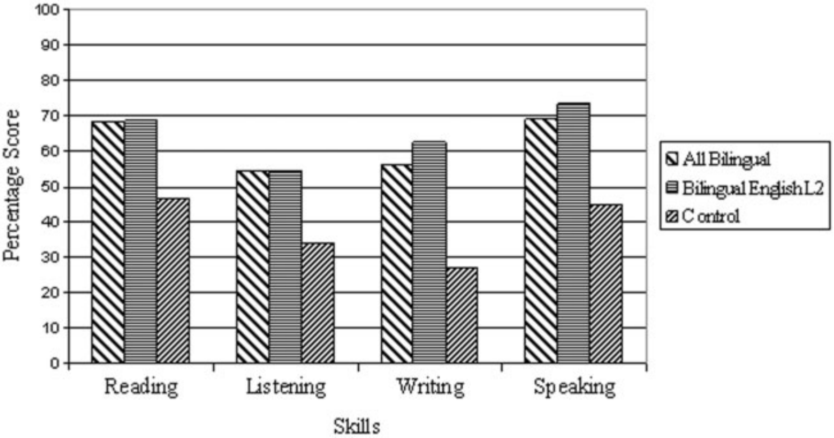

Figure 1: Linguistic evaluation: all bilingual, bilingual English L2 and control

Considering that the only feature which distinguishes these two groups is that the bilingual learners have had one and a half years of CLIL, the difference is striking. As in previous studies (e.g. Burmeister and Daniel 2002) results here demonstrate a non-linear correlation between exposure and competence.

In turn, this gives rise to a need for a closer examination of language competence levels in CLIL settings. It has already been suggested that CLIL engenders a greater lexical range (Dalton-Puffer 2007) and this study suggests that the advantage extends to structural variety and pragmatic efficiency, hence encompassing language growth at lexico-grammatical and discourse levels. To date there has been little comparative research focusing on discrete grammar in bilingual and mainstream language environments although one exception is Järvinen (2005), who explored the acquisition of relativization and found that it appeared to emerge earlier for CLIL students than for their peers in the control groups. Previous research has also demonstrated increased accuracy when production is focused on discourse topics which engage students' attention due to contextual significance, here content-based topics, thereby reflecting the authenticity of the academic domain (Clachar 1999; Butler and Hakuta 2004). This suggests that attention allocation can contribute to the acquisition of lexico-grammar while processing academic content in CLIL-type contexts.

The same proactive engagement with language is in evidence at the level of discourse pragmatics. CLIL learner L2 output features rhetorical moves and discourse patterns such as hedging and tentative language, hypothesising, impersonal structures and metaphorical grammar, typical of academic discourse but not addressed within primary or early secondary L2 syllabi. This suggests a considerable degree of positive transfer in the manipulation and maintenance of cohesion and coherence (Lorenzo and Moore, forthcoming). This is also consistent with studies in bilingual scenarios where academic functions such as formal definitions and picture descriptions have been found to lend themselves to cross-language transfer (Bialystok 2004).

The data and cross-references discussed above may contribute to preliminary steps towards the formation of a theory of learning in CLIL scenarios. Apart from increased exposure it is likely that other factors contribute, chief among them cognitive considerations surrounding cognitive inhibition (Bialystok 2005) and the in-depth processing of language stimuli which appears to result from attention to meaningful input (Lee and VanPatten 2003; Kroll and De Groot 2005; Wong 2005). In CLIL scenarios, this is facilitated through the embedding of target language in contextualised subject matter materials—thereby providing significant semantic scaffolding. A *primacy of meaning principle* operating in real and authentic L2 use would appear to be the norm in formal CLIL settings. This is likely bolstered by conative questions relating to the corollary effects of increases in motivation caused by significant learning environments like CLIL programmes (see below).

If a theory of learning is proposed, the I in CLIL—Integration—demands that the question of a theory of language also be addressed. In light of the degree of competence observed in the results, this paper holds that language theories which favour the concept of language as semiosis may render a more adequate analysis of language integrated with content. Functional systemic principles examining the cognitive outcomes of content and language integration may be more explanatory of the true nature of the language, or to be more precise of the interlanguage, revealed in the linguistic evaluation (Mohan and Beckett 2003; Mohan and Slater 2005). Functional approaches would claim that what is required is a clear concept of semantics as a layer for language structuring in language education. This belief would appear to apply to CLIL and in turn may serve to strengthen it as a language approach. (See Halliday and Hasan (2006) for a recent discussion of the origins of functional systemics and Mohan and Slater (2005) for a review of controversy in content and language integrated models as opposed to focus on form models.)

Acquisitional routes and individual differences in CLIL programmes

Figure 2 sets out the results for the three L2s of the bilingual evaluation project (and see Table A3 in the Appendix). It shows that the French learners obtained marginally higher scores for receptive skills and the English learners for productive skills. Nonetheless, globally speaking, the average scores for the three languages are comparable within the diagnostic levels of the CEFR.

It should be borne in mind, however, that the English learners have had but one and a half years of CLIL instruction, while the French and German learners have been in bilingual programmes since the beginning of primary education. While it is possible that the English L2 sections have benefited from the insights obtained over the course of the earlier French and German

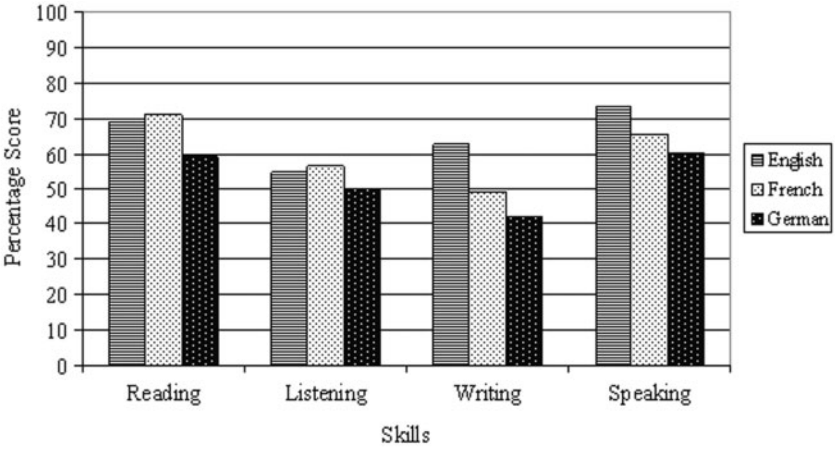

Figure 2: Comparison of average scores in the three L2 (English, French and German)

experimental schemes and/or that global English as a *lingua franca* inspires more productive attitudinal and motivational stances than its continental neighbours (see below), the parity in results also opens the door to the perennial 'age factor' debate.

Logistically speaking, bilingual education can be early, middle or late start and much research has been directed at comparative evaluations (for a useful overview, see Genesee 2004). There has been significant discussion, within psycholinguistic and second language acquisition fields regarding the advantages of early starts (Muñoz 2006, 2008; Nikolov and Mihaljević Djigunović 2006) and promising findings within neurologically-oriented research into cerebral development seem to imply cognitive benefits for early bilingualism (Van de Craen *et al*. 2007). Nonetheless, the results here presented appear to imply that, in CLIL programmes, middle or late introduction can result in competences similar to those obtained in early introduction. It is also worth pointing out here that other studies have found similar advantages for late and low frequency programmes (see Wesche 2002 on the former and Marsh 2002 on the latter). This may be attributable to the fact that increasing cognitive and meta-cognitive abilities and more advanced L1 academic proficiency—as typical of later primary or early secondary learners—can offset the neurologically psycholinguistic advantages of an early start. It also seems logical that the quality and quantity of input/exposure be just as important as age (Muñoz 2008), and CLIL implies both more and more meaningful L2. If subsequent research continues to demonstrate potential for later starts, it is likely to significantly aid the CLIL cause. Decisions regarding start points for bilingual programmes are ultimately framed by budgetary considerations and implementing full CLIL at early primary can be costly. The results here presented suggest that later starts, on condition that they are framed within a sound manipulation of exposure time, can optimise resources.

Results also indicate that CLIL may offer a solution to the long-standing problem of disaffection in foreign, particularly non-world, language learning in European secondary schools (Dörnyei and Csizer 2002). Attainment levels demonstrate that motivational processes in CLIL-type learning differ from mainstream foreign language learning. Research into motivation posits that, in instructed L2 learning, integrativeness—one of the key constructs in goal-oriented behaviour—has little to do with inter-ethnic contact. Nonetheless, the likelihood of exchanges with native speakers is considered key in communicative approaches to foreign language teaching. In CLIL scenarios, however, the identification process between students and the language rests upon the link between language and subject matter, rather than on some nebulous future need. In other words, when French is the language of the history lessons, this supersedes the view of it as the language of the French nation. Satisfaction and engagement levels, as reported in the learner opinion questionnaires, seem to support this interpretation (and see Merisuo-Storm 2007; Seikkula-Leino 2007). What the results obtained seem to imply is that when the learning situation inculcates an identification process between learner and language, and this results in a revision of learner self-concept, both high motivation levels and successful competence outcomes can be achieved. It follows that a theoretical model of bilinguality for CLIL should be aligned with socio-educational models of bilingual acquisition (Masgoret and Gardner 2003) rather than with models where factors unrelated to the language learning situation (such as ethno-linguistic vitality) are highlighted (for a review of models of bilinguality, see Bourhis 1990).

L2 use in CLIL classrooms

This type of research needs to be wary of a tendency to over-rely on quantitatively formulated evaluations of bilingual education, frequently based on learner test scores, to the detriment of more qualitatively oriented explorations of praxis (Leung 2005). Regarding L2 use, the teacher questionnaires were interested not only in the amount of time spent using the L2 in CLIL classrooms but also in pedagogic questions concerning stages of the lesson and functional questions exploring the type of language employed. This section will briefly review findings relating to each of these concerns.

To begin with, Figure 3 presents the data relating to the quantitative use of the L2 in the classroom distinguishing between primary and secondary and between teachers and language assistants.

Regarding the pedagogic question of staging, the teacher questionnaires focused on six key stages in CLIL teaching: *Introducing the Topic, Conducting Activities, Clarifying and Dealing with Problems, Providing Feedback and Evaluation, Conducting Consolidation and Revision* and *Making Links to Other*

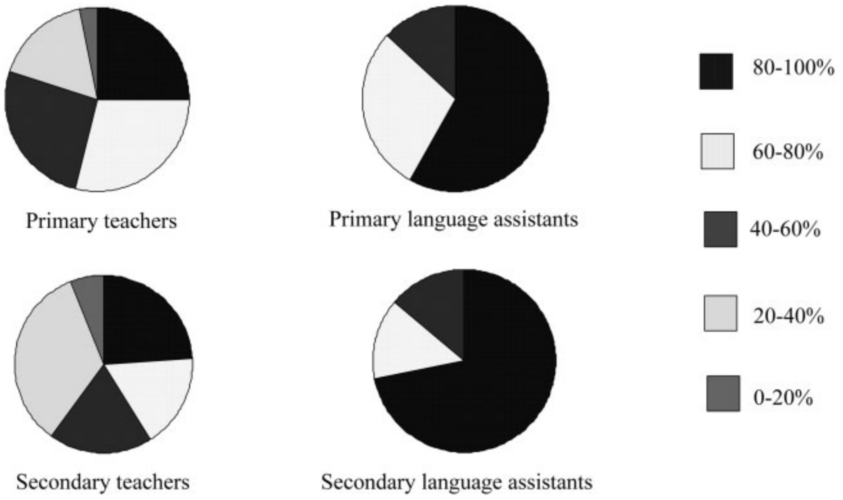

Figure 3: L2 use as percentage of classroom time

Table 3: L2 Use and classroom stages (secondary teachers) (figures in percentages)

	Frequency of L2 use	Topic introduction	During activities	Clarify and deal with problems	Feedback and evaluation	Consolidation and revision	Making links to other areas
Content teachers	Always	16.8	26.5	1.8	15.9	17.7	1.8
	Often	28.3	52.2	17.7	34.5	45.1	30.1
	Sometimes	38.1	20.4	52.2	42.5	32.7	52.2
	Never	15.0	0.0	26.5	5.3	2.7	12.4
L2 teachers	Always	53.7	50.0	7.4	40.7	31.5	20.4
	Often	27.8	42.6	33.3	38.9	50.0	55.6
	Sometimes	11.1	0.0	50.0	11.1	13.0	13.0
	Never	1.9	1.9	3.7	1.9	0.0	0.0
Language assistants	Always	76	73.3	43.3	53.5	60.0	53.5
	Often	13.3	20.0	23.3	20.0	26.7	20.0
	Sometimes	6.7	6.7	23.3	10.0	10.0	20.0
	Never	3.3	0.0	10.0	10.0	0.0	3.3

Areas and respondents were asked to signal the frequency of their L2 use in each case. Table 3 provides a breakdown of results for secondary teachers (figures for primary were comparable).

The results show that content teachers are more likely to employ the L2 in explicitly content-centred teaching: during activities, consolidation and revision and to a lesser degree topic introduction, an aspect which appears to be shared between content and language teachers. In general, language teachers are more likely to use the L2 for feedback and evaluation than their subject specialist counterparts. Overall the stage which is least likely to involve L2 usage is that of clarifying and dealing with problems; even language assistants, who otherwise prefer to maximise target language use, are less likely to use the L2 in this scenario.

The section dealing with functional aspects of use focused on five macro discourse areas: *Formulaic Language*; *Giving Instructions for Activities*; *Telling Anecdotes*; *Error Correction* and *Classroom Management*. Table 4 presents the results for primary teachers (again the figures for secondary were comparable).

As might be expected, all three teacher categories tend to use the L2 when it comes to formulaic language, the language specialists employing the L2 more than 70 per cent of the time. There is also an overall tendency to use the L2 for

Table 4: Functional aspects of L2 use (primary teachers) (figures in percentages)

	Frequency of L2 use	Formulaic language	Giving instructions for activities	Telling anecdotes	Error correction	Classroom management
Content teachers	Always	36.3	19.5	3.5	8.8	12.4
	Often	44.2	51.3	13.3	27.4	36.3
	Sometimes	15.9	29.2	42.5	48.7	44.2
	Never	2.7	0.0	38.9	14.2	7.1
L2 teachers	Always	79.6	50.0	16.7	18.5	40.7
	Often	13.0	40.7	31.5	48.1	40.7
	Sometimes	1.9	0.0	38.9	25.9	11.1
	Never	0.0	1.9	5.6	0.0	1.9
Language assistants	Always	73.3	56.7	50.0	46.7	46.7
	Often	23.3	33.3	16.7	36.7	20.0
	Sometimes	3.3	10.0	20.0	13.3	23.3
	Never	0.0	0.0	10.0	3.3	6.7

classroom management and setting up activities, although the content teachers seem to alternate more frequently. Overall, respondents report that they use the L2 in error correction around half of the time. When it comes to the recounting of anecdotes, however, teachers report less L2 use. Regarding the use of more colloquial language, as implicit in the telling of anecdotes, it is interesting that a Dutch survey into CLIL content teacher attitudes also found that they were least comfortable in this domain (Wilkinson 2005). It has been suggested that native-speaker language specialists engage in more conversational face-to-face exchanges than non-native content teachers (Dalton-Puffer and Nikula 2006) and this would seem to be one of the areas where the language assistants of the Andalusian programme are proving their worth.

Discourse analysis has repeatedly demonstrated the rigid hierarchy typical of classroom discourse and its roles (Sinclair and Coulthard 1975; Markee 2000). CLIL teaching, however, does not conform to the stereotypical educational scenario: while the latter is monolingual, focused on one subject at a time and fronted by a sole teacher, CLIL is bilingual, intertwines subjects and is co-taught. Taking a sociolinguistic stance, and positing CLIL as a community of practice, the question of teacher classroom roles can be perceived as a triadic symbiosis. Regarding the L1/L2 mix from a more quantitatively inclined perspective, the three instructional actors of the CLIL classroom appear to be providing a range of bilingual experiences: the language assistants come close to providing full immersion, the language teachers represent semi-immersion and the content teachers apply judicious code-switching. In tandem, and further evidenced by results detailing the types of materials and activities that the content and language teachers each use more frequently, it seems that each is dealing with a specific area of language expertise: the language assistants foster conversational style language, the language teachers focus on sentence-level grammar and the content teachers work at the textual level. If this observation holds, it means that CLIL has the potential to provide an extremely rich language learning environment.

CLIL educational effects beyond the L2

There is widespread agreement among bilingual section teaching staff (including L1 teachers and coordinators) that CLIL is beneficial to the educational process in general, an opinion echoed by parents and learners alike. Teacher questionnaires examined this aspect in more detail and demonstrate that the consensus appears to be that some aspects benefit more than others. Tables 5 and 6 provide a more detailed breakdown of attitudes in primary and secondary sectors.

As can be seen above, there is a general consensus that CLIL enhances cohesion within schools. One of the greatest challenges for bilingual education undoubtedly lies in the successful integration of language and content. In order for such a venture to succeed, it is vital that it be operating at both curricular planning (top-down) and classroom praxis (bottom-up) levels. It is therefore significant that the teaching body as a whole considers that interdepartmental cooperation and cohesion is improved in bilingual sections. Coordinator interviews and teacher questionnaires revealed that teacher involvement in CLIL planning is high and characterised by engaged collaboration between content and L2 teachers and language assistants. Aside from European models designed specifically for CLIL (Coyle 1999; Lorenzo 2007) teachers have looked to North American and Australian experiences with minority language learners and sheltered instruction for insights (see for example, Short 1993; Brisk 1998; Swain 2000; Carder 2008).

Turning to the question of content, we find it promising that CLIL appears to be contributing to new forms of language awareness among both content and language teachers. The fact that CLIL involves content and language teachers working together to design and plan integrated lessons has led to a heightened appreciation of the interface between content and language. This appears to be leading content teachers to an acknowledgement both of the ubiquitous nature of language and to the fact that the successful transmission of subject matter content relies heavily on its linguistic selection and grading.

Table 5: The degree and nature of change implied in CLIL on a series of educational aspects—results from primary institutes (figures in percentages)

Educational aspect	Degree of change					Nature of change				
	None	Minimal	Moderate	Significant	No reply	Much worse	Worse	Better	Much better	No reply
Subject area objectives	11	14	42	12	22	1	0	53	8	38
Methodologies	4	8	39	28	20	0	1	49	19	31
Content focus	5	18	45	12	21	0	1	52	14	34
L1 learning	12	21	29	11	28	0	2	45	9	43
L2 learning	1	3	23	48	25	0	0	35	35	29
Content learning	2	8	45	24	20	0	1	51	18	31
Classroom (peer) cohesion	7	25	22	22	25	2	2	41	14	42
Interdepartmental cohesion	3	14	33	30	20	1	2	48	20	29

Table 6: The degree and nature of change implied in CLIL on a series of educational aspects—results from secondary institutes (figures in percentages)

Educational aspect	Degree of change					Nature of change				
	None	Minimal	Moderate	Significant	No reply	Much worse	Worse	Better	Much better	No reply
Subject area objectives	6	24	48	11	11	0	1	66	5	28
Methodologies	1	12	52	26	8	0	1	68	14	16
Content focus	6	24	50	11	10	0	3	62	8	26
L1 learning	6	18	35	15	26	0	1	46	11	41
L2 learning	0	2	25	52	20	0	0	39	36	25
Content learning	3	13	35	29	20	0	4	47	17	32
Classroom (peer) cohesion	2	10	37	39	12	0	4	49	25	21
Interdepartmental cohesion	1	8	41	42	8	0	2	55	28	15

In turn, language teachers are becoming aware that planning for advanced literacy is just as important as basic communicative L2. The gains reported in content focus, content learning and subject area objectives can be attributed to this increase in coherence. Nevertheless, it became apparent that many language teachers are still attempting to align language structures with content in a somewhat erratic manner (no doubt a legacy of their structurally biased professional development) and this area remains fuzzy.

From a language learning perspective, the tables above demonstrate that there is wide consensus regarding the benefits which CLIL implies for L2 learning. When it comes to the L1, however, both coordinator interviews and questionnaires administered to the L1 teachers suggest that CLIL tends to be regarded primarily as a means to improve second (foreign) language development. L1 teachers appear reluctant to participate in integration; some even considered CLIL a competitor to L1 learning, in the belief—nourished by a pedestrian view of bilingualism—that different languages represent opposing forces, growing at each other's expense. Not only does such a view pose risks to the entire education system, partisan attitudes amongst language departments also pose a serious hurdle to successful CLIL implementation, as they represent an overly narrow interpretation of an approach which offers much wider potential.

In the situated context of this research, CLIL implies a new language model and it both coincides with and has contributed to a move away from the *ars gramatica* and towards a genre-based approach to language study—all language study, be it first or subsequent languages—which is not restricted to Andalusia (Bhatia 2004; Martin 2004; Hyland 2008). This conflates with the concept of *Language Across the Curriculum* (LAC) a movement which, although quashed by political opposition in the 1970s when it first emerged (Stubbs 2000), has recently been enjoying something of a renaissance and is currently being actively promoted by European language planning agencies (Vollmer 2006; Beacco and Byram 2007).

CONCLUSION

This article began by outlining the renaissance of European educational bilingualism under the contemporary banner of CLIL (Content and Language Integrated Learning). It then introduced the Andalusian *Plan to Promote Plurilingualism*, within which the research here presented and discussed was conducted. A brief review of European CLIL research helped to position the research project within a wider continental ambit. The four primary meta-concerns which shaped the research—competence development; curricular organisation; classroom praxis and levels of satisfaction—were then outlined and clarified. The section dealing with methodological questions covered participants, instruments and data collection and analysis. The results here presented narrow the focus to four key research questions which we believe are of significant import in current European CLIL-related research: Linguistic outcomes and competence levels; acquisitional routes and individual differences; L2 use in CLIL classrooms; and educational effects beyond the L2. Findings relating to each of these areas were presented and discussed.

In isolation, several of the questions addressed above offer significant contributions to current Applied Linguistics research: confirmation that CLIL learners show greater gains than their monolingual peers; the evidence regarding incidental learning and positive transfer through content-focused instruction; the fact that later start learners are demonstrating competences comparable with early start learners and the observation that team teaching between content and language specialists is providing for a wider range of discourse input are all relevant in the contemporary arena. In conjunction, however, these results suggest that CLIL is an approach which may hold significant potential for European education planning. Not only does it promote the integration of content and language, CLIL also fosters greater inter-departmental collaboration and conflates with other language development initiatives such as Language Across the Curriculum, the genre-based approach and multi-disciplinary curricula.

In essence CLIL has evolved as, and still remains, a grassroots initiative: *A European solution to a European need*. This has, however, left it bereft of sound supporting theory regarding the nature of language and the nature of its acquisition. On the basis of empirical results, this article has attempted to establish some primary connections between observed CLIL learning outcomes and existing and robust linguistic and learning theories. This should be interpreted as a work-in-progress and future descriptive research will contribute to this task.

As a final point it should be noted that while some of the results obtained in the research here discussed coincide with claims made for CLIL at other latitudes in the continent, it is still too early to infer any generalised outcomes for European CLIL. It is possible that, in the long term, CLIL-type initiatives might contribute to the formulation of a common European ideology of language. Such a paradigm would, of necessity, be rooted in the historical tradition of educational multilingualism in the continent. Where it was once believed that the quintessential cultural endeavour of Europe across time lay in the search for the perfect language (Eco 1995), this quest is now considered utopian and dated; nowadays the goal has become the propagation of plurilingual competences and multicultural values and CLIL may well have a significant contribution to make in this endeavour.

ACKNOWLEDGEMENTS

We would like to thank Leticia Rodríguez for her invaluable help in the statistical analysis of this research. We would also like to thank the three anonymous reviewers for their insightful feedback on our first version of this article.

APPENDIX

Table A1: The overall results of the linguistic evaluation

Skill	Group	Number of cases	Mean	Standard deviation	Significance value *t*-test	95% confidence interval of the mean	
Reading	Bilingual	1320	68.50	17.17			
	Control	448	46.26	22.14	0.00***	19.98	24.49
Listening	Bilingual	1274	54.71	16.37			
	Control	421	33.94	17.77	0.00***	18.93	22.61
Writing	Bilingual	1295	56.24	26.70			
	Control	446	26.82	29.99	0.00***	26.27	32.57
Speaking	Bilingual	348	69.07	20.25			
	Control	119	45.04	20.92	0.00***	19.77	28.29

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table A2: Results of the English L2 evaluation

Skill	Group	Number of cases	Mean	Standard deviation	Significance value *t*-test	95% confidence interval of the mean	
Reading	Bilingual	754	68.90	16.99			
	Control	448	46.26	22.14	0.00***	20.25	25.02
Listening	Bilingual	731	54.59	17.03			
	Control	421	33.94	17.77	0.00***	18.58	22.73
Writing	Bilingual	752	62.71	26.82			
	Control	446	26.82	29.99	0.00***	32.51	39.28
Speaking	Bilingual	186	73.46	21.11			
	Control	119	45.04	20.92	0.00***	23.56	33.28

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table A3: Comparative results for English, French and German bilingual sections in the three L2

Skill	Group	Number of cases	Mean	Standard deviation	Significance value *t*-test	95% confidence interval of the mean	
Reading	English	754	68.90	16.99			
	French	423	70.90	15.93	0.05*	−3.98	−0.02
Listening	English	731	54.59	17.03			
	French	400	56.53	14.45	0.04*	−3.82	−0.06
Writing	English	752	62.71	26.82			
	French	400	49.13	23.43	0.00***	10.59	16.59
Speaking	English	186	73.46	21.11			
	French	120	65.40	16.52	0.00***	3.81	12.31
Reading	English	754	68.90	16.99			
	German	143	59.32	18.68	0.00***	6.48	12.67
Listening	English	731	54.59	17.03			
	German	143	50.20	17.17	0.00**	1.34	7.46
Writing	English	752	62.71	26.82			
	German	143	42.13	24.10	0.00***	15.86	25.32
Speaking	English	186	73.46	21.11			
	German	42	60.10	21.43	0.00***	6.24	20.49
Reading	French	423	70.90	15.93			
	German	143	59.32	18.68	0.00***	8.14	15.01
Listening	French	400	56.53	14.45			
	German	143	50.20	17.17	0.00***	3.17	9.50
Writing	French	400	49.13	23.43			
	German	143	42.13	24.10	0.00**	2.48	11.52
Speaking	French	120	65.40	16.52			
	German	42	60.10	21.43	0.15	−1.97	12.58

$^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

NOTES

[1] An English version of the Plan is available at: http://www.juntadeandalucia.es/averroes/html/portal/com/bin/contenidos/B/InnovacionEInvestigacion/ProyectosInnovadores/Plurilinguismo/Portada/1182945265640_wysiwyg_planing.pdf

REFERENCES

Ackerl, C. 2007. 'Lexico-Grammar in the essays of CLIL and non-CLIL students: error analysis of written production,' *ViewZ* (Vienna English Working papers) 16/3: 6–11. URL: http://www.univie.ac.at/Anglistik/Views_0703.pdf. Last accessed 29th October 2009.

Adams, J. N. 2003. *Bilingualism and the Latin Language*. Cambridge University Press.

Admiraal, W., G. Westhoff, and **K. de Bot.** 2006. 'Evaluation of bilingual secondary education in the Netherlands: students' language proficiency in English,' *Educational Research and Evaluation* 12/1: 75–93.

Ager, D. 2001. *Motivation in Language Planning and Language Policy*. Multilingual Matters.

Arnau, J. and **J. Artigal.** (eds) 1998. *Immersion Programs. A European Perspective*. Universidad de Barcelona.

Baetens Beardsmore, H. (ed.) 1993. *European Models of Bilingual Education*. Multilingual Matters.

Beacco, J.-C. and **M. Byram.** 2007. *Guide for the Development of Language Policies in Europe – from Linguistic Diversity to Plurilingual Education*. Language Policy Division, Council of Europe. URL: http://www.coe.int/t/dg4/linguistic/Source/Guide_Main_Beacco2007_EN.doc. Last accessed 29th October 2009.

Bhatia, T. K. and **W. C. Ritchie.** (eds) 2004. *The Handbook of Bilingualism*. Blackwell.

Bhatia, V. K. 2004. *Worlds of Written Discourse: A Genre-based View*. Continuum.

Bialystok, E. 2004. 'The impact of bilingualism on language and literacy development' in T.K. Bhatia and W.C. Ritchie (eds): *The Handbook of Bilingualism*. Blackwell, pp. 577–601.

Bialystok, E. 2005. 'Consequences of bilingualism for cognitive development' in J.F. Kroll and A. De Groot (eds): *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, pp. 417–33.

Bourhis, R. 1990. 'Social and individual factors in language acquisition: some models of bilingual proficiency' in B. Harley, P. Allen, J. Cummins, and M. Swain (eds): *Development of Second Language Proficiency*. Cambridge University Press, pp. 134–45.

Braunmüller, K. and **G. Ferraresi.** 2003. *Aspects of Multilingualism in European Language History*. John Benjamins.

Brisk, M. E. 1998. *Bilingual Education: From Compensatory to Quality Schooling*. Lawrence Erlbaum Associates.

Burmeister, P. and **A. Daniel.** 2002. 'How effective is late partial immersion? Some findings of a secondary school program in Germany' in P. Burmeister, T. Piske, and A. Rohde (eds): *An Integrated View of Language Development: Papers in Honour of Henning Wode*. Wissenschaftlicher Verlag Traer, pp. 499–512.

Burmeister, P., T. Piske, and **A. Rohde.** (eds) 2002. *An Integrated View of Language Development: Papers in Honour of Henning Wode*. Wissenschaftlicher Verlag Traer.

Butler, Y. G. and **K. Hakuta.** 2004. 'Approaches to bilingualism and second language acquisition' in T. K. Bhatia and W. C. Ritchie (eds): *The Handbook of Bilingualism*. Blackwell, pp. 114–46.

Carder, M. 2008. 'The development of ESL provision in Australia, Canada, the USA and England, with conclusions for second language models in international schools,' *Journal of Research in International Education* 7/2: 205–31.

Casal, S. and **P. Moore.** 2009. 'The Andalusian bilingual sections scheme: evaluation and consultancy,' *International CLIL Research Journal* 1/2: 36–46.

Clachar, A. 1999. 'It is not just cognition: the effect of emotion on multiple level discourse processing in second language writing,' *Language Sciences* 21: 31–60.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press. URL: http://www.coe.int/T/DG4/Linguistic/Source/Framework_EN.pdf. Last accessed 29th October 2009.

Coyle, D. 1999 'Teacher education for multilingual education: a CLIL teacher training curriculum.' Paper presented at *The Multilingual Challenge* conference held in Brussels (Thematic Network Project in the Area of Languages – Sub project 6: Language Teacher Training and Bilingual Education), pp. 70–9.

Cummins, J. 1999. 'Alternative paradigms in bilingual education research: Does theory have a place?,' *Educational Researcher* 28/7: 26–32.

Dalton-Puffer, C. 2007. *Discourse in Content and Language Integrated Classrooms*. John Benjamins.

Dalton-Puffer, C. 2008. 'Outcomes and processes in content and language integrated learning (CLIL): Current research from Europe' in W. Delanoy and L. Volkmann (eds): *Future Perspectives for English Language Teaching*. Carl Winter, pp. 139–57.

Dalton-Puffer, C. and **T. Nikula.** 2006. 'Pragmatics of content-based instruction: teacher and student directives in Finnish and Austrian classrooms,' *Applied Linguistics* 27/2: 241–67.

de Mejía, A. M. 2002. *Power, Prestige and Bilingualism*. Multilingual Matters.

Dörnyei, Z. and **K. Csizer.** 2002. 'Some dynamics of language attitudes and motivation: results of a longitudinal nationwide survey,' *Applied Linguistics* 23/4: 421–62.

Eco, U. 1995. *The Search for the Perfect Language (the Making of Europe)*. Blackwell.

European Commission. 1995. *White Paper on Education and Training.* URL: http://ec.europa.eu/education/doc/official/keydoc/lb-en.pdf. Last accessed 29th October 2009.

Eurydice. 2006. *Content and Language Integrated Learning at School in Europe*. Eurydice European Unit.

Genesee, F. 2004. 'What do we know about bilingual education for majority language students?' in T.K. Bhatia and W.C. Ritchie (eds): *The Handbook of Bilingualism*. Blackwell, pp. 547–76.

Halliday, M. A. K. and **R. Hasan.** 2006. 'Retrospective on SFL and literacy' in R. Whittaker, M. O'Donnell, and A. McCabe (eds): *Language and Literacy: Functional Approaches*. Continuum, pp. 15–45.

Housen, A. 2002. 'Processes and outcomes in the European schools model of multilingual education,' *Bilingual Research Journal* 26/1: 45–64.

Hyland, K. 2008. 'Genre and academic writing in the disciplines,' *Language Teaching* 41: 543–62.

Jäppinen, A.-K. 2005. 'Thinking and content learning of mathematics and science as cognitional development in content and language integrated learning (CLIL): teaching through a foreign language in Finland,' *Language and Education* 19/2: 148–69.

Järvinen, H.-M. 2005. 'Language learning in content-based instruction' in A. Housen and M. Pierrard (eds): *Investigations in Instructed Second Language Acquisition*. Mouton de Gruyter, pp. 433–56.

Johnson, R. K. 1997. 'The Hong Kong education system: late immersion under stress' in R.K. Johnson and M. Swain (eds): *Immersion Education: International Perspectives*. Cambridge University Press, pp. 171–89.

Johnson, R. K. and **M. Swain** (eds) 1997. *Immersion Education: International Perspectives*. Cambridge University Press.

Junta de Andalucía. 2005. *Plan de Fomento del Plurilingüismo en Andalucía*. Junta de Andalucía. URL: http://www.juntadeandalucia.es/boja/boletines/2005/65/d/5.html. Last accessed 29th October 2009.

Kroll, J. F. and **A. de Groot.** 2005. *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press.

Lamsfuß-Schenk, S. 2002. 'Geschichte und Sprache – ist der bilinguale Geschichtsunterricht der Königsweg zum Geschichtsbewusstsein?' in S. Breidbach, G. Bach and D. Wolff (eds): *Bilingualer Sachfachunterricht: Didaktik, Lehrer-/Lernerforschung und Bildungspolitik zwischen Theorie und Empirie*. Peter Lang, pp. 191–206.

Lee, J. and **B. VanPatten.** 2003. *Making Communicative Language Teaching Happen.* 2nd edition. McGraw-Hill.

Leung, C. 2005. 'Language and content in bilingual education,' *Linguistics and Education* 16: 238–52.

Lewis, E. G. 1976. 'Bilingualism and bilingual education: the ancient world to the renaissance' in J.A. Fishman (ed.): *Bilingual Education: An International Sociological Perspective*. Newbury House, pp. 150–200.

Lorenzo, F. 2007. 'An analytical framework of language integration in L2 content-based courses,' *Language and Education* 21: 503–13.

Lorenzo, F. and **P. Moore.** Forthcoming. 'On the natural emergence of language structures in CLIL. Towards a language theory of European educational bilingualism' in C. Dalton-Puffer, T. Nikula, and U. Smit (eds): *Language Use in Content-and-language Integrated Learning (CLIL)*. John Benjamins.

Markee, N. 2000. *Conversation Analysis*. Lawrence Erlbaum Associates.

Marsh, D. 2002. *Content and Language Integrated Learning. The European Dimension*. University of Jyväskyla Press.

Martin, J. R. 2004. 'Genre and literacy: modelling context in educational linguistics' in D. Wray (ed.): *Literacy: Major Themes in Education*. Taylor and Francis, pp. 297–326.

Masgoret, A.-M. and **R. C. Gardner.** 2003. 'Attitudes, motivation, and second language learning: a meta-analysis of studies conducted by Gardner and associates,' *Language Learning* 53: 123–63.

Merisuo-Storm, T. 2007. 'Pupils' attitudes towards foreign-language learning and the development of literacy skills in bilingual education,' *Teaching and Teacher Education* 23/2: 226–35.

Mewald, C. 2007. 'A comparison of oral foreign language performance of learners in CLIL and mainstream classes at lower secondary level in Lower Austria' in C. Dalton-Puffer and U. Smit (eds): *Empirical Perspectives on CLIL Classroom Discourse*. Peter Lang, pp. 139–73.

Mohan, B. 1986. *Language and Content*. Addison Wesley.

Mohan, B. and **G. Beckett.** 2003. 'A functional approach to research on content-based language learning: recasts in causal explanations,' *The Modern Language Journal* 87/3: 421–32.

Mohan, B. and **T. Slater.** 2005. 'A functional perspective on the critical "theory/practice" relation in teaching language and science,' *Linguistics and Education* 16: 151–72.

Muñoz, C. (ed.) 2006. *Age and the Rate of Foreign Language Learning*. Multilingual Matters.

Muñoz, C. 2008. 'Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning,' *Applied Linguistics* 29: 578–96.

Nikolov, M. and **J. Mihaljević Djigunović.** 2006. 'Recent research on age, second language acquisition and early foreign language learning,' *Annual Review of Applied Linguistics* 26: 234–60.

Pakir, A. 1993. 'Two tongue-tied: bilingualism in Singapore,' *Journal of Multilingual and Multicultural Development* 14: 73–90.

Rodgers, D. 2006. 'Developing content and form: encouraging evidence from Italian content-based instruction,' *The Modern Language Journal* 90/3: 373–86.

Seikkula-Leino, J. 2007. 'CLIL learning: achievement levels and affective factors,' *Language and Education* 21/4: 328–41.

Serra, C. 2007. 'Assessing CLIL in primary school: a longitudinal study,' *The International Journal of Bilingual Education and Bilingualism* 10/5: 582–602.

Short, D. J. 1993. 'Assessing integrated language and content instruction,' *TESOL Quarterly* 27/4: 627–56.

Sinclair, J. and **M. Coulthard.** 1975. *Towards an Analysis of Discourse*. Oxford University Press.

Snow, M. A. 1998. 'Trends and issues in content-based instruction,' *Annual Review of Applied Linguistics* 18: 243–67.

Stohler, U. 2006. 'The acquisition of knowledge in bilingual learning: an empirical study on the role of content in language learning,' *ViewZ* (Vienna English Working Papers) 15/3: 41–6. URL: http://www.univie.ac.at/Anglistik/views15_3_clil_special.pdf. Last accessed 29th October 2009.

Stryker, S. B. and **B. Leaver** (eds) 1997. *Content-based Instruction in Foreign Language Education: Models and Methods*. Georgetown University Press.

Stubbs, M. 2000. 'Society, education and language: the last 2,000 (and the next 20?) years of language teaching' in H. Trappes-Lomax (ed.): *Change and Continuity in Applied Linguistics*. BAAL and Multilingual Matters, pp. 15–34.

Swain, M. 2000. 'French immersion research in Canada: recent contributions to SLA and applied linguistics,' *Annual Review of Applied Linguistics* 20: 199–212.

Tollefson, J. W. 2002. *Language Policies in Education. Critical Issues*. Lawrence Erlbaum.

Van de Craen, P., E. Ceuleers, and **K. Mondt.** 2007. 'Cognitive development and bilingualism in primary schools: teaching maths in a CLIL environment' in D. Marsh and D. Wolff (eds): *Diverse Contexts – Converging Goals. CLIL in Europe*. Peter Lang, pp. 185–200.

Vollmer, H. (ed.) 2006. *Language across the Curriculum.* Language Policy Division, Council of Europe. URL: http://www.coe.int/t/dg4/linguistic/Source/Prague07_LangCom_VollmerEd_EN.doc. Last accessed 29th October 2009.

Vollmer, H. 2008. 'Constructing tasks for content and language integrated assessment' in J. Eckerth and S. Siekmann (eds): *Research on Task-based Language Learning and Teaching. Theoretical, Methodological and Pedagogical Perspectives*. Peter Lang, pp. 227–90.

Wesche, M. 2002. 'Early French immersion: how has the original Canadian model stood the test of time?' in P. Burmeister, T. Piske and A. Rohde (eds): *An Integrated View of Language Development: Papers in Honour of Henning Wode.* Wissenschaflicher Verlag Traer, pp. 357–79.

Wilkinson, R. 2005. 'The impact of language on teaching content: views from the content teacher.' Paper presented at the *Bi and Multilingual Universities – Challenges and Future Prospects Conference*. Helsinki, 2 September 2005. URL: http://www.palmenia.helsinki.fi/congress/bilingual2005/presentations/wilkinson.pdf. Last accessed 29th October 2009.

Wong, W. 2005. *Input Enhancement: From Theory and Research to the Classroom*. McGraw-Hill.

Zydatiß, W. 2007. *Deutsch-Englische Züge in Berlin (DEZIBEL). Methoden und Ergebnisse des Evaluationsprojekts zum bilingualen Sachfachunterricht an Gymnasien*. Lang.

Applied Linguistics: 31/3: 443–464 © Oxford University Press 2009
doi:10.1093/applin/amp059 Advance Access published on 23 December 2009

The Role of Language Aptitude in First Language Attrition: The Case of Pre-pubescent Attriters

EMANUEL BYLUND, NICLAS ABRAHAMSSON and KENNETH HYLTENSTAM

Centre for Research on Bilingualism, Stockholm University

While language aptitude has been investigated actively within second language research, there is a current dearth of research on the effects of aptitude in cases of attrition. The aim of the present investigation was to explore the role of language aptitude for L1 proficiency in speakers who experienced a break with their L1 setting prior to puberty. Twenty-five L1 Spanish—L2 Swedish bilinguals residing in Sweden participated in the study, and 15 native speakers of Spanish living in Chile were recruited as controls. The L1 proficiency was measured by means of a grammaticality judgement test (GJT) and language aptitude data were obtained through the *Swansea Language Aptitude Test* (Meara *et al*. 2003). Results showed a positive correlation between GJT performance and language aptitude. More specifically, the bilinguals with above-average aptitude were more likely to score within the native range on the GJT than those with below-average aptitude. It was also seen that among the participants with below-average aptitude, GJT scores were related to daily L1 use. In view of these findings, we suggest that language aptitude has a compensatory function in language attrition, helping the attriter to retain a high level of L1 proficiency despite reduced L1 contact.

In language acquisition, the time span running from birth to puberty stands out as a decisive period: studies on hearing impaired children (Mayberry and Lock 2003), feral children (Curtiss 1977), traumatic aphasia (Lenneberg 1967), and second language (L2) acquisition (e.g. Johnson and Newport 1989) have provided considerable evidence that languages are more readily learned prior to the onset of puberty. Research has also shown that even though exposure to a given language before puberty is a crucial condition, it might not be sufficient if nativelike proficiency should be attained; findings from e.g. Ruben (1999) and Hyltenstam *et al*. (2009) indicate that even minimal delays in language exposure from birth may compromise nativelikeness (for a discussion, see Hyltenstam and Abrahamsson 2003).[1]

Parallel to the age pattern documented in language acquisition, puberty seems to represent an important turning point in first language (L1) attrition as well: findings from this field of research show that if L1 contact is reduced prior to puberty, the L1 system may undergo severe loss (e.g. Ventureyra *et al.*

2004; Hyltenstam *et al.* 2009), whereas a break in the L1 speech community after the age of puberty seems to result in only minor effects on L1 maintenance (e.g. Köpke 1999; Yeni-Komshian *et al.* 2000).[2] Undoubtedly, the time frame spanning from birth up to puberty seems important not only for the acquisition of a given language, but also for the retention of it. The findings provided by attrition research thus indicate that nativelike attainment is not only dependent on early exposure, but also on continued, intense contact and use. This becomes particularly evident in the case of L1 attrition, where exposure to the language from birth meets the early exposure condition, but nativelike proficiency may be compromised due to reductions in L1 contact. Given the importance of this time span for language development, it seems important to identify the factors that eventually lead to the attainment of nativelike proficiency in situations of reduced L1 contact. While research to date on pre-pubescent attriters has focused primarily on the effects that degree of L1 contact (e.g. Hakuta and D'Andrea 1992) and L2 proficiency (Yeni-Komshian *et al.* 2000) exert on L1 maintenance, there is a dearth of knowledge about the role of *language aptitude* in attrition. Given that language aptitude has been demonstrated to be an important factor in L2 ultimate attainment, accounting for a large part of the variation among adult learners (DeKeyser 2000) as well as to some degree among early learners (Abrahamsson and Hyltenstam 2008), there is reason to believe that aptitude may constitute a factor in language attrition as well. Consequently, it is the aim of the current article to explore the role of language aptitude in pre-pubescent L1 attrition.

BACKGROUND

Pre-pubescent L1 attrition and outcome variability

A robust finding in the study of language attrition is the age-related differences in attrition outcome. While attrition in adults is generally low and manifests primarily in lexical retrieval difficulties (e.g. Olshtain and Barzilay 1991), children may attrite to such an extent that no remnants of the L1 appear to be left (Ventureyra *et al.* 2004). Not only do the children differ from adults as to the degree to which their language skills may possibly attrite, but also exhibit greater outcome variability. This variability becomes most salient by contrasting the findings from studies on immigrant children with studies on international adoptees: while in the former group attrition may be manifested in morphological reduction (e.g. Seliger 1991) and L2 convergence strategies (Montrul 2004a), the latter group is typically reported to experience an apparently complete loss of the L1 system (e.g. Isurin 2000; Pallier *et al.* 2003). The greater amount of variability found among early attriters can also be appreciated in the studies having examined L1 maintenance across a wider age range. Since these studies have used the same test instruments to collect data from speakers representing both pre- and post-pubescent attrition onsets, they provide a more solid basis for examining age-related outcome variability than does a comparison of different studies based on different research methodologies.

One of the studies that has investigated the potential effects of age in cases of attrition is Yeni-Komshian *et al.*'s (2000) examination of L1 pronunciation proficiency in 240 L1 Korean–L2 English bilinguals. A general finding in this study was that the lower the speaker's age of onset (AO) of L2 acquisition, the lower the score on the L1 pronunciation task. Specifically, the onset of puberty, operationalized as age 12, turned out to be a good predictor for native Korean pronunciation: almost all participants whose AO was greater than 12 converged with the Korean monolinguals, whereas the participants whose AO was less than 12 displayed a greater distribution in the pronunciation scores. Some speakers in this group fell within the Korean monolingual range, whereas others were characterized as having a heavy accent. In exploring the factors underlying this pattern, Yeni-Komshian *et al.* found that among the pre-pubescent attriters, pronunciation proficiency in the L2 correlated negatively with L1 pronunciation skills ($r = -0.47$, $p < 0.0001$) (the same did not hold for the post-pubescent group). It was also seen that amount of L1 contact (determined by self-reports) correlated positively with accuracy in L1 Korean pronunciation (separate analyses for each age group were not carried out in this case).

Bylund (2009a) found similar age-related effects while examining descriptions of goal-oriented motion events produced by 31 L1 Spanish–L2 Swedish bilinguals living in Sweden. The AO of the participants ranged from 1 to 19 years and their mean length of residence (LoR) was 32 years. The participants watched a set of videoclips projecting motion events and were asked to provide an oral online description for each clip (cf. von Stutterheim 2003). The results showed that deviations from the Spanish monolingual, preferred patterns of describing goal-oriented motion events were strongly associated with the onset of puberty (operationalized as age 12): all participants whose AO was greater than 12 years converged fully with Spanish monolingual behaviour, whereas those participants whose AO was less than 12 exhibited greater variability in their performance.

A similar pattern of variability tied to age differences was documented in Hakuta and D'Andrea's (1992) study on L1 maintenance and loss in 234 US high school students from Mexican backgrounds. L1 Spanish proficiency was measured by means of a productive vocabulary test, a grammaticality judgement test (GJT), and a cloze test. The results showed that the older the participants were at the onset of English acquisition, the better had they maintained their L1. More specifically, Hakuta and D'Andrea documented an increase in Spanish proficiency in the AOs ranging up to 10 years, after which the effects of AO on L1 proficiency levelled out. From the scatter plot (p. 82), it can also be appreciated that there is a great deal of variation in the test scores among the early AOs. Since the study did not include a control group, conclusions cannot be drawn about the participants' L1 proficiency in relation to that of Spanish-speaking monolinguals. Aside from the age function, Hakuta and D'Andrea found that the more frequently the Spanish was spoken at home, the higher the participants' scores were on the language tests.

As it becomes evident from the studies reviewed above, different kinds of factors may contribute to the varying proficiency outcomes in speakers experiencing a break with the L1 environment prior to puberty. First of all, it should be noted that in this group, failure to score within the native range may not always be ascribed to attrition. This is, for example, the case with some of the participants in Bylund (2009a): in this study, the deviations from the native norm attested in speakers having arrived in the L2 setting at the age of (say) one year were not manifestations of loss, since the speakers at the time of arrival could not possibly have acquired adult patterns of event conceptualization (cf. Sebastián and Slobin 1994). In cases where a person's separation from the L1 speech community occurs before the age by which a particular feature of the L1 would in normal circumstances have been fully acquired, that person's lack of convergence with L1 monolingual patterns can be attributed to incomplete acquisition rather than to attrition. Hence, one of the factors at play in the variable outcome in pre-pubescent speakers is incomplete acquisition. However, incomplete acquisition alone cannot account for the outcome variability in this age group. Consider that, for example, the findings of Bylund (2009a) and Yeni-Komshian *et al.* (2000) showed that age 12 was equally important for the retention of features that were acquired at different ages (event conceptualization patterns are fully mastered by the end of the first decade of life and global pronunciation is acquired by approximately age five). That is to say, even though a person's conformity with L1 monolingual patterns will be influenced by whether his or her break from the L1 environment took place before the age by which a given feature of the L1 is fully mastered, that person's conformity with L1 monolingual patterns may be even more decisively affected by whether his or her break from the L1 environment took place before the onset of puberty. The importance of puberty for L1 maintenance can also be appreciated in studies having examined speakers who experienced a break with the L1 environment at around puberty or later, such as the participants of Köpke (1999) and Schmid (2002), who were 11–29 and 14–36 years of age, respectively, at the time of arrival in the L2 setting. Results from these studies show that, first, the degree of attrition was relatively low, that is, no way near the drastic attrition levels found in pre-pubescent attriters; and second, L1 maintenance did not vary as a function of age (in Schmid's study, it was correlated with ethnic persecution).

Taken together, these findings suggest that there is a heightened susceptibility to attrition up until the onset of puberty. One can expect that, as a consequence of this susceptibility, L1 maintenance among pre-pubescent attriters is to a greater extent dependent on advantageous socio-psychological circumstances that can function as a counterweight to their proneness to attrition (cf. Bylund 2009b). This possibility may in part explain the attrition variability among pre-pubescent attriters: given the individual variation in socio-psychological circumstances, the heightened susceptibility to attrition is compensated to different degrees in different speakers, thus giving rise to the varying levels of attrition found across individuals. The individual variation that arises from socio-psychological factors is (as mentioned above) most clearly manifested in the differences between immigrant and adoptee attriters: a speaker in the former group may experience a break with the L1 tradition at a very early age, say two years, but due to favourable circumstances such as continuous L1 exposure, possibly along with positive attitudes towards L1, this speaker may eventually attain nativelike L1 proficiency. An adoptee attriter, on the other hand, whose L1 contact is effectively cut off (which is usually the case with international adoptees, cf. Hene 1993) at say age nine, will in spite of having attained a considerably high level of L1 proficiency most probably suffer from drastic attrition as a consequence of the disadvantageous circumstances for language maintenance. Although the pre-pubescent groups in the studies reviewed above were not treated separately with respect to this degree of L1 contact, the findings showed that L1 proficiency was positively correlated with L1 contact.

Another source of attrition variability is, as suggested by the findings of Yeni-Komshian *et al.* (2000), L2 proficiency. Results from their study showed an inverse relationship between L1 and L2 proficiency levels in bilinguals who had acquired the L2 before puberty. This finding is taken as evidence that pre-pubescent bilinguals are likely to end up with nativelike proficiency in *one* of their two languages, due to interaction and/or interference effects (see also Flege 1999). The idea that there is an inverse relationship between L1 attrition and L2 attainment is also found in the studies by Pallier *et al.* (2003) and Ventureyra *et al.* (2004) on international adoptees. Similar to Flege and colleagues, these researchers suggest that L2 attainment is distorted through L1 interference in pre-pubescent bilinguals. This suggestion may be traced to the idea that the brain has only a limited capacity for languages and that the 'addition of a second language automatically leads to a decrease of proficiency in the L1' (Hoffman 1991: 129; see also Cummins 1981).

Effects of language aptitude in language learning

In its most generic conception, language aptitude is defined as an innate, relatively fixed, talent to acquire and process language structure. The degree of language aptitude found within normal populations has been shown to vary significantly and, moreover, appears to be relatively unrelated to other individual factors such as general intelligence or personality (Novoa *et al.* 1988; Schneiderman and Desmarais 1988; Ross *et al.* 2002; Skehan 2002; Dörnyei and Skehan 2003). The four constituents of language aptitude identified as particularly important are (i) phonetic/phonemic coding ability, that is, the capacity to identify speech sounds and to make sound-symbol associations; (ii) grammatical sensitivity, that is, the capacity to identify the grammatical functions of words in a sentence; (iii) rote learning ability, that is, the capacity to in a rapid and efficient way associate lexical forms with meaning (i.e. to learn and remember new words effortlessly); and (iv) inductive learning ability, that is, the capacity to infer grammatical rules of a set of previously unknown language materials (for further discussion of these components, see Carroll and Sapon 1959; Carroll 1981).[3]

To date, research on language aptitude has focused primarily on foreign language learning. An exception to this trend is Skehan's (1989) and Skehan and Ducroquet's (1988) studies on the role of language learning aptitude for L1 development (in a monolingual L1 setting). Using data from an earlier study on L1 acquisition (obtained during the early 1970s), by Wells (1981, 1985), these studies examined the relationship between language aptitude and varying rates of L1 development (obtained during 1983–84). The results indicated significant correlations between degree of aptitude and certain aspects of L1 development, such as rates of auxiliary development and pronominalization (for further details, see Skehan 1989).

The studies on language learning aptitude from the field of L2 research (SLA) usually report a positive correlation (generally $r > 0.40$) between the degree of aptitude and the proficiency attained in the foreign language (e.g. Carroll and Sapon 1959; Skehan 1986). The finding that aptitude is an important factor in foreign language learning is further substantiated by studies that have assessed the effects of aptitude as well as motivation or attitudes towards the language to be learnt (e.g. Reves 1982). The results from these studies generally indicate that degree of aptitude is the most reliable predictor of language learning success (for an overview, see Dörnyei and Skehan 2003).

Besides investigations that have been conducted in the context of instructed language learning, a small number of studies have recently been carried out on the role of language aptitude in naturalistic L2 learning (e.g. DeKeyser 2000; Harley and Hart 2002; Abrahamsson and Hyltenstam 2008). A general finding from these studies is that language aptitude also plays a crucial role when learning takes place in a naturalistic setting. This is in line with Skehan's (1989) suggestion that aptitude will be a sizeable factor in informal settings—perhaps even more important than in formal ones—since informal settings place greater demands on the speakers' capacity to discover grammatical regularities and phonetic patterns merely from language exposure. The studies by DeKeyser (2000) and Abrahamsson and Hyltenstam (2008) provide information about the role of language aptitude in naturalistic settings and additionally take into account the AO of L2 learning: DeKeyser found that those adult learners who exhibited a high L2 proficiency level performed above average in an aptitude test (a Hungarian version of a subtest ('Words in Sentences'; Ottó 1996) of the *Modern Language Aptitude Test*, MLAT). Such a trend was not found among the child learners, whose L2 proficiency seemed to be unrelated to their aptitude scores. Framing these findings within the Fundamental Difference Hypothesis (Bley-Vroman 1989), DeKeyser suggested that if the implicit learning mechanisms are lost at around puberty, then language learning past this age must draw on explicit learning mechanisms, such as conscious reflection on linguistic structure. Consequently, near-native attainment in adult L2 learners would require a high degree of language aptitude.

Using a demanding GJT, Abrahamsson and Hyltenstam (2008) set out to test DeKeyser's claim that a high degree of language aptitude is crucial for native-like L2 proficiency in adult learners, but not for child learners. The results from this study showed that, first, in line with DeKeyser's finding, those adult learners who exhibited nativelike intuitions regarding grammaticality had a high aptitude, and second, that aptitude served as a significant predictor of native-like attainment even among the pre-pubescent learners (AO < 12) (aptitude was measured with the *Swansea Language Aptitude Test (LAT)v.2.0*; Meara *et al.* 2003). More specifically, many of the pre-pubescent learners who scored below the native range on the GJT also exhibited a below-average degree of aptitude, whereas the opposite held for those scoring within the native GJT range. There was a statistically significant correlation between aptitude and GJT scores among the pre-pubescent learners ($r = 0.70$, $p < 0.001$). On the basis of these findings, Abrahamsson and Hyltenstam concluded that a high degree of aptitude is a necessary, though not always sufficient, condition for nativelike proficiency among post-pubescent learners, whereas it is an advantageous, though not necessary, condition for pre-pubescent learners. Abrahamsson and Hyltenstam also showed that there was no significant correlation between the native-speaking controls' aptitude degrees and grammaticality judgement scores (DeKeyser's study included no native controls). This finding thus suggests that for native speakers aptitude is not important for L1 grammatical sensitivity.

Additional information about language aptitude may be found in research on exceptionality in language learning. There is reason to believe that polyglots, who belong to the select category of exceptionally multilingual individuals who have learnt several languages post-puberty, have a high level of aptitude (Hyltenstam forthcoming). Empirical data supporting this claim come from a detailed case study (Novoa *et al.* 1988). The learner ('C. J.'), a 29-year-old male in the USA, studied French, German, Spanish, and Latin during his high school years. In college he majored in French and spent a year in France at age 20. During that year he visited Germany only briefly and, merely listening to German 'was enough for him to recover his lost fluency' (p. 295). After graduation, he took up a governmental position in Morocco, and learnt Moroccan Arabic. After that he reactivated his Spanish and learnt Italian 'in a ''matter of weeks'''(p. 295). Native listeners to each of his languages confirmed that his abilities in each language were nativelike. The learner was given a large battery of neuropsychological tests, a language aptitude test and tests for visuo-spatial functions, musical ability, memory, and personality. The results showed that he was particularly adept at the acquisition of new codes, fluency, and vocabulary access. He excelled in formal aspects of language, but was average with respect to his performance on semantic and conceptual tasks. On the specific language aptitude test he was given (*MLAT*), he scored at or near ceiling on most components.

Aims of the present study

The purpose of the present study was to explore the role of language aptitude in pre-pubescent L1 attrition among Spanish–Swedish bilinguals. In particular, our aim was to examine the relationship between language aptitude and L1 grammatical intuition and processing. Due to the absence of previous studies on this matter, at this stage the effects of language aptitude on attrition can only be speculated on. Two competing hypotheses may be formulated concerning language aptitude and L1 attrition and L2 learning in a L2 setting. The first hypothesis stems from two assumptions: first, that aptitude promotes high levels of L2 proficiency (e.g. DeKeyser 2000), and second, that L2 proficiency is inversely related to L1 proficiency (cf. Yeni-Komshian *et al.* 2000; Pallier *et al.* 2003). From these assumptions follows the prediction that a high degree of aptitude *indirectly* would have negative consequences for L1 proficiency, since it would facilitate elevated levels of L2 proficiency at the expense of L1. That is, high aptitude would allow for greater flexibility in the language processing system, which would not only promote acquisition but also open up for attrition.

The second hypothesis, based on the assumption that a high level of L2 proficiency need not entail a decrease in L1 proficiency, suggests that aptitude makes possible the attainment of a high level of L2 proficiency while simultaneously facilitating L1 retention. In other words, language aptitude would reinforce L1 maintenance in situations conducive to attrition. The current study adheres to the second hypothesis. The reason for this position is twofold: first, based on the findings and interpretations of our prior research (Hyltenstam *et al.* 2009), we believe there is reason to question the hypothesis that predicts an inverse relationship between L1 and L2 proficiency.[4] Accordingly, even if advanced L2 proficiency is attained as a function of aptitude, this does not necessarily mean that L1 proficiency is negatively affected. Second, evidence from research on exceptional cases of language learning (e.g. Novoa *et al.* 1988; Hyltenstam forthcoming) suggests that persons with a propensity for picking up languages do not lose one language as soon as another one is learned; rather, these individuals seem to be able to keep a high level of proficiency in various languages simultaneously. Even though these findings mostly concern post-pubescent learners (as opposed to pre-pubescent bilinguals), we believe they may be generalized to other bilingual groups and taken as support for the second hypothesis. In the light of this reasoning, we predict that a generalized capacity to handle language structure should be reflected not only in the ability to acquire new languages, but also in an ability to retain a language in situations of reduced contact.

It should be noted that the scope of the present study is limited to explore the relationship between verbal analytical ability and L1 proficiency in situations of attrition, resorting to the current techniques available for operationalizing and measuring language aptitude. In other words, the study will not be concerned with investigating the nature of language aptitude *per se*.

METHODOLOGY

Participants

Twenty-five L1 Spanish–L2 Swedish bilinguals participated in the study. The majority (70 per cent) of the participants were of Chilean origin whereas the rest were born in other Latin American countries with no specific concentration. The AO of L2 acquisition of the participants ranged from 1 to 11 years (mean = 5.72) (this measure was used as an index of the break with the former monolingual L1 setting). The participants' LoR in Sweden ranged from 12 to 34 years (mean = 24.6) and their chronological age at the time of testing was between 20 and 41 years (mean = 30.2). All participants had completed upper secondary school and the majority of them also had academic degrees. The self-reported daily use of Spanish among the participants ranged between 5 and 50 per cent (mean = 25 per cent). A common denominator among the participants was that they were near-native speakers of Swedish: in a listener experiment, the participants had passed for native speakers of Swedish by the majority of a panel of native listener judges (for details, see Abrahamsson and Hyltenstam 2009).

Fifteen native speakers of Chilean Spanish were recruited as controls. This group was born and raised in a monolingual Spanish-speaking setting (Santiago de Chile), and none of these persons had lived abroad for any appreciable length of time. Pure monolingualism was, however, not a criterion for participation and some of the controls had elementary foreign language skills in, for example, English. The control group was matched with the bilingual participants with regard to educational level and chronological age. Because both the bilingual and control groups had Spanish as a L1, we have chosen not to use the term 'native speaker' to differentiate the controls from the bilinguals. Instead, we will refer to this group as 'monolingual controls'.

Procedure and materials

The participants were tested individually in a sound-treated room. The test sessions generally lasted 3.5 hours including two 20-minute breaks with sandwiches, fruit, and refreshments. The data collection involved, apart from measures of grammatical intuition and processing ability, different tests of pronunciation, speech perception, grammatical and semantic inferencing, and formulaic language. The test sessions with the bilingual participants took place at Stockholm University, and the monolingual controls were tested at the Pontificia Universidad Católica in Santiago de Chile. A native speaker of Chilean Spanish administered the tests in both Stockholm and Santiago de Chile. All participants were paid in return for their efforts.

Prior to testing, all participants underwent a hearing test (OSCILLA SM 910 screening audiometer) in which a loss of up to 10 dB for one frequency was considered acceptable.

Grammaticality judgement test

Grammatical intuition and processing ability were measured with an aural GJT. The test included 80 sentences of which 40 contained ungrammatical constructions regarding one of the following features: gender agreement, verb agreement, and verbal clitics.[5] According to previous research on the acquisition of L1 Spanish, these structures are acquired early on and mastered by the age of approximately five years (e.g. López Ornat 1994; Montrul 2004b). Research has also shown that these structures in Spanish may be vulnerable in situations of language contact and attrition (Lipski 2004; Montrul 2008). The sentences were carefully designed so that the participants' judgements would not be dependent on dialectal variation. Prior to application in the current study, the test sentences had been piloted extensively with both monolingual and bilingual speakers of Spanish, with the intention to remove any possible ambiguities or non-targeted deviations. Faulty sentences contained one error only (for samples, see Appendix 1). The participants were instructed to focus on the structure of the sentence and not on its content (the difference between these was illustrated through an example). The sentences were presented through earphones in different random orders and the participants indicated the grammaticality of each item by pressing a red button for 'incorrect' or a green button for 'correct'. The test was designed and run in *E-Prime* (v.1.0; Schneider *et al.* 2002a, 2002b). The sentences had been recorded in an anechoic chamber by a female native speaker of Chilean Spanish.[6]

Language aptitude test

Measures of the bilingual speakers' language aptitude were obtained through the *Swansea Language Aptitude Test* (LAT, v.2.0; Meara *et al.* 2003). This test was developed through a series of research projects at the University of Wales, Swansea, and is described by Meara and associates as being based on the MLAT (Meara 2005).[7] The Swansea LAT comprises five subtests: phonetic memory (LAT A), lexical–morphological analytical skills (LAT B), grammatical inferencing skills (LAT C), aural memory for unfamiliar sound sequences (LAT D), and the capacity to form sound–symbol correspondences (LAT E). The subtests include linguistic materials from either artificial language systems or languages with which the participants were unlikely to be familiar. The test was administered on computer, LAT A, D, and E also through earphones. The test took about 40–60 minutes to complete, and the maximum test score was expressed as 100 per cent.

RESULTS

L1 grammatical intuition and processing

The results from the grammaticality judgements showed that the bilinguals attained a significantly lower score than the controls. The average score for the bilinguals was 59.5 (SD 9.47) and for the controls 65.9 (SD 6.25). This difference was statistically significant (t (38) = −2.25, p = 0.03). In Figure 1, the test results from the GJT are laid out according to the AO of L2 acquisition of the bilinguals.

It was found that 15 out of the 25 (i.e. 60 per cent) participants performed within the controls' range (the lower limit of which is indicated by the horizontal line), whereas 10 (i.e. 40 per cent) fell outside of the range of nativelike intuitions about grammaticality. It can also be seen that the conformity with the controls' behaviour seems to be randomly distributed with regard to AO: on the one hand, there are participants with AO 1 scoring within the monolingual range, and on the other, there are those with AO 10 falling below the monolingual range. A correlational test (Pearson's) confirmed that there was virtually no correlation between AO and GJT scores (r = 0.04, p = 0.84). This pattern also seems to suggest that the age by which the tested features were

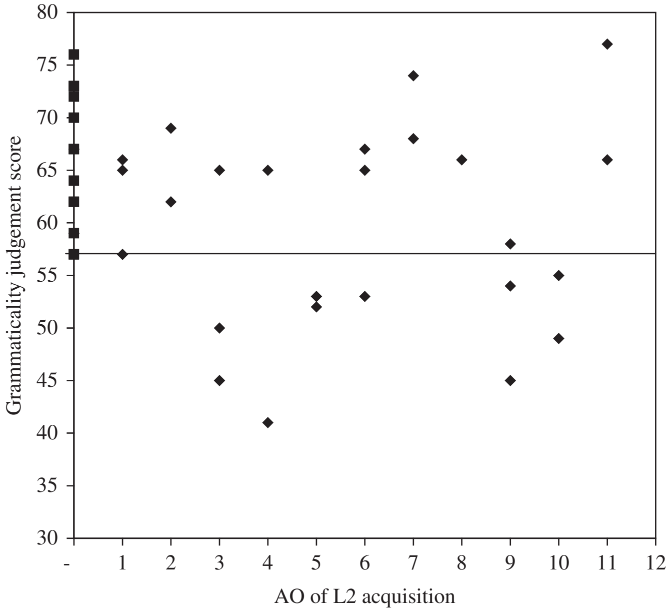

Figure 1: Grammaticality judgement scores according to AO. Diamonds = bilinguals; squares = monolingual controls (note that the controls are found on the y-axis); horizontal line = controls' lower range

acquired does not play in whether the bilinguals' intuitions about grammaticality will converge with monolingual behaviour. As seen in Figure 1, AO 5 (i.e. the age by which these features are thought to be mastered) does not mark a change in GJT performance. Instead, 7 out of the 15 participants who had supposedly acquired the GJT structures by the time of the break with the L1 environment fell below the monolingual range, whereas among those who could be assumed not to have completely acquired these structures only 3 out of 10 exhibited such behaviour.

Another possible source of the varying GJT scores could have been LoR in the L2 context (cf. Köpke and Schmid 2004), which among the participants ranged from 12 to 42 years. There was, however, virtually no correlation at all between LoR and GJT scores ($r = -0.02$, $p = 0.89$). We also ran a correlational test on the relationship between GJT performance and the bilinguals' daily use of Spanish. However, similar to LoR, daily L1 use did not turn out to be a reliable predictor of the GJT results when the group was analysed as a whole ($r = 0.15$, $p = 0.48$).

Language aptitude and L1 grammatical intuition and processing

The average score among the bilinguals on the language aptitude test was 58.2 (SD 9.47).[8] The highest scoring participant obtained 76.3 and the lowest scoring obtained 38.3 (maximum score possible = 100). There was a positive and statistically significant correlation between the bilinguals' GJT scores and their language aptitude ($r = 0.52$, $p < 0.01$). That is to say, the higher the score on the aptitude test, the higher the score on the GJT. The relationship between grammatical proficiency and aptitude is depicted in Figure 2. This figure is a replication of Figure 1, but information about the participants' degree of language aptitude has been added: Those participants who scored above average (i.e. >58.2) on the aptitude test are represented by filled diamonds, whereas those who scored below average are represented by empty diamonds.

A certain pattern can be observed between the degree of language aptitude and nativelike L1 proficiency. The pattern indicates that almost all of the participants who had above-average aptitude scores performed within the range of monolingual controls on the GJT (the horizontal line indicates the lower limit of this range). The exception to this trend is represented by a participant with AO 9 whose aptitude score, 59 points, was 0.8 above average (this was actually the lowest aptitude score obtained among the participants with above-average aptitude). This person had reported an average degree of daily use of Spanish (i.e. 25 per cent).

As for the participants with below-average aptitude scores, Figure 2 shows that 9 out of 13 (or 69.2 per cent) fell outside of the range of nativelike intuitions about grammaticality. Among those four participants whose aptitude was below average but still scored within the native range on the GJT, one had an AO of two years. Considering this low AO as well as the degree of aptitude, it may seem unexpected that this participant still attained a nativelike

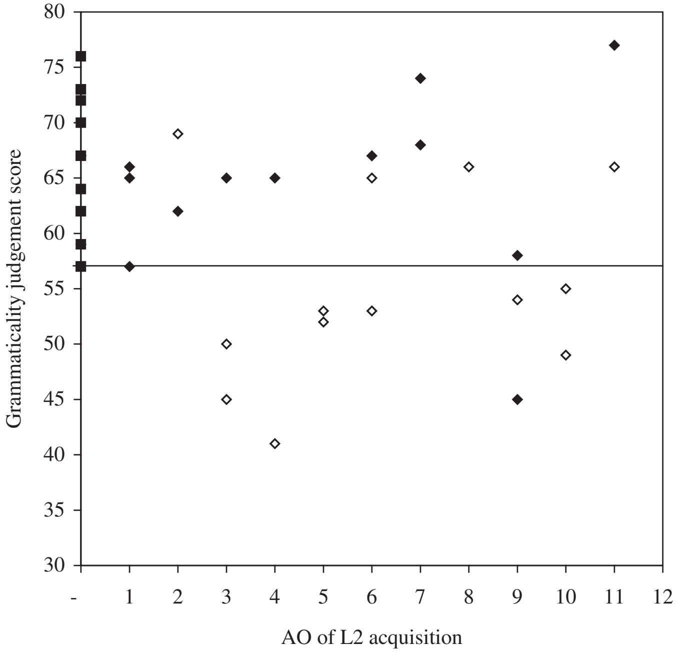

Figure 2: Grammaticality judgement scores according to AO and language aptitude. Black diamonds = bilinguals with above-average aptitude; white diamonds = bilinguals with below-average aptitude; squares = monolingual controls (note that the controls are found on the y-axis); horizontal line = controls' lower range

GJT score. However, this participant turned out to be the only one in the group who, after her/his arrival in Sweden, had returned to the country of origin for an extended period of time: After finishing high school, this person returned to the former L1 setting to study at the university and to work for 12 years. This person also reported a 30 per cent daily use of Spanish. None of the other three participants (AO 6, 8, and 11) with low-degree aptitude who scored within native range on the GJT had spent any appreciable length of time in their country of origin (or in any other Spanish-speaking environment) after moving to Sweden. It turned out that, nevertheless, these three individuals had a self-reported daily L1 use at 40 per cent. In order to check whether L1 use could have influenced the GJT performance of those participants with below-average aptitude, a correlation was run between the GJT scores and self-reported daily L1 use. The results showed that there was indeed a significant correlation between these measures ($r = 0.60$, $p = 0.03$), indicating that in this group daily L1 use had a positive effect on GJT performance. The same did not hold for the participants with above-average aptitude: in this case, no significant correlation was found between daily L1 use and the GJT scores ($r = -0.22$, $p = 0.49$).

DISCUSSION

Language aptitude and L1 attrition

The adopted hypothesis predicted that language aptitude would have a positive effect on L1 proficiency in situations of reduced L1 contact. The results corroborated this prediction. Language aptitude turned out to be a reliable predictor of GJT performance: a significant positive correlation was found between the participants' degree of language aptitude and their performance on the GJT. AO of L2 acquisition, on the other hand, did not correlate with the GJT scores. The absence of AO effects in the present study is fairly consistent with previous research showing that major effects of AO are typically not found among groups of pre-pubescent attriters (e.g. Yeni-Komshian *et al.* 2000; Bylund 2009a). The same held for LoR: no correlation was found between this variable and the participants' GJT scores. This finding too is in line with the prediction that no major LoR effects on attrition are to be expected when LoR exceeds 10 years (de Bot *et al.* 1991; de Bot and Clyne 1994).

How, then, should the correlation between degree of language aptitude and grammaticality intuitions be interpreted? Why does a speaker with a high degree of language aptitude attrite to a lesser extent than does a speaker with a low degree of aptitude? We suggest that the reduction in L1 contact—which is the primary catalyst for attrition—needs to be taken into account in order to characterize the function of aptitude in attrition. The importance of L1 contact for L1 maintenance has been emphasized in attrition research throughout the years (e.g. Andersen 1982; Sharwood Smith and van Buren 1991; Köpke and Schmid 2004; Paradis 2007).[9] Andersen (1982), for example, in his seminal paper stated that without a reduction in L1 contact 'it is unlikely that there will be much attrition at all' (p. 90). In order to maintain full L1 proficiency, the speaker is in need of evidence confirming that L1 is the way it is (Sharwood Smith and van Buren 1991). In the absence of L1 contact, the speaker's L1 proficiency will be affected in such a way that he or she no longer is able to make the same kind of linguistic distinctions made by proficient native speakers of that language (Andersen 1982: 91). We propose that the role of language aptitude in attrition relates to the need for L1 contact, in the sense that a speaker with a high degree of aptitude is to a lesser extent dependent on continuous L1 contact in order to maintain L1 proficiency. This interpretation is applicable to both attrition and incomplete acquisition. In the case of L1 attrition, a speaker with a high degree of aptitude will cope with the decreased L1 input without any drastic manifestations of loss, whereas a speaker with a low degree of aptitude will be more affected by the changes in the linguistic setting and thus attrite to a greater extent. As for incomplete acquisition, a high degree of aptitude will help the speaker to, first, maintain the proficiency level acquired by the time of the break with the L1 environment, and second, continue developing the L1 on the basis of the sparse input available. An incomplete learner with a low degree of aptitude, on the other hand, will probably have a harder time coping with reductions in L1 contact and is consequently unlikely to retain the acquired knowledge and continue acquiring the language to the same extent as would a speaker with a high degree of language aptitude.

The current findings indeed lend support to the interpretation that the need for L1 contact is connected to the speaker's degree of language aptitude. Among the participants with below-average aptitude, a positive significant correlation was found between self-reported daily L1 use and GJT scores. Among the participants with above-average aptitude, on the other hand, these variables turned out not to be significantly correlated. This result thus indicates that a speaker with a low degree of aptitude is to a greater extent dependent on L1 contact in order to retain/attain nativelike grammatical intuitions, whereas a speaker with a high degree of language aptitude is less dependent on L1 contact to retain or attain a high level of proficiency.

One could ask, however, to what extent language aptitude may compensate for reduced L1 contact? A finding from the present study might offer some information in this regard: it was found that a participant whose degree of aptitude was slightly above average scored below the range of the monolingual controls. In spite of the fact that this behaviour was only documented in one case, it suggests that although an above-average level of aptitude is certainly important for nativelike proficiency among pre-pubescent bilinguals, it may not always be a sufficient condition.

Another remark to be made regarding the relation between aptitude and L1 contact concerns the fact that the participants in the current study had stayed in contact with the L1 (albeit to varying degrees). A question that is relevant to our discussion is therefore how language aptitude would affect L1 proficiency if L1 contact were reduced to zero. A case in point is the linguistic situation of international adoptees, which is most frequently characterized by a complete absence of L1 contact (Hene 1993). Is it possible that a high level of aptitude, even under such extreme circumstances, would facilitate L1 retention?[10] Although a high degree of aptitude could be beneficial for L1 retention even among speakers who experience a complete cut-off in L1 contact, some degree of L1 contact may be necessary for language aptitude to really come into play, allowing nativelike proficiency to be maintained/attained. This question is obviously open to further research.

Besides L1 contact, other individual factors such as attitudes and motivation towards L1 maintenance (cf. Köpke 1999; Yağmur *et al*. 1999) could also come into play, compensating for a low degree of aptitude. There is, nevertheless, one fundamental way in which language aptitude differs from other individual factors: aptitude is relatively fixed and does not vary as a function of external circumstances (e.g. Politzer and Weiss 1969; Skehan 1998). In contrast, other individual factors such as L1 contact, attitude, or motivation may vary as external circumstances change (cf. Schmid 2002). A given variable's stability over time may be a particularly important feature in explaining pre-pubescent attrition: due to the fact that pre-pubescent attriters experience a reduction in L1 contact during a period when they are allegedly more susceptible to changes in the linguistic setting (Köpke and Schmid 2004; Bylund 2009b), research on this group would benefit from data about the attriter's individual circumstances *during* this period in order to determine which variables may or may not have contributed to the attrition outcome. Given the more changeable nature of attitude and motivation (as well as L1 contact), the indices that are obtained for these variables at the time of testing cannot with certainty be said to be representative for the whole time span in the L2 environment. One way of solving this problem would be to ask the participants to describe in detail the linguistic situation during their childhood and in this way obtain diachronic indices (a difficult enterprise which may still not render a reliable result; see Schmid 2004 for further discussion). The same methodological problems are not present in the case of language aptitude: due to the stability of this variable, the level of aptitude measured in a participant at the time of testing, will be the same as it was (say) 15 years ago, or more importantly, by the time of the break with the L1 setting. Hence, due to its stable nature language aptitude seems to constitute a reliable variable in predicting the outcome of attrition.

Language aptitude and bilingual proficiency

The last aspect to be treated in the discussion concerns the first of the two hypotheses about aptitude in attrition that were formulated in the beginning of the article. Assuming an inverse relationship between L1 and L2 proficiency levels, the first hypothesis predicted that if language aptitude leads to increased L2 proficiency this should have a negative effect on L1 proficiency. Since the findings of the current study showed the opposite, this hypothesis was not supported. Actually, the result that aptitude was positively correlated with L1 proficiency not only disconfirms the hypothesis, but it also raises doubts about the premises on which the hypothesis was based. The documented positive effects of language aptitude on L2 ultimate attainment (DeKeyser 2000; Abrahamsson and Hyltenstam 2008), together with the findings of the present study, seem to suggest that a person with a high degree of aptitude not only attains a high proficiency level in the L2, but he/she is also able to retain/develop a high level of L1 proficiency despite limited L1 contact. This state of affairs seems to be at odds with the suggestion about inversely related proficiency levels.

The reason that such different views on L1–L2 proficiency interaction arise may in part be ascribed to the scarcity of SLA or attrition investigations that empirically assess the proficiency level in more than one language (Yeni-Komshian *et al.* 2000 being, to the best of our knowledge, one of the few exceptions) (for further discussion, see Schmid and Köpke 2007). As a consequence, certain constructs on interacting proficiency levels have tended to build upon theoretical reasoning about constraints on memory and processing abilities (see, e.g. Bever 1981; Hoffman 1991; Pallier *et al.* 2003; Francis 2005). According to such conceptions, the languages of the bilingual may be seen as interacting containers where a proficiency increase in one language may consume capacity from the other and vice versa (cf. Hyltenstam *et al.* 2009). A criticism that could be presented against such constructs of bilingual proficiency is that they do not take into account the notion of exceptionality. That is to say, the idea that a major capacity to handle and process language structure may lead to high levels of proficiency in both L1 and L2 does not seem possible in such a framework. In the light of the findings from the present study, we suggest that models aiming at explaining variation in L1 and L2 proficiency as a function of an inverse relationship between the two could benefit from taking into consideration the role that language aptitude may play in bilingual proficiency.

CONCLUSIONS

The aim of the present study was to explore the effects of language aptitude in pre-pubescent L1 attrition. The findings demonstrated that speakers with an above-average degree of aptitude were more likely to exhibit nativelike grammatical intuitions than were speakers with a below-average degree of aptitude. It was also found that nativelike grammatical intuitions among speakers with below-average aptitude were connected to amount of daily L1 use. This effect was not found among the speakers with a high degree of aptitude. Our interpretation of these results is that language aptitude has a compensatory function in situations of reduced L1 contact, in that the speaker's degree of aptitude to a certain extent regulates his/her dependency on L1 contact to achieve and maintain L1 proficiency.

The fact that the present study found effects of language aptitude on L1 retention/attainment suggests that it may be valuable to continue exploring the role of aptitude in attrition. An important task for future research will be to confirm the level of generalizability of the current findings. In order to do so, future studies would benefit from, first, examining larger participant groups, and second, correlating aptitude with other types of language proficiency measures (e.g. production data). Such a development would stand the possibility to provide further knowledge not only about the role of language aptitude in attrition, but also about the nature of language aptitude *per se*.

APPENDIX 1

Six examples out of 80 grammaticality judgement sentences, grouped by structure type. (a) = grammatical sentences, (b) =*ungrammatical sentences. Target structures are underlined, and the correct structure for the ungrammatical sentences is given in [ ]. Translations have been done in a word-to-word fashion.

VERB AGREEMENT

(a) El profesor observó entusiasmado que las clases de historia eran muy populares entre todos los alumnos (‘The professor observed enthusiastic that the history classes were very popular among all the students’)

(b) *El reportaje sobre la Universidad de California llamaron [llamó] la atención en todos los estados federales. (*‘The reporting about the University of California attracted attention in all the federal states’)

GENDER AGREEMENT (ADJECTIVES IN PREDICATIVE POSITION)

(a) La actriz que desempeñó el papel de viuda en la última película del gran director italiano fue nominada a varios premios prestigiosos. (‘The actress that played the role of widow in the last film of the great Italian director was nominated for various prestigious prizes’)

(b) *Las cartas del rey, que en su versión original contenían mucha crítica contra el trabajo del obispo, fueron censurados [censuradas] por la Iglesia. (*‘The king’s letters, that in their original version presented much critique towards the bishop, were censored by the Church’)

VERBAL CLITICS

(a) Según los expertos, la producción de vino blanco en California destaca por su calidad que es reconocida dentro y fuera de los Estados Unidos. (‘According to the experts, the production of white wine in California stakes out because of its quality that is acknowledged inside and outside the United States’)

(b) *Las manzanas cultivadas en Argentina distinguen [se distinguen] de las europeas por su excelente calidad. (*‘The apples grown in Argentina distinguish from the Europeans because of their excellent quality’)

ACKNOWLEDGEMENTS

This work was supported by the Bank of Sweden Tercentenary Foundation (grant no. 1999-0383:01) and the Swedish Research Council (grant no 421-2004-1975). We wish to thank Johan Roos and Alejandra Donoso for running the experiments; Katrin Stölten and the staff at the phonetics laboratory at Stockholm University for their assistance in recording the GJT sentences; Alejandra Donoso and Carlos Henderson for constructive feedback on a previous version of the GJT; three anonymous reviewers for insightful comments on an earlier draft of this article. None of these persons, however, is responsible for any remaining errors of fact or interpretation.

NOTES

[1] It is important to note that even though the ultimate cause of age-related differences has not yet been demonstrated convincingly (for a discussion, see DeKeyser and Larsen-Hall 2005), there is general agreement about the fact that such differences exist.

[2] Attrition is defined as 'a non-pathological decrease in language proficiency that had previously been acquired by an individual' (Köpke and Schmid 2004: 4).

[3] Although these components were identified by Carroll and Sapon (1959) around 50 years ago, it should be noted that subsequent research has not 'reconceptualized aptitude in any significant manner' but has mostly worked 'within the aptitude agenda set by Carroll' (Skehan 2002: 73).

[4] In Hyltenstam *et al.* (2009), it was shown that L2 speakers, who due to adoption had undergone a complete L1 loss, exhibited non-native L2 features to the same degree as immigrant speakers with the same AOs. This finding would contradict the suggestion that the attainment of L2 nativelikeness is only possible if L1 is completely lost (e.g. Pallier *et al.* 2003).

[5] The test in fact contained 100 sentences of which 20 were designed to test tense/aspect contrasts. These sentences were removed from analysis with the intention of leaving aside items that were anomalous due to contextual rather than morphosyntactic criteria. See Bylund and Jarvis (forthcoming) for a report on the participants' ability to discriminate aspectual contrasts.

[6] The reading of the sentences aimed at following the orthography; for example, features typical for Chilean Spanish such as aspiration of /s/ were not present in the material.

[7] According to DeKeyser (2000) and Skehan (2002), the MLAT is still the most widely recognized and commonly used measurement battery of aptitude.

[8] An anonymous reviewer asks if the aptitude scores obtained by the participants correspond to a generally high, low, or average degree of aptitude, and further suggests that given the participants' near-native L2 proficiency, one may expect that they had an elevated degree of aptitude. One way of addressing this question would have been to compare the participants' aptitude degrees with those of monolingual speakers. Unfortunately, this was not possible in the present study since we do not have measures on the controls' aptitude. However, previous findings from Abrahamsson and Hyltenstam (2008) show that the aptitude degrees among near-native prepubescent L2 speakers is normally distributed and not significantly different from monolingual speakers'.

[9] However, see Schmid (2007) for a critical assessment of the role of L1 contact in L1 attrition.

[10] Parenthetically, it could be pointed out that the answer to this question may be related to some methodological difficulties: as attrition in adoptees is commonly manifested in an apparent complete loss (i.e. an inability to recall or even recognize L1 features), the search for L1 remnants must necessarily involve some sort of reactivation/relearning activities (Bylund 2009b; Hyltenstam *et al.* 2009). This, in turn,

may cause problems as to how to interpret the relationship between aptitude and the proficiency level in the forgotten L1 post-relearning. If a correlation is found between these two measures, it could be either due to the fact that aptitude had facilitated retention, or simply a result of aptitude rendering (re)learning more successful (i.e. the fact that the participant once was a proficient speaker of the target language is not a factor), or a combination of both.

REFERENCES

Abrahamsson, N. and **K. Hyltenstam.** 2008. 'The robustness of aptitude effects in near-native second language acquisition,' *Studies in Second Language Acquisition* 30: 481–509.

Abrahamsson, N. and **K. Hyltenstam.** 2009. 'Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny,' *Language Learning* 59: 249–306.

Andersen, R. W. 1982. 'Determining the linguistic attributes of language attrition' in Lambert R. and B. Freed (eds): *The Loss of Language Skills*. Newbury House, pp. 83–117.

Bever, T. 1981. 'Normal acquisition processes explain the critical period for language learning' in Diller K. (ed.): *Individual Differences and Universals in Language Learning Aptitude*. Newbury House, pp. 176–8.

Bley-Vroman, R. 1989. 'What is the logical problem of foreign language learning?' in Gass S. and J. Schachter (eds): *Linguistic Perspectives on Second Language Acquisition*. Cambridge University Press, pp. 41–68.

de Bot, K. and **M. Clyne.** 1994. 'A 16-year longitudinal study of language attrition in Dutch immigrants in Australia,' *Journal of Multilingual and Multicultural Development* 15/1: 17–18.

de Bot, K., P. Gommans, and **C. Rossing.** 1991. 'L1 loss in an L2 environment: Dutch immigrants in France' in Seliger H. and R. Vago (eds): *First Language Attrition*. Cambridge University Press, pp. 87–98.

Bylund, E. 2009a. 'Effects of age of L2 acquisition on L1 event conceptualization patterns,' *Bilingualism: Language and Cognition* 12: 305–22.

Bylund, E. 2009b. 'Maturational constraints and first language attrition,' *Language Learning* 59: 687–715.

Bylund, E. and **S. Jarvis.** (forthcoming). 'L2 effects on L1 event conceptualization,' *Bilingualism: Language and Cognition*.

Carroll, J. B. 1981. 'Twenty-five years of research in foreign language aptitude' in Diller K. (ed.): *Individual Differences and Universals in Language Learning Aptitude*. Newbury House, pp. 83–118.

Carroll, J. B. and **S. Sapon.** 1959. *Modern Language Aptitude Test. Form A*. Psychological Corporation.

Cummins, J. 1981. *Bilingualism and Minority Language Children*. Ontario Institute for Studies in Education.

Curtiss, S. 1977. *Genie: A Psycholinguistic Study of a Modern-Day ''Wild Child''*. Academic Press.

DeKeyser, R. 2000. 'The robustness of critical period effects in second language acquisition,' *Studies in Second Language Acquisition* 22: 499–533.

DeKeyser, R. and **J. Larsen-Hall.** 2005. 'What does the critical period really mean?' in Kroll J. and A. Groot (eds): *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, pp. 89–108.

Dörnyei, Z. and **P. Skehan.** 2003. 'Individual differences in second language learning' in Doughty C. and M. Long (eds): *Handbook of Second Language Acquisition*. Blackwell, pp. 612–30.

Flege, J. E. 1999. 'Age of learning and second language speech' in Birdsong D. (ed.): *Second Language Acquisition and the Critical Period Hypothesis*. Lawrence Erlbaum, pp. 101–32.

Francis, N. 2005. 'Research findings in attrition: Implications for the critical period hypothesis,' *Language Learning* 55/3: 491–531.

Hakuta, K. and **D. D'Andrea.** 1992. 'Some properties of bilingual maintenance and loss in Mexican background high-school students,' *Applied Linguistics* 13: 72–99.

Harley, B. and **D. Hart.** 2002. 'Age, aptitude, and second language learning on a bilingual exchange' in Robinson P. (ed.): *Individual Differences and Instructed Language Learning*. John Benjamins, pp. 302–30.

Hene, B. 1993. *Utlandsadopterade Barns och Svenska Barns Ordförståelse. En Jämförelse mellan Barn i Åldern 10-12 år.* [Word Knowledge in Foreign Adoptees. A Comparison between Children in the Ages 10-12 Years]. SPRINS-gruppen 41. Gothenburg University.

Hoffmann, C. 1991. *An Introduction to Bilingualism*. Longman.

Hyltenstam, K. (forthcoming). 'The polyglot – on exceptional ability to achieve high-level proficiency in numerous languages' in Hyltenstam K. (ed.): *High-Level Proficiency in Second Language Use*. Mouton de Gruyter.

Hyltenstam, K. and **N. Abrahamsson.** 2003. 'Maturational constraints in SLA' in Doughty C. J. and M. H. Long (eds): *The Handbook of Second Language Acquisition*. Blackwell, pp. 539–88.

Hyltenstam, K., E. Bylund, N. Abrahamsson, and **H.-S. Park.** 2009. 'Dominant language replacement: The case of international adoptees,' *Bilingualism: Language and Cognition* 12: 121–40.

Isurin, L. 2000. 'Deserted islands or a child's first language forgetting,' *Bilingualism: Language and Cognition* 3: 151–66.

Johnson, J. and **E. Newport.** 1989. 'Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language,' *Cognitive Psychology* 21: 60–99.

Köpke, B. 1999. 'L'attrition de la première langue chez le bilingue tardif: implications pour l'étude psycholinguistique du bilinguisme', [First Language Attrition in Late Bilinguals. Implications for the Psycholinguistic Study of Bilingualism], PhD dissertation, Université de Toulouse-Le Mirail.

Köpke, B. and **M. Schmid.** 2004. 'Language attrition: The next phase' in Schmid M., B. Köpke, M. Keijzer, and L. Weilemar (eds): *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*. John Benjamins, pp. 1–43.

Lenneberg, E. 1967. *Biological Foundations of Language*. John Wiley & Sons, Ltd.

Lipski, J. 2004. 'La lengua española en los Estados Unidos: avanza a la vez que retrocede,' [The Spanish Language in the United States: Simultaneous progression and regression], *Revista Española de Lingüística* 33: 231–60.

López Ornat S. (ed.). 1994. *La Adquisición de la Lengua Española*. [Spanish Language Acquisition]. Siglo XXI.

Mayberry, R. and **E. Lock.** 2003. 'Age constraints on first versus second language acquisition: Evidence for linguistic plasticity and epigenesis,' *Brain and Language* 87: 369–84.

Meara, P. 2005. 'Llama Language Aptitude Tests. The Manual,' University of Wales, Swansea.

Meara, P., J. Milton, and **N. Lorenzo-Dus.** 2003. *Swansea Language Aptitude Tests (LAT) v2.0*. Lognostics.

Montrul, S. 2004a. 'Subject and object expression in Spanish heritage speakers,' *Bilingualism: Language and Cognition* 7: 125–42.

Montrul, S. 2004b. *The Acquisition of Spanish: Morphosyntactic Development in Monolingual and Bilingual L1 Acquisition and Adult L2 Acquisition*. John Benjamins.

Montrul, S. 2008. *Incomplete Acquisition in Bilingualism*. John Benjamins.

Novoa, L., D. Fein, and **L. K. Obler.** 1988. 'Talent in foreign languages: a case study' in Obler L. K. and D. Fein (eds): *The Exceptional Brain: Neuropsychology of Talent and Special Abilities*. Guilford Press, pp. 294–302.

Olshtain, E. and **M. Barzilay.** 1991. 'Lexical retrieval difficulties' in language attrition' in adult, H. Seliger, and R Vago (eds): *First Language Attrition*. Cambridge University Press, pp. 139–50.

Ottó, I. 1996. *'Hungarian language aptitude test: Words in sentences'*. Unpublished manuscript, Department of English Applied Linguistics, Eötvös Loránd University, Budapest.

Pallier, C., S. Dehaene, J.-B. Poline, D. LeBihan, A.-M. Argenti, E. Dupoux, and **J. Mehler.** 2003. 'Brain imaging of language plasticity in adopted adults: Can a second language replace the first?,' *Cerebral Cortex* 13: 155–61.

Paradis, M. 2007. 'L1 features predicted by a neurolinguistic theory of bilingualism' in Köpke B., M. Schmid, M. Keijzer, and S. Dostert. (eds): *Theoretical Perspectives*. John Benjamins, pp. 121–34.

Politzer, R. and **L. Weiss.** 1969. *An Experiment in Improving Achievement in Foreign Language Learning through Learning of Selected Skills Associated with Language Aptitude*. Stanford University.

Reves, T. 1982. 'What makes a good learner?,' PhD dissertation, Hebrew University.

Ross, S., N. Yoshinaga, and **M. Sasaki.** 2002. 'Aptitude-exposure interaction effects on Wh-movement violation detection by pre and post-critical period Japanese bilinguals' in Robinson P. (ed.): *Individual Differences and Instructed Language Learning*. John Benjamins, pp. 267–99.

Ruben, R. 1999. 'Persistency of an effect. Otitis media during the first year of life with nine years follow-up,' *International Journal of Pediatric Otorhinolaryngology* 49: 115–18.

Schmid, M. 2002. *Language Attrition, Maintenance and Use: The Case of German Jews in Anglophone Countries*. John Benjamins.

Schmid, M. 2004. 'A new blueprint for language attrition research' in Schmid M., B. Köpke, M. Keijzer, and L. Weilemar (eds): *First Language Attrition: Interdisciplinary Perspectives on Methodological Issues*. John Benjamins, pp. 349–63.

Schmid, M. 2007. 'The role of L1 use for L1 attrition' in Köpke B. *et al.* (eds): *Language Attrition: Theoretical Perspectives*. John Benjamins, pp. 135–54.

Schmid, M. and **B. Köpke.** 2007. 'Bilingualism and attrition' in Köpke B., M. Schmid, M. Keijzer, and S. Dostert (eds): *Language Attrition: Theoretical Perspectives*. John Benjamins, pp. 1–8.

Schneider, W., A. Eschman, and **A. Zuccolotto.** 2002a. *E-Prime User's Guide*. Psychology Software Tools.

Schneider, W., A. Eschman, and **A. Zuccolotto.** 2002b. *E-Prime Reference Guide*. Psychology Software Tools.

Schneiderman, E. and **C. Desmarais.** 1988. 'The talented language learner: Some preliminary findings,' *Second Language Research* 4: 91–109.

Sebastián, E. and **D. Slobin.** 1994. 'The development of linguistic forms: Spanish' in Berman R. and D. Slobin (eds): *Relating Events in Narrative*. Erlbaum, pp. 239–84.

Seliger, H. 1991. 'Language attrition, reduced redundancy, and creativity' in Seliger H. and R. Vago (eds): *First Language Attrition*. Cambridge University Press, pp. 227–40.

Sharwood Smith, M. P. and **van Buren.** 1991. 'First language attrition and the parameter setting model' in Seliger H. and R. Vago (eds): *First Language Attrition*. Cambridge University Press, pp. 17–30.

Skehan, P. 1986. 'The role of foreign language aptitude in a model of school learning,' *Language Testing* 3: 188–221.

Skehan, P. 1989. *Individual Differences in Second Language Learning*. Arnold.

Skehan, P. 1998. *A Cognitive Approach to Learning Language*. Oxford University Press.

Skehan, P. 2002. 'Theorizing and updating aptitude' in Robinson P. (ed.): *Individual Differences and Instructed Language Learning*. John Benjamins, pp. 69–93.

Skehan, P. and **L. Ducroquet.** 1988. *A Comparison of First and Foreign Language Learning Ability*. Working Documents 8, ESOL Department. London University.

Ventureyra, V., C. Pallier, and **H. Yoo.** 2004. 'The loss of first language phonetic perception in adopted Koreans,' *Journal of Neurolinguistics* 17: 79–91.

von Stutterheim, C. 2003. 'Linguistic structure and information organisation: The case of very advanced learners' in Foster-Cohen S. and S. Pekarek Doehler (eds): *EuroSLA Yearbook*. John Benjamins, pp. 183–206.

Wells, G. 1981. *Learning Through Interaction*. Cambridge University Press.

Wells, G. 1985. *Language Development in the Preschool Years*. Cambridge University Press.

Yağmur, K., K. de Bot, and **H. Korzilius.** 1999. 'Language attrition, language shift and ethnolinguisitc vitality of Turkish in Australia,' *Journal of Multilingual and Multicultural Development* 20/1: 51–69.

Yeni-Komshian, G., J. Flege, and **S. Liu.** 2000. 'Pronunciation proficiency in the first and second languages of Korean-English bilinguals,' *Bilingualism: Language and Cognition* 3/2: 131–49.

Applied Linguistics: 31/3: 465–470 © Oxford University Press 2010
doi:10.1093/applin/amq014 Advance Access published on 25 April 2010

FORUM

The Soft Ideological Underbelly of the Notion of Intelligibility in Discussions about 'World Englishes'

KANAVILLIL RAJAGOPALAN

State University at Campinas (UNICAMP), Brazil
Email: rajagopalan@uol.com.br

The term 'intelligibility' is widely viewed as denoting an ideologically neutral concept and therefore useful in speculating about the future of the English language, especially in the context of its expansion at the current exponential rate and the danger or otherwise of its breaking up into mutually incomprehensible languages, the way Latin did in the Middle Ages. It has also been bandied about in the context of English language teaching, especially to speakers of other languages. In this piece, I question the status of intelligibility as an ideologically innocent concept and argue that the adjective *intelligible* is analogous to others such as *beautiful*, *ugly*, *easy*, *difficult*, *primitive*, *civilized*, and so forth, which are also sometimes used with respect to languages, and which we have long learned to regard with suspicion on the grounds that they invariably presuppose the standpoint of someone who furtively manages to remain invisible.

Intelligibility seems to have become a buzzword these days, especially among scholars who are getting increasingly worried about the rate at which English is spreading right across the world like wildfire. It is being touted today as the one key litmus test for the integrity of the English language in its new role as the language of international communication or what some people call, not without some impish humor (as well as unintended irony), 'elf' (acronym for English as a Lingua Franca), and also as a guarantor of its continued existence (cf. Jenkins 2000, 2007).

Just how important the notion of intelligibility has become in scholarly discussions can be gauged by the fact that the journal *World Englishes* devoted a sizeable portion of an entire issue in 2008 to a 'Symposium on the intelligibility and cross-cultural communication in World Englishes' (initially organized as part of the 12th annual gathering of the International Association for World Englishes in Nagoya, Japan in 2006). In his editorial introduction, Kachru (2008: 293) lavished praise on Larry Smith as an early proponent of the idea and said that '[w]hat motivated Smith's research on intelligibility was his concern about methodologies of teaching English in the USA and beyond'.

Kachru made an extremely important point when he suggested that the question of intelligibility is most relevant to language pedagogy. While intelligibility might interest a language historian who is engaged in speculating about the future of the English language and who wonders whether or not the language is destined to go the Latin way (Rajagopalan 2009), that is disintegrating into a number of distinct and mutually unintelligible or only partially intelligible languages, it interests the English language teacher, especially someone involved in teaching the language to those for whom it is a second or foreign language. Their concern was expressed in no uncertain terms by Perren (1956: 3) more than half a century ago when he wrote:

> So far little attempt has been made to deal with the phonetic origin of errors in spoken English in either training colleges or schools. There is a danger that an 'East African English'—characterized by its own pronunciation, intonation and sentence patterns—will become normal among educated Africans.

In the present-day context, the question of intelligibility is generally raised against the backdrop of a widespread consensus building up among English language teaching professionals across the world over 'the need for educators to re-align themselves in the face of the changing ownership of English' (Holliday 2005: ix) and the idea of 'polycentricity' and its significance in contemporary globalized contexts of language use (Blommaert 2006). Intelligibility, it is held, is what will guarantee that what a rancher in Texas says will be minimally understandable to a primary school teacher somewhere in a remote corner of, say, Chennai or vice versa.

This is by all means a welcome change from the days, not so long ago, when it used to be claimed that the only way to attain intelligibility across the board was to accept the native speaker as the model, 'as the ultimate state at which first and second language learners may arrive and as the ultimate goal in language pedagogy' (Van der Geest 1981: 317). Bansal's classic work on the intelligibility of Indian English (Bansal 1969) unquestioningly accepted this idea, since it was R.P. (British 'Received Pronunciation') that served as its loadstar (Nelson 2008). The concept of 'linguistic competence' of the imaginary, idealized native speaker on which such claims were based soon gave way to the broader, richer notion of 'communicative competence'. In effect what this meant was that Chomsky's *homo syntacticus* was replaced by a speaker who is tethered to his/her native culture and immersed in it neck-deep. He/she became the key figure, the centerpiece, around which to devise teaching methodologies.

Park and Wee (2009: 393) rightly pointed out that 'a language-ideological turn is important for understanding the status of English in the world'. To this one might add the rider that it is even more so when it comes to the teaching of English all over the world. Given that the English language has for some time been the hottest selling commodity on the world linguistic market, it is hardly surprising that its ownership and/or 'leasing rights' should be so tightly contested. Concepts like 'authenticity' and cultural or situational 'appropriacy' that were freely floated around in the wake of the burgeoning orthodoxy referred to above did contribute toward ensuring special trading privileges for those who could claim the status of consummate native speakers (i.e. legal owners) of English.

Inconvenient details, such as Chomsky's having presented his prized notion of 'native speaker-hearer' with the modifier 'ideal' and also the other worldly description 'in a completely homogeneous speech community, who knows its language perfectly' and so forth, were brushed aside so as to prepare the grounds for what I have referred to elsewhere as 'the apotheosis of the native speaker' (Rajagopalan 1997). Once enthusiasm had died down over the Chomskyan paradigm, the new kid on the block was 'communicative competence' and, hot on its heels, communicative language teaching, where it was not only the native who was at the epicenter but also his native circumstances as well. Once again, the fact that not every learner of English as a second or foreign language was interested in communicating to or making friends with a so-called native speaker seemed not to stand in the way of those who swore by the new orthodoxy.

But times have changed. With just a handful of exceptions, the scholarly community has rallied massively behind the idea that the English language is no longer in the hands of this or that nation or group of nations. Instead, it has become fashionable to speak in terms of Englishes, in the plural. The journal *World Englishes* is now almost 30 years old and seems to have overcome initial reactions of smirk and simper from the wider public at large. People no longer balk at such expressions as 'East African English' or 'Singaporean English' the way they used to.

Amidst all this new-found enthusiasm over English as an international language, one also notices some rearguard action from sectors within the English language teaching (ELT) enterprise worldwide that are unwilling to see their erstwhile privileges slip away between their very fingers. And these sectors with their own vested interests have found a powerful ally in the notion of intelligibility. Ironically enough, this notion plays into the hands of those very people whose 'gatekeeping practices' designed 'to hamper [the] acceptance [of ELF] as legitimate English' have been duly noticed and decried (Jenkins 2007: 238).

Although, on the face of it, it would seem that intelligibility is an ideologically neutral concept, it turns out upon closer inspection that it is not so. Jenkins herself toys with the notion of 'non-reciprocal intelligibility' which clearly brings out how politically suffused the whole concept of intelligibility is. But she fails, I think, to perceive the ideologically loaded nature of the very concept. For, no matter how one tries to define intelligibility from a neutral standpoint, the question that cries out for an answer is: 'intelligible for who?' This becomes evident as we consider the following remark by Kirkpatrick (2007: 200), as he speaks of a CD with recordings of samples of varieties of

English and their corresponding transcripts appended at the end of his book *World Englishes*:

> As a general rule, listeners might like to listen to these excerpts before consulting the transcripts. In this way the relative intelligibility/unintelligibility of these varieties can be better appreciated. Please note that the excerpts here are simply intended to give listeners the opportunity to listen to a selection of different varieties of English.

The question is who is to decide whether a given stretch of language production is intelligible or unintelligible? Could it not be the case that what someone dismisses as unintelligible may well sound perfectly intelligible to another? What criteria are we supposed to bring to bear on cases where one person's judgement is at odds with another's? It is here that we begin to sense that with a concept such as intelligibility nurtured in a context where the so-called native varieties no longer rule the roost, the figure of the native speaker creeps back in, only this time through the back door and that too most stealthily.

Now, history is full of more blatant cases of unilateral claims of authority to pontificate on intelligibility. In the early 1990s, as the Foreign and Commonwealth Office in the UK began releasing piles of stationery pertaining to the last days of the Raj stacked away in the name of official secrecy, a number of files recording daily bureaucratic transactions were made public. One file carried the following remark scribbled in pen on a type-written memo: 'What Mr. Chatterjee writes, Mr. Mukherjee seems to understand; and what Mr. Mukherjee writes, Mr. Banerjee seems to be perfectly happy with. But this is certainly not what Captain Simpson understands as English' (cited from memory).

True, there have been many attempts to either tone down or sidestep the centrality of the native speaker in assessing intelligibility. Smith and Rafiqzad's remark (1979: 375, reported in Nelson 2008: 301) that, in their study, 'the native speaker was always found to be among the least intelligible speakers' is best seen as a distracter or diversionary after-dinner joke or, at the very least, the proof of the poor command of language of those who answered the questionnaire or the skewed sampling used in administering it.

Intelligible is an evaluatory adjective like *beautiful*, *ugly*, *easy*, *difficult*, *primitive*, *civilized*, and so on. All of them automatically invoke the figure of an evaluator. Those of us who can go back in time and recall our earliest lessons in General Linguistics, especially Introduction to the Scientific Study of Language or Linguistics 101, will easily remember that our first exposure to Linguistics was a lengthy harangue aimed at extolling the character of linguistics as a scientific discipline, opposed to the work of 'old-fashioned' traditional grammarians. And part of what was meant by saying that the discipline was scientific was made clear by highlighting the fact that it eschewed evaluatory remarks about language and used instead non-evaluatory, purely 'descriptive' ones. Thus, contrary to popular perception, no language, howsoever exotic it might sound, was easy or difficult, beautiful or ugly, in and of itself. Tamil might sound somewhat exotic to a speaker of Hungarian, but not to a speaker of, say, Malayalam or Telugu. By the same token, no variety is intelligible or otherwise in and of itself. Rather, it all depends on who is making the remark and about what language or variety.

It seems truly amazing that many of us who have been won over to the notion of intelligibility as that which would ultimately guarantee the survival of English as a lingua franca or as a universal language of communication among nations have nevertheless failed to see that, in spite of our best intentions and all-too frequent public disclaimers to the contrary, we have not fully got rid of some of the old habits of thinking, along with their deeply ingrained ideological implications.

Before wrapping up our discussion, we must consider the one question that seems to demand an answer at this juncture: if it is not intelligibility, what other notion can take its place? Paucity of space prevents me from attempting an elaborate answer. I shall, however, point out two clues to coming up with one. First, it is perhaps time that we overcame the temptation to think that it is the availability of a common language that will guarantee mutual intelligibility. I contend that it is precisely the other way around: it is a willingness or need to understand one another that makes it possible for us to postulate the existence of a common language in the first place (Rajagopalan 2001). Secondly, it might be useful to think of intelligibility, not in essentialist terms by looking for a minimum common core, or a highest common factor, but rather something like a lowest common denominator. Or, to put it differently, by thinking of a World English (in the singular) where different regional varieties display some sort of family resemblance with one another and the speakers can, whenever need arises, communicate with one another by learning to cope with whatever initial difficulty they may encounter.

Finally, a remark—a food for thought—to round off our discussion: the very question of 'intelligibility' or lack thereof might well be the offspring of our initial decision to start with 'World Englishes' in the plural, rather than 'World English' in the singular.

REFERENCES

Bansal, R. K. 1969. *The Intelligibility of Indian English: Measurement of the Intelligibility of Connected Speech, and Sentence and Word Material, Presented to Listeners of Different Nationalities*. Central Institute of English.

Blommaert, J. 2006. *Discourse*. Cambridge University Press.

Holliday, A. 2005. *The Struggle to Teach English as an International Language*. Oxford University Press.

Jenkins, J. 2000. *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford University Press.

Jenkins, J. 2007. *English as a Lingua Franca: Attitude and Identity*. Oxford University Press.

Kachru, B. 2008. 'The first step: the Smith paradigm for intelligibility in world Englishes,' *World Englishes* 27/(3/4), 293–6.

Kirkpatrick, A. 2007. *World Englishes. Implications for International Communication and*

English Language Teaching. Cambridge University Press.

Nelson, C. L. 2008. 'Intelligibility since (1969),' *World Englishes* 27/(3/4), 297– 308.

Park, J. S. and **L. Wee.** 2009. 'The three circles redux: a market theoretic perspective on World Englishes,' *Applied Linguistics* 30/3: 389–406.

Perren, G. E. 1956. 'Some problems of oral English in East Africa,' *English Language Teaching* XI/1: 3–10.

Rajagopalan, K. 1997. 'Linguistics and the myth of nativity: comments on the controversy over ''new/non-native Englishes'',' *Journal of Pragmatics* 27/2: 225–31.

Rajagopalan, K. 2001. 'The politics of language and the concept of linguistic identity,' *CAUCE: Revista de Filologia y su Didáctica* 24: 17–28.

Rajagopalan, K. 2009. ' ''World English'' and the Latin analogy: where we get it wrong,' *English Today* 25/2: 49–54.

Smith, L. E. and **R. Rafiqzad.** 1979. 'English for cross-cultural communication: the question of intelligibility,' *TESOL Quarterly* 13: 371–80.

Van der Geest, T. 1981. 'How to become a native speaker: one simple way' in F. Coulmas (ed.): *A Festschrift for Native Speaker*. Mouton, pp. 317–53.

REVIEWS

Alexandra Georgakopoulou: SMALL STORIES, INTERACTION AND IDENTITIES. John Benjamins, 2007.

Labov and Waletzky's (1967) model of narrative, along with its subsequent reformulation in Labov (1972) and the less frequently cited Labov and Fanshel (1977), is arguably one of the best known constructs in discourse analysis, having influenced subsequent generations of scholars within linguistics and beyond, becoming in effect what Latour (1987) has termed a 'black box', widely accepted and used, sometimes uncritically, for a whole range of analytic purposes. When for example I started my doctoral research in the early 1980s, I wanted to take a discourse perspective on Second Language Acquisition (SLA) and chose to focus on narrative assuming it, post-Labov, to be a relatively well-understood phenomenon. How wrong I was and how much more there is, as I subsequently discovered, to narrative analysis. Roberts and Campbell (2005) identify a Labovian schematic pattern in the guidelines for eliciting narratives in job interviews, suggesting that its influence has gone well beyond linguistics as such. In this monograph, Georgakopoulou takes what one might call a post-Labovian stance on narrative analysis. She argues that Labov's analysis has become an orthodoxy or canon, which, while perhaps working well for the narratives elicited in interview contexts, prototypically personal, past experience stories of non-shared events, ignore other narrative-like phenomena, particularly those characteristic of conversation. Her analysis of small stories provides as she puts it an antidote to conventional narrative analysis based on interview data. Small stories are narrative fragments, snippets, shifts into narrative, allusively evoking a narrative telling, what Hymes (1996) has termed 'fleeting moments of narrative orientation to the world'.

These small stories, told in conversations between members of a female Greek friendship group and in a smaller corpus of their e-mails, are investigated using a narrative as talk-in-interaction approach. These small stories are overwhelmingly joint constructions, *shared stories*. They both draw on and contribute to the articulation of a shared communicative history among group members, typically discussing and arguing over the detail of actual and potential meetings with men. They involve not just *narratives of personal experience*, but also *breaking news* (in which tellers share unfolding events as they occur, giving the narrative a commentary-like quality), *projections* (narratives of what might predictably or hypothetically occur), which Georgakopoulou finds to be in fact more prevalent in her conversational data than personal narratives of past experience. These tend to co-occur, with shared stories typically embedded in and used to interpret breaking news. So along with the breaking news and projections, there is an accumulation of shared stories that can be drawn on, referred to, and argued over in ongoing talk, providing an invaluable resource in the co-construction of shared identity among group members. To investigate these over time, the study combines a talk-in-interaction approach, drawing on conversation analysis and interactional sociolinguistics, but with an ethnographic dimension, situating the talk and the stories told both in emergent talk and in the communication practices and shared concerns of the young female co-conversationalists. As such it lines up with sociolinguistic studies of teenage friendship groups such as Bucholtz (1999) and indeed studies of older female friendship groups such as Coates (1996), but with a distinctive focus on narrative.

Is the focus on these small stories, narrative fragments or snippets, a turn away from narrative structure? Georgakopoulou's analysis continues to emphasize the identification and description of narrative structure, but for the intellectual tradition she is coming from, structure is emergent and locally occasioned (as the contrast between the different opportunities and affordances of the conversational data and the e-mail data incidentally demonstrates). Narrative is viewed as 'a sequentially organized activity in which structure emerges on-line and is negotiated by the participants' (p. 57).

What is the interpersonal work that is achieved by the telling and re-telling of these small stories? As suggested above, Georgakopoulou analyses them for their contribution to the ongoing identity work in which the group is continually engaged. She makes a distinction on the one hand between an approach to narrative and identity which sees narrative as a privileged mode for self-revelation, to be scrutinized by the analyst who reads and reconstructs the identity traces sitting as it were behind the text and on the other, the theoretical position that she adopts, which is to ask how identities are talked up in discourse, 'made visible, worked up and attended to in the sequential organization of talk' (p. 89). The analysis aims to make a connection between the identity work performatively engaged in through storytelling and larger social roles, particularly here gender roles, that are being discursively talked up in conversation. To do this, she uses the notion of situational identity adapted from Zimmerman (1998) as a meso construct to connect the local role of a teller with the larger scale social roles that are indexed in the story told.

In contrast to the traditional account of narrative as self-revelation, particularly associated of course with autobiography, these data show how identity construction can be dialogic and relational, interactionally achieved. In order to link the micro activity engaged in through talk, to the larger scale social identities, the analysis in Chapter 5 invokes the literature on positioning, understood in the sense of taking up positions in discourse, positioning others and also of course being positioned. The young women in the study position each other in talk, for example as more or less knowledgeable, more or less savvy about men, but the primary focus of their other positioning is on the men around them, those they are interested in, those they are not, those who are available or not, desirable or not, in their social world. The linguistic devices drawn on to position them are *nicknames* and *assessments* evoking 'soft' or 'hard' masculinity, stereotypically male, hard men, or softer men, these providing *membership categorization devices* and evoked through *category-bound activities* as well as *stylizations*, for example of voice quality or dialect.

The final twist of the analysis is a well-motivated connection made between the other positioning through which the men on their horizons are constructed in talk and a performative construction of the gendered self: 'constructing gendered positions for men is an integral process for the participants' constitution of their own gendered selves: they learn about self through representation of the other, through looking into the boundaries between self and other', constructing identities not only as heterosexual women but also as members of the female friendship group.

This is a rich and intriguing study. The argument comes off in a way that is intricately accountable to the data, though at times as a reader one has a sense of holding one's breath below as the author negotiates the tightrope of a particularly tricky bit of analysis, wondering how she is going to pull it off. At the end of the book, Georgakopoulou is in a position to say some interesting and convincing things about identity construction, the connection between the micro interactional and the social, but she has reached the insights by the novel route of interactional analysis rather than simply reading off identities from a monologic text. A short review also cannot give an impression of the richness and vividness of the data, presented both in the original Greek and in an English translation. Sustained throughout the book as well are theoretical arguments about the study of narrative and identity that challenge and advance our thinking.

Reviewed by Mike Baynham
University of Leeds, UK
E-mail: mike.baynham@education.leeds.ac.uk
doi:10.1093/applin/amq012 Advance Access published on 22 April 2010

REFERENCES

Bucholtz, M. 1999. '"Why be normal?": language and identity practices in a group of nerd girls,' *Language and Society* 28: 203–23.

Coates, J. 1996. *Women Talk.* Blackwell.

Hymes, D. 1996. *Ethnography, Linguistics and Narrative Inequality. Towards an Understanding of Voice*. Taylor and Francis.

Labov, W. 1972. *Language in the Inner City.* University of Pennsylvania Press.

Labov, W. and **D. Fanshel.** 1977. *Therapeutic Discourse.* Academic Press.

Labov, W. and **J. Waletzky.** 1967. 'Narrative analysis: oral versions of personal experience' in J. Helm (ed.): *Essays on the Verbal and Visual Arts*. University of Washington Press.

Latour, B. 1987. *Science in Action.* Harvard University Press.

Roberts, C. and **S. Campbell.** 2005. 'Fitting stories into boxes: textual and rhetorical constraints on candidates' performances in British job interviews,' *Journal of Applied Linguistics* 2/1: 45–73.

Zimmerman, D. H. 1998. 'Identity, context and interaction' in C. Antaki and S. Widdicombe (eds): *Identities in Talk*. Sage.

Thomas Farrell: REFLECTIVE LANGUAGE TEACHING: FROM RESEARCH TO PRACTICE. Continuum, 2007.

Two concepts currently in vogue in language teacher education are collaboration and reflection. Both are related, although reflection is probably the superordinate concept. It is possible to be reflective without collaborating, although I would argue that collaboration can greatly increase the quality of reflection. Without reflection, collaboration is of little value. Farrell covers both concepts in his book. While the focus is on reflection, collaboration also features prominently, particularly in the chapters on teacher development groups, classroom observations, and critical friendships.

I should declare my position at the outset. This is a book that I wish I had written! While it covers some of the same terrain as that traversed by a book I did co-author (Bailey *et al.*, 2001), it does so from a different angle. Additionally, each book covers topics not covered by the other, and is therefore complementary. In fact, I used both in a professional development program that I taught not so long ago. Concepts are presented with admirable clarity, and Farrell's voice as well as his extensive experience in language teacher education and development are evident on almost every page.

Each chapter in the book follows a set pattern, which gives a sense of coherence to the volume: an introduction, review of the literature, a case study relating to the topic at hand drawn from Farrell's own experience, a section entitled *From research to practice*, which sets out practical ideas for getting started on implementing the topic, and a chapter scenario, which is, in effect, another mini-case study based on someone else's experience. Each of the main sections in the book is followed by a set of questions for the reader to reflect on. The book thus becomes a training manual, not only just for developing skills in action research, keeping journals, classroom observation, and so on, but also for developing a reflective attitude on the part of the reader. In this way, the medium becomes the message.

While the substantive focus of the majority of chapters is on classroom management and methodology, that is, on teaching/learning, one chapter is devoted to language proficiency. Given the fact that the majority of foreign language instructors are non-native speakers of the language they teach, this is a critically important issue. In the case of English, the explosion in the demand for English globally has driven many institutions, both public and private, to employ as English teachers, practitioners whose own command of the language may be inadequate. I realize that this begs the question of exactly what *is* an adequate command of English for teaching purposes. Putting that aside, it is good to see books such as this dealing directly with an issue which is too often either ignored or overlooked.

Although the subtitle of the book is 'From research to practice', the focus is firmly on practice, and the heart of the volume resides in the rich array of case studies and scenarios as well as the reflection points that punctuate each chapter. This is a book for practitioners: teachers and teacher educators—a ‘how to’ volume. I imagine that researchers will be somewhat underwhelmed by the research sections that initiate each chapter and that are intended to summarize what the research has to say about the topic at hand. For me this is not a problem. This book is unashamedly practice oriented. However, enough signposts are provided to the relevant research literature for readers who want to look in greater detail at the empirical basis of particular topics.

Reviewed by David Nunan
Hong Kong, China and Anaheim, USA
E-mail: david.nunan@gmail.com

doi:10.1093/applin/amq013 Advance Access published on 25 April 2010

REFERENCE

Bailey, K., A. Curtis, and **D. Nunan.** 2001. *Pursuing Professional Development: The Self as Source.* Heinle/Cengage.

Christiane Dalton-Puffer: DISCOURSE IN CONTENT AND LANGUAGE INTEGRATED LEARNING (CLIL) CLASSROOMS. John Benjamins, 2007.

During the last two decades or so, the use of English in teaching different subjects, such as mathematics or music, which are taught at all levels of mainstream schools where English is not the first language of students/teachers, has been a very controversial issue. Some of the questions asked include: Where do we put the emphasis on in those classes, language or content? How about the role of foreign language in the construction of meaning? What are the roles of teachers? An in-depth analysis of Content and Language Integrated Learning (CLIL) classroom instruction is therefore both necessary and inevitable. This book addresses this need by providing a comprehensive analysis of classroom discourse in CLIL classrooms. The manuscript is based on a research project that was designed, within the light of constructivist and participatory learning theories, as a predominantly qualitative study of naturalistic classroom interactions in CLIL contexts in Austria. Therefore, the book makes a significant attempt at answering several important questions regarding the features of CLIL classrooms by placing the main emphasis on language use.

The book consists of 10 chapters that can be broadly divided into three major parts. The first two chapters provide a brief overview and historical development of CLIL with a specific focus on the European context and the rationale for the current research project. The author, Christiane Dalton-Puffer, is also chapter. This is a book for practitioners: teachers and teacher educators—a 'how to' volume. I imagine that researchers will be somewhat underwhelmed by the research sections that initiate each chapter and that are intended to summarize what the research has to say about the topic at hand. For me this is not a problem. This book is unashamedly practice oriented. However, enough signposts are provided to the relevant research literature for readers who want to look in greater detail at the empirical basis of particular topics.

Reviewed by David Nunan
Hong Kong, China and Anaheim, USA
E-mail: david.nunan@gmail.com
doi:10.1093/applin/amq013 Advance Access published on 25 April 2010

REFERENCE

Bailey, K., A. Curtis, and **D. Nunan.** 2001. *Pursuing Professional Development: The Self as Source.* Heinle/Cengage.

Christiane Dalton-Puffer: DISCOURSE IN CONTENT AND LANGUAGE INTEGRATED LEARNING (CLIL) CLASSROOMS. John Benjamins, 2007.

During the last two decades or so, the use of English in teaching different subjects, such as mathematics or music, which are taught at all levels of mainstream schools where English is not the first language of students/teachers, has been a very controversial issue. Some of the questions asked include: Where do we put the emphasis on in those classes, language or content? How about the role of foreign language in the construction of meaning? What are the roles of teachers? An in-depth analysis of Content and Language Integrated Learning (CLIL) classroom instruction is therefore both necessary and inevitable. This book addresses this need by providing a comprehensive analysis of classroom discourse in CLIL classrooms. The manuscript is based on a research project that was designed, within the light of constructivist and participatory learning theories, as a predominantly qualitative study of naturalistic classroom interactions in CLIL contexts in Austria. Therefore, the book makes a significant attempt at answering several important questions regarding the features of CLIL classrooms by placing the main emphasis on language use.

The book consists of 10 chapters that can be broadly divided into three major parts. The first two chapters provide a brief overview and historical development of CLIL with a specific focus on the European context and the rationale for the current research project. The author, Christiane Dalton-Puffer, is also the researcher in the project. The book has a language focus, although the author makes it clear that content and language are equally important and in fact inseparable. Therefore, the aim is not to stress the importance of one over the other or to model how to teach science or history. Rather, the general goal is to examine the use of language in CLIL learning settings and provide pedagogical implications. Chapter 3 presents information about the research setting and data collection and analysis methods. The data were collected in seven schools and cover 14 classes, 305 students, and 10 teachers, at secondary level, with the exclusion of schools that receive special funding or other institutional support. The data sources included 'recordings of naturalistic CLIL classroom discourse, teacher interviews in order to access participants' theories about second language learning in general and their own CLIL classrooms in particular, and document analysis of (semi-) official publications on [*sic*: CLIL], as well as the study of specialist SLA, SLL, and CLIL literature' (p. 49).

Chapters 4–8, where the author draws heavily on Conversation Analysis, deal with data analysis and discussion of findings. Each chapter has an overarching question and looks at different features of CLIL classrooms. For example, in Chapter 4, general mechanisms of interactive talk such as topic management, turn taking, repair, and its functions in teaching content subjects are analyzed. Repair is revisited in Chapter 8, which also examines feedback and correction. Chapter 5 addresses the role of questions asked by teachers and students in constructing knowledge. These chapters mainly aim to show 'what the language looks like as a consequence of its employment in content teaching' (p. 65). By analyzing several episodes through conversation analysis techniques, Dalton-Puffer successfully indicates, for example, how repair works in the co-construction of curricular work. The next chapter, Chapter 6, differs from Chapters 4 and 5 as the focus here is on 'academic language functions' (p. 128) with an emphasis on definitions, explanations, and hypothesizing, which were scarcely used in the CLIL classes observed. The author suggests that teachers be encouraged and supported to design and use tasks and activities that promote language and thinking skills in students; these were lacking in the CLIL classes observed. The next chapter, Chapter 7, focuses on classroom politeness and directives and it is concluded that 'the experience of directives that CLIL lessons provide the students with is predominantly passive' (p. 204).

In the remaining two chapters, the author further discusses her findings. Chapter 9 is devoted to answering the question 'how should we teach?' (p. 257) in CLIL classrooms. The study findings are interpreted and discussed within the light of several second language learning theories such as Krashen's monitor model, Long's interaction hypothesis, Swain's output hypothesis, Givon's discourse hypothesis (Givon 1979), and sociocultural theory. Canale and Swain's model of communicative competence is also addressed in this chapter. The final chapter, Chapter 10, presents implications for pedagogical actions and directions for further research. Both the restrictions of CLIL environments and their positive potential are discussed. Dalton-Puffer asks a very crucial question in this final chapter: 'But why should we be doing CLIL at all if there are no language goals present?' (p. 295). She strongly argues that 'language curricula for CLIL should be developed, and language goals in speaking, writing, reading, and listening concretized' (p. 295) in order for CLIL classes to be effective in terms of both teaching content and using the target language.

The broad scope of empirical questions, the prolonged engagement in the research setting, and the approaches used in analyzing discourse make the study a strong one and the book an ideal model for researchers who aim to design large-scale research projects. Goals, research subjects, methodology, and findings of the project are clearly reported. Another major strength of the book is that the text is richly illustrated with close analyses of samples from classroom discourse data recorded in a variety of classroom settings that includes but is not limited to history, biology, and music. The excerpts are well chosen; they tend to be short and engaging. This text can be a useful resource for teachers who are involved in content-based instruction, especially in English as a foreign language (EFL) settings, second language acquistion (SLA) researchers, discourse analysts, and policy makers. It will be particularly helpful for graduate students who want to learn about qualitative research design. As a graduate student, I found the researcher's methodological reflections particularly helpful, for example where the author explains how she negotiated access to the field and developed relationships with the teachers. Her model which discusses role pairs available in the teacher–researcher relationships such as equal-spy, giver-taker, helper-client has also been eye-opening in terms of enabling the reader to realize the wide variety of roles a researcher might take on in conducting qualitative research. Yet, I wish that several student participants had also been interviewed. Given the fact that 'the interaction' under analysis in the text took place between students and teachers and/or student groups, it would be useful to hear students' own voices. This would further help explain some of the study's findings and support the author's hypotheses.

In addition, although the author presents the reader with an overview of conversation analysis, speech acts, and genre analysis as the main tools which she employs in her research and this enables us to see her position as a discourse analyst, a more elaborate section where she specifically talks about her understanding and own definition of 'discourse' would be a plus. This is necessary given the fact that discourse is defined in different ways by various scholars across different disciplines. Nevertheless, the book contributes substantially to the literature on CLIL classrooms, in which it serves as an excellent reference point.

Reviewed by Hayriye Kayi
University of Texas at Austin, USA
E-mail: hkayi@mail.utexas.edu

doi:10.1093/applin/amq015 Advance Access published on 26 April 2010

REFERENCE

Givon, T. 1979. ''From discourse to syntax: Grammar as a processing strategy'' in T. Givon (ed.): *Syntax and Semantics* Academic Press.

Paula Kalaja, Vera Menezes and Ana Maria F. Barcelos (eds): NARRATIVES OF LEARNING AND TEACHING EFL. Palgrave Macmillan, 2008.

This book represents a diversity of research perspectives on English as a foreign language (EFL) narratives in three different continents (South America, Asia, and Europe). Fully aware of the complexity of the field, the editors adopt the metaphor of a *kaleidoscope* to capture the multi-layered and interactive nature of the processes implied in learning and teaching EFL. Each turn of the narrative kaleidoscope gives an insight on the intricacies of learning/teaching context. The conscious move from the 'learning as acquisition' to 'learning as participation' depicts a learner who actively deploys strategies and constructs hypotheses. It highlights an emotional and reflexive dimension of the teaching process, thus presenting participants who try to cope with or resist institutional frameworks. Teaching and learning identities contribute simultaneously to accounts of lived, personal, subjective experiences. The volume's introduction supplies an excellent conceptual map and a description of national EFL contexts in Japan, Finland, and Brazil. Although most contributors are EFL teachers, the book's organization manifests mostly an orientation towards research. The aim is to provide 'a glimpse of the unfolding perspectives and alternative views on narrativising learning and teaching EFL' (p. 232). The book's division into four sections reflects the means of data collection.

Part II on 'written narratives' focuses on language learning histories of Japanese and Brazilian university students learning English (Chapter 2 by Murphey and Carpenter and Chapter 3 by Barcelos) and describes the ideological divide between private and public EFL contexts in Brazil from a teacher's perspective (see especially Chapter 4 by Dutra and Mello and Chapter 5 by Miccoli). The chapters show how narrative inquiry presents an emic perspective on learning/teaching by disclosing personal beliefs, expectations, and strategies, as well as emotional responses to EFL experiences. Whereas Murphey and Carpenter argue for affective relationships between learners and teachers, engaging them in the construction of affinities and shared spaces, Barcelos locates the learners' frustrations with dominant EFL learning ideologies in Brazil and calls for political action towards better quality education which transforms EFL practice. Dutra and Melo, as well as Miccoli, stop short of demanding political transformation and suggest an inside-out change which is based on reflective teaching and less teacher-centred pedagogy.

REFERENCE

Givon, T. 1979. “From discourse to syntax: Grammar as a processing strategy” in T. Givon (ed.): *Syntax and Semantics* Academic Press.

Paula Kalaja, Vera Menezes and Ana Maria F. Barcelos (eds): NARRATIVES OF LEARNING AND TEACHING EFL. Palgrave Macmillan, 2008.

This book represents a diversity of research perspectives on English as a foreign language (EFL) narratives in three different continents (South America, Asia, and Europe). Fully aware of the complexity of the field, the editors adopt the metaphor of a *kaleidoscope* to capture the multi-layered and interactive nature of the processes implied in learning and teaching EFL. Each turn of the narrative kaleidoscope gives an insight on the intricacies of learning/teaching context. The conscious move from the ‘learning as acquisition’ to ‘learning as participation’ depicts a learner who actively deploys strategies and constructs hypotheses. It highlights an emotional and reflexive dimension of the teaching process, thus presenting participants who try to cope with or resist institutional frameworks. Teaching and learning identities contribute simultaneously to accounts of lived, personal, subjective experiences. The volume’s introduction supplies an excellent conceptual map and a description of national EFL contexts in Japan, Finland, and Brazil. Although most contributors are EFL teachers, the book’s organization manifests mostly an orientation towards research. The aim is to provide ‘a glimpse of the unfolding perspectives and alternative views on narrativising learning and teaching EFL’ (p. 232). The book’s division into four sections reflects the means of data collection.

Part II on ‘written narratives’ focuses on language learning histories of Japanese and Brazilian university students learning English (Chapter 2 by Murphey and Carpenter and Chapter 3 by Barcelos) and describes the ideological divide between private and public EFL contexts in Brazil from a teacher’s perspective (see especially Chapter 4 by Dutra and Mello and Chapter 5 by Miccoli). The chapters show how narrative inquiry presents an emic perspective on learning/teaching by disclosing personal beliefs, expectations, and strategies, as well as emotional responses to EFL experiences. Whereas Murphey and Carpenter argue for affective relationships between learners and teachers, engaging them in the construction of affinities and shared spaces, Barcelos locates the learners’ frustrations with dominant EFL learning ideologies in Brazil and calls for political action towards better quality education which transforms EFL practice. Dutra and Melo, as well as Miccoli, stop short of demanding political transformation and suggest an inside-out change which is based on reflective teaching and less teacher-centred pedagogy.

Part III on ‘self-narratives’ engages in a reflexive analysis of the lived emotional struggles of teachers who face student resistance (Chapter 7 by Sakui and Cowie) and respond to other people’s learning experiences (Chapter 6 by Karlsson). The focus here is on EFL contexts in Finland and Japan. Narrative inquiry helps to reveal the tacit knowledge implied in both processes, yet Karlsson takes this further by linking teachers and students in a ‘recycling’ process of learning experiences, this way creating a ‘kaleidoscopic picture of their learners and themselves, not a microscopic one’ (p. 87). In Part III, on ‘oral narratives’, Cotterall (Chapter 8) applies narrative research to explore individual learners’ management of evolving motivations over time. This is the key issue of a life history project by Murray in Japan (Chapter 9) which illustrates how English learners ‘become members of a variety of communities of practice, both immediate and imagined’ (p. 131). Identities in the making are examined by Block (in Chapter 10) and Chik and Benson (Chapter 11). Block stresses, in a story of an adult EFL learner in Spain, that identity work develops in close connection to immediate communities of practice, rather than in relation to the English language. He warns against ‘overemphasising individual agency’ (p. 143) and urges us to consider it as both ‘constitutive of and constituted by social structure’ (p. 143). The same point is illustrated by Chik and Benson’s description of the destabilizing sense of identity in the case of a postgraduate student from Hong Kong who is being positioned as an EFL learner at a UK university.

Finally, the part of the book which deals with ‘multimodal narratives’ focuses on the use of photographs, drawings, and multi-modal resources to examine the role of EFL in Finnish teenagers’ everyday lives (Chapter 12 by Nikula and Pitkänen-Huhta), beliefs about EFL teaching/learning among learners in Finland (Chapter 13 by Kalaja, Alanen, and Dufva), and EFL learning experiences in Brazil (Chapter 14 by Menezes). These chapters reveal how the identities of learner and language user ‘get intertwined in [the] participants’ stories’ (p. 171) and agree on the cross-influences between learning activities inside and outside the classroom (p. 212). The accounts are viewed as multi-voiced meta-narratives. They are analysed within the variety of their meaning-making potentials and situated within particular socio-cultural contexts (p. 198).

The book is shaped both by a kaleidoscopic account of EFL experiences from multiple contexts (formal, informal, national) and a variety of disciplinary fields. The contributors in it draw on a wide range of theoretical constructs, including the philosophy of language and education, sociolinguistics, applied linguistics, social psychology, psychotherapy, etc., while insights from discourse analysis, literacy studies, cultural–historical analysis, critical pedagogy, action research, etc. evidence a poststructuralist approach to the analysis of language and language learning. The volume’s contributors view language as inherently social and therefore consider learning from within a participant perspective of socialisation (pp. 170–172). The spotlight is ‘from the student’s perspective and within the context of the student’s life’ (p. 156). The use of learning histories, journals, diaries, field-notes, and semi-structured interviews underline the book's orientation to qualitative and ethnographic methods, while participant observation and collaborative research are used as strategies of triangulation. One can also note a distinction between analyses of narratives (Chapters 2–5, 8–10, and 12–14) and narrative analysis (Chapters 6, 7, and 11). The volume's contributors differ in their use of narrative as a method: for some it is 'both the phenomenon and the method' (p. 85), while for others it is 'a form of representation rather than a mode of analysis' (p. 62). However, the chapters share the view that narratives provide participants with agency, situating experiences within time and space. Narratives are thus particularly valuable for the analysis of identities, since 'identities are discursively constructed through [them]' (p. 172).

Along with the socially oriented traditions on language and education (Martin-Jones and Heller 1996; Block 2003; Pavlenko and Blackledge 2004), the volume authors discuss identities rather than an identity, 'in part conditioned by social and cultural factors (including gender, nationality, ethnicity, class and language repertoire) and by the ways others see us [...], in part by individual agency and negotiated through ongoing narratives' (p. 156). Block (this volume) explains that these variables 'do not stand independently of one another in the larger general identity' (p. 143). Traces of identity (including previous educational experiences) interact in a language classroom (p. 106), which becomes a space where not only language is taught but also identities are negotiated, imposed, accepted, and resisted (Chapters 7, 10, and 11).

As the authors consider learning as a socio-historically situated phenomenon, and a language classroom as a culture (p. 65), they also stress how learners and teachers participate in, negotiate, and construct language and literacy practices. They develop shared ways of doing and making meaning and become members of real and imaginary communities of practice (Lave and Wenger 1998; Murray this volume) which are not contained within the classroom but extend into family life, the consumption of literature, popular culture, and sports (Chapters 3, 8, 9, 11, and 14). While highlighting the multi-faceted role which communities of practice can play in language learning, Murray, however, stays 'unclear how educators can apply the notion to classroom and other learning contexts' (p. 138). It is worth noting here that Rogoff *et al.* (2002) developed the notion of a 'community of learners' from a different perspective and have successfully implemented it in a school setting.

Overall, the volume clearly illustrates the social turn in second language acquisition (SLA) (Block 2003), in which learners are seen as 'complex social beings' (Canagarajah 2004). It not only offers many practical considerations about EFL teaching and learning, but also discloses some dominant and tacit language learning ideologies. As a collection of narratives, it addresses the 'urgent need for accounts from inside' (Baynham and de Fina 2005), for instance, by dispelling myths of 'passive' Japanese learners, by bridging the EFL gap between private and public schools in Brazil, by validating the role of informal and incidental learning across the three continents, and ultimately by depicting autonomous learners and reflective teachers, as well as providing understandings of the dynamic nature of their socio-historical contexts.

The volume acknowledges the fragmented, partial, and situated nature of any interpretation of social phenomena. It argues for a three-dimensional perspective of narrative inquiry, made up by temporality, human agency, and place (p. 88). Although most authors link participants' experiences to larger social patterns, the aspect of socio-historical and discursive embeddedness could have been explored more explicitly by addressing indexicalities of peer cultures, institutional, and societal discourses and by looking at experiences as critical resources for social identification (Norton and Toohey 2004; Blommaert 2005; Bartlett 2007). The view that educational sites are 'power-laden' (Canagarajah 2004: 117) helps us understand teacher/learner identities better. Language learning needs to be viewed as constituted and constrained by relationships of dominance, as language counts as 'a set of unequally distributed resources' (Heller 2007: 2).

Reviewed by Olga Solovova
Centre for Social Studies, Coimbra, Portugal
E-mail: olga@ces.uc.pt

doi:10.1093/applin/amq016 Advance Access published on 26 April 2010

REFERENCES

Bartlett, L. 2007. 'Bilingual literacies, social identification, and educational trajectories,' *Linguistics and Education* 18/1: 215–231.

Baynham, M. and **A. de Fina** (eds). 2005. *Dislocations/Relocations. Narratives of Displacement.* St Jerome Publishing.

Block, D. 2003. *The Social Turn in Second Language Acquisition.* Georgetown University Press.

Blommaert, J. 2005. *Discourse: A Critical Introduction.* Cambridge University Press.

Canagarajah, S. 2004. 'Subversive identities, pedagogical safe houses, and critical learning' in B. Norton and K. Toohey (eds): *Critical Pedagogies and Language Learning*Cambridge University Press, pp. 116–37.

Heller, M. (ed.) 2007. *Bilingualism: A Social Approach.* Palgrave Macmillan.

Lave, J. and **E. Wenger.** 1998. *Communities of Practice: Learning, Meaning, and Identity.* Cambridge University Press.

Martin-Jones, M. and **M. Heller** (eds). 1996. *Education in Multilingual Settings: Discourse, Identities, and Power. Part II: Contesting Legitimacy. Linguistics and Education,* vol. 8/2.

Norton, B. and **K. Toohey** (eds). 2004. *Critical Pedagogies and Language Learning.* Cambridge University Press.

Pavlenko, A. and **A. Blackledge** (eds). 2004. *Negotiation of Identity in Multilingual Contexts.* Multilingual Matters.

Rogoff, B., C. Goodman Turkanis, and **L. Bartlett.** 2002. *Learning Together: Children and Adults in a School Community.* Oxford University Press.

Applied Linguistics: 31/3: 482–485
doi:10.1093/applin/amq018

NOTES ON CONTRIBUTORS

Niclas Abrahamsson is an Associate Professor in Bilingualism and Director of the Centre for Research on Bilingualism, Stockholm University. His research interests include maturational constraints and critical-period effects, language learning aptitude, and phonological/phonetic aspects of first and second language acquisition. His work has appeared in *Bilingualism: Language and Cognition, Language Learning, Studia Linguistica, Studies in Second Language Acquisition, TESOL Quarterly*, and in Doughty and Long's (eds. 2003) *The Handbook of Second Language Acquisition* (with Kenneth Hyltenstam). *Address for correspondence*: Centre for Research on Bilingualism, Stockholm University, SE-106 91 Stockholm, Sweden. <*niclas.abrahamsson@biling.su.se*>

Mike Baynham is a Professor of TESOL at the University of Leeds. He is a co-convenor with Stef Slembrouck of the AILA research network on Language and Migration. He has a long-standing research interest in narrative: he and Anna de Fina edited *Dislocations/Relocations: Narratives of Displacement* (St-Jerome Publishing, 2005). *Address for correspondence*: School of Education, University of Leeds, Leeds LS2 9JT, UK. <*mike.baynham@education.leeds.ac.uk*>

Emanuel Bylund, PhD in Bilingualism and PhD in Spanish, is a researcher at the Centre for Research on Bilingualism, Stockholm University. His research interests include maturational constraints in language acquisition and loss, language aptitude, and the relationship between language and thought. His work has appeared in *Bilingualism: Language and Cognition, International Journal of Bilingualism, Language Learning, Revue Romane*. *Address for correspondence*: Centre for Research on Bilingualism, Stockholm University, SE-106 91 Stockholm, Sweden. <*manne.bylund@biling.su.se*>

Sonia Casal (Universidad Pablo de Olavide) has been training teachers working in bilingual schools all over Spain for the last five years. She is the co-ordinator of the Masters in Bilingual Education at Universidad Pablo de Olavide. Her research interests focus on cooperative learning, the integrated curriculum and bilingual education and she has participated in international conferences on the subject in Spain, Germany, Italy, Estonia, and Brazil. She also has published chapters in edited collections for Cambridge Scholars and Peter Lang. <*scasmad@upo.es*>

Debra Friedman is an Assistant Professor of Second Language Studies in the Department of Linguistics and Germanic, Slavic, Asian, and African Languages at Michigan State University. She specializes in the application of discourse analysis and language socialization approaches to the study of language classroom interaction, with a focus on the sociocultural, political, and ideological aspects of language education and language education policy. *Address for correspondence*: Department of Linguistics and Germanic, Slavic, Asian, and African Languages, Michigan State University, A-701 Wells Hall, East Lansing, MI 48824, USA. <*fried106@msu.edu*>

Claudia Gentile is affiliated with *Mathematica Policy Research* and currently working as a consultant for Educational Testing Service (ETS). She worked as a researcher in the Research and Development (R&D) division at ETS for more than 15 years. *Address for correspondence*: 228 Carton Avenue, Ewing, NJ 08618, USA. <*claudiagentile7@yahoo.com*>

Richard Herrington is Statistical Consultant Team Leader in the Computing and IT Center at the University of North Texas with a PhD in Psychology. *Address for correspondence*: University of North Texas, PO Box 311307, Denton, TX 76203, USA.

Kenneth Hyltenstam is Professor of Bilingualism at Stockholm University. His research includes contributions on variation and developmental sequences, typological markedness and maturational constraints in second language acquisition, code-switching among bilinguals suffering from Alzheimer's dementia, and minority language maintenance and shift. He is presently exploring exceptional language learning abilities in hyperpolyglots. His work has appeared in *Bilingualism: Language and Cognition, Language Learning, Studia Linguistica, Studies in Second Language Acquisition, and TESOL Quarterly*. *Address for correspondence*: Centre for Research on Bilingualism, Stockholm University, SE-106 91 Stockholm, Sweden. <*kenneth.hyltenstam@biling.su.se*>

Robert Kantor is currently working as Principal Assessment Specialist in the English Language Learning group in the R&D division at ETS. He has been involved in research and development activities for TOEFL writing assessment. He has produced a significant number of TOEFL/ETS research reports and research articles in refereed journals, such as *The Modern Language Journal, Language Testing*, and *International Journal of Testing*. *Address for correspondence*: MS 40-N Rosedale Road, Princeton, NJ 08541, USA. <*rkantor@ets.org*>

Hayriye Kayi is a PhD candidate in Foreign Language Education with a specialization in Teaching English as a Second Language at the University of Texas at Austin. Her research interests include critical-micro analysis of ESL classroom discourse, positioning theory, and identity. *Address for correspondence*: Foreign Language Education, Department of Curriculum and Instruction, the University of Texas at Austin, 1 University Station, D 6500, Austin, TX 78712-0379, USA. <*hkayi@mail.utexas.edu*>

Jenifer Larson-Hall is an Assistant Professor of Applied Linguistics at the University of North Texas and author of the recent book, *A Guide to Doing Statistical Analysis in Second Language Research*. *Address for correspondence*: University of North Texas, PO Box 311307, Denton, TX 76203, USA. <*jenifer@unt.edu*>

Yong-Won Lee is Assistant Professor in the Department of English Language and Literature at Seoul National University (SNU). Prior to assuming a faculty position at SNU, he worked as a researcher in the TOEFL research groups at Educational Testing Service (ETS) for eight years. He has produced a significant number of TOEFL/ETS research reports and research articles in refereed journals, such as *Language Testing, Language Assessment Quarterly, and International Journal of Testing*. He is currently serving on the editorial advisory boards of *Language Assessment Quarterly* and the *Korean Association for the Study of English Language and Linguistics Journal*. *Address for correspondence*: Department of English Language and Literature, Seoul National University, Gwanak-ro 599, Sillim9-dong Gwanak-gu, Seoul 151-745, South Korea. <*ylee01@snu.ac.kr*>

Francisco Lorenzo (Universidad Pablo de Olavide) has been involved in bilingual education for over 20 years and has given seminars in the UK, the USA, Germany, Poland, and Russia. He has articles in peer-reviewed journals such as *Language Learning* and *Language and Education* and chapters in edited collections with Peter Lang and John Benjamins. He is a member of the AILA CLIL ReN, the Thematic Network of Languages (European Commission), LANQUA (University of Southampton), and MOLAN (Frei Universität, Berlin). He is about to publish the first handbook on bilingual education in Spanish. <*fjlorber@upo.es*>

Pat Moore (Universidad Pablo de Olavide) has lived and taught in France, Portugal, Greece, Brazil, China, and now Spain. Her research focuses on L2 oracy (teaching, learning, and evaluation). She is the coordinator of the CLIL Tasks and Activities module in the Bilingual Education Masters and of the Bilingual Education module in the Secondary Education Masters. She has published CLIL-related articles and chapters in RESLA (Revista Española de Lingüística Aplicada), ViewZ (Vienna English Working Papers), ICRJ (International CLIL Research Journal) and in a forthcoming John Benjamins collection. <*pfmoox@upo.es*>

David Nunan is Emeritus Professor of Applied Linguistics at the University of Hong Kong and Vice-President of Academic Affairs, Anaheim University, California. <*david.nunan@gmail.com*>

Kanavillil Rajagopalan is a Professor of Linguistics at the State University at Campinas (Unicamp), Brazil. He was born in India and pursued his studies in that country, as well as in the UK, Brazil, and the USA. He has published four books and upwards of 300 papers in journals and has contributed chapters to various edited volumes, encyclopedias, handbooks, etc. His academic interests include the philosophy of language, linguistic pragmatics, postcolonial literature, and the spread of this new linguistic phenomenon called 'World English'. *Address for correspondence*: Rua Cardoso de Almeida 898, Apto: 51, Perdizes, São Paulo – SP, 05013-001 Brazil. <*rajagopalan@uol.com.br*>

Olga Solovova is a doctoral student at the Centre for Social Studies at the University of Coimbra. She develops research on multi-lingual communicative practices by Eastern European school children within the broader social process of integration of immigrants in Portuguese society. Publications include 'Multilingual dynamics among Portuguese-based migrant contexts in Europe' (*Journal of Pragmatics*, with C. Keating), 'A new stereotype in the making: Imigrantes de Leste in the Press' (in Barker, A (ed.), *The Power and Persistence of Stereotyping*, University of Aveiro, 2005) and 'Progressive markers in the creoles of Cabo Verde and Guiné-Bissau: the outcome of different sociolinguistic histories (in Fernandez, M. *et al.* (eds), *Los criollos de base ibérica*, ACBLPE, 2003). *Address for correspondence:* Centre for Social Studies, Colegio de S.Jeronimo, Apartado 3087 3001-401, Coimbra, Portugal. <*olga@ces.uc.pt*>

Ivor Timmis is Reader in English Language Teaching at Leeds Metropolitan University, where he teaches on the MA in Materials Development for ELT and supervises PhD students. He has a strong interest in both the descriptive and pedagogic aspects of spoken language research and in the relevance of corpus findings for language teaching. <*i.timmis@leedsmet.ac.uk*>