Routledge Studies in Applied Linguistics

DEFINING AND ASSESSING LEXICAL PROFICIENCY

Agnieszka Leńko-Szymańska

Defining and Assessing Lexical Proficiency

This comprehensive account of performance-based assessment of second language (L2) lexical proficiency analyses and compares two of the primary methods of evaluation used in the field and unpacks the ways in which they tap into different dimensions of one model of lexical competence and proficiency. It also juxtaposed performance-based assessment with discrete-point tests of vocabulary.

This book builds on the latest research on performance-based assessment to systematically explore the qualitative method of using human raters and the quantitative method of using statistical measures of lexis and phraseology. Supported by an up-to-date review of the existing literature, both approaches' unique features are highlighted but also compared to one another to provide a holistic overview of performance-based assessment as it stands today at both the theoretical and empirical level. These findings are exemplified in a concluding chapter, which summarises results from an empirical study looking at a range of lexical and phraseological measures and human raters' scores of over 150 essays written by both L2 learners of English and native speakers as well as their vocabulary tests results. Taken together, the volume challenges existing tendencies within the field, which attempt to use one method to validate the other, by demonstrating their propensity to capture very different aspects of lexical proficiency, thereby offering a means by which to better conceptualise performance-based assessment of L2 vocabulary in the future.

This book will be of interest to students and researchers working in second language acquisition and applied linguistics research, particularly those interested in issues around assessment, vocabulary acquisition, and language proficiency.

Agnieszka Leńko-Szymańska is Assistant Professor at the Institute of Applied Linguistics at the University of Warsaw, Poland.

Routledge Studies in Applied Linguistics

Grounded Theory in Applied Linguistics Research
A Practical Guide
Gregory Hadley

Project-Based Language Learning with Technology
Learner Collaboration in an EFL Classroom in Japan
Michael Thomas

Metacognition in Language Learning and Teaching
Edited by Åsta Haukås, Camilla Bjørke and Magne Dypedahl

Language Management and Its Impact
The Policies and Practices of Confucius Institutes
Linda Mingfang Li

Multiliteracies, Emerging Media, and College Writing Instruction
Santosh Khadka

Cantonese as a Second Language
Issues, Experiences and Suggestions for Teaching and Learning
Edited by John Wakefield

The Social Lives of Study Abroad
Understanding Second Language Learners' Experiences through Social Network Analysis and Conversation Analysis
Atsushi Hasegawa

Defining and Assessing Lexical Proficiency
Agnieszka Leńko-Szymańska

For more information about this series, please visit: www.routledge.com/Routledge-Studies-in-Applied-Linguistics/book-series/RSAL

Defining and Assessing Lexical Proficiency

Agnieszka Leńko-Szymańska

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2020
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-0-367-33792-6 (hbk)
ISBN: 978-0-429-32199-3 (ebk)

Typeset in Sabon
by Apex CoVantage, LLC

In loving memory of
my brother
Krzysztof Leńko
and my son
Jan Szymański

till we all meet again


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Acknowledgements

Writing this book has been a long and laborious process. I would not have succeeded if it had not been for the help and support of numerous people. Carrying this project through to its completion finally gives me an opportunity to express my deep gratitude to them.

My special thanks go to Professor Barbara Lewandowska-Tomaszczyk, who has been my mentor for the last 30 years. Her first review of the manuscript reminded me that any applied linguistic research must have a proper grounding in a sound theoretical model of language.

I would like to thank Dr. habil. Ewa Gruszczyńska and Dr. Anna Szczęsny, Director and Deputy Director of the Institute of Applied Linguistics, and Dr. habil. Elżbieta Gajek, Head of the Department of Second Language Studies, for taking over many of my administrative duties, so that I could concentrate on completing this monograph. I would also like to thank Dr. habil. Łucja Biel, Dr. habil. Ewa Gruszczyńska, Dr. habil. Michał Paradowski, Dr. habil. Markus Eberharter and Dr habil. Radosław Kucharczyk for their advice concerning the choice of the right academic career tactics and their explanations of the meanders of the Polish *habilitation* process. Many thanks to all my colleagues from the Institute of Applied Linguistics for their encouragement and support.

My appreciation also goes to all my friends, who have seconded me in my endeavours. I need to mention specifically Kate Woodward, who has considerately coached me on how to choose my priorities and remain focused on them. I am deeply grateful to my almost lifelong friend Dr. Joanna Kazik, who has always been there for me in difficult times. Although this book is not a novel, Joanna will probably find herself on its pages. I would also like to express my appreciation to Patrycja Gołębicka for her relentless concern and help and Barbara Zakrzewska for her trustworthy aid in mundane everyday chores.

I would also like to express my gratitude to my British relatives, the Wojciechowskis, my American foster family Benjamin and Dr. Susan Uchitelle, and the former Fulbright US scholar at University of Łódź,

Professor David Pichaske for taking me under their wings when I was still a student, and for showing me that where there is a will there is a way.

And finally, my deepest gratitude to my family for their love, patience, understanding and support. Without you, none of this would have been possible or worthwhile.

Introduction

Assessment is an inherent part of second language learning. Information on learners' current linguistic competence and proficiency level is of key importance in both language instruction and research. In education, feedback on students' progress and their current linguistic ability is valuable to the students themselves, but also to their teachers and to administrative bodies. This information is also frequently used for gatekeeping purposes, as many educational programmes or job descriptions specify a required standard of foreign language skills for their applicants. Moreover, a lot of other decisions related to language instruction are based on information about students' level. Organising language courses, creating or selecting teaching materials and even developing or choosing appropriate exams requires specifying a target group of learners, and their proficiency level is the most important characteristic of that group. Instruction is most effective if students work with materials and on tasks suitable for their current level and—provided they study in a group—if other members of that group represent the same or similar linguistic ability. In research into second language acquisition (SLA), in turn, any description of the learner's language is only relevant if it takes his or her proficiency into account. Interlanguage is very dynamic, which means that the linguistic system underlying the learner's performance is constructed, deconstructed and restructured all the time. The traits and processes shaping interlanguage may function differently at different stages of advancement. Thus, determining the linguistic competence and the proficiency level of the subjects studied is the first step in any SLA research.

The learner's linguistic command is an abstract concept and it can only be assessed indirectly by studying and analysing samples of his or her performance. There have been different approaches to what kind of performance is most useful and adequate for evaluation purposes and how it should be elicited from learners. On the one hand, the indirect discrete-point approach has emerged, which acts on the assumption that language can be broken down to its component parts and these parts can

be elicited and assessed separately. This approach gave rise to second language tests which consist of many different types of items such as multiple choice, gap-filling, transformation or reordering. These tasks address specific linguistic problems such as the control of the form, meaning and use of particular words or of particular grammatical constructions or else the familiarity with particular aspects of discourse structure. Within this approach the learner's proficiency level is assessed based on the number of questions that he or she can correctly solve, which is an indication of what proportion of the vast pool of specific linguistic elements (vocabulary, grammar or discourse building elements), judged as representing different levels of difficulty, he or she has mastered. However, more recently a new trend in language assessment has proposed that evaluation of linguistic proficiency should be based on samples of the learner's natural use of language for genuine communication. This is the pragmatic approach, which requires that language tests and exams engage learners in authentic language use: both comprehension and production. Thus, modern language tests elicit from learners, among other things, samples of performance in the form of extended written and spoken production.

One way to assess the learner's proficiency based on his or her extended language production is using human raters. They are expected to scrutinise the totality of a learner's written or spoken performance and make a judgement of its quality, which is supposed to be a reflection of the learner's linguistic ability. Raters are equipped with guidelines in the form of assessment scales. In order to ensure consistent interpretation of these scales, rates training is provided. This method of assessment is frequently used in language testing. Researchers in second language acquisition, on the other hand, have proposed the methods of assessing learner data which are based on so-called developmental indices (Wolfe-Quintero, Inagaki, & Kim, 1998). These indices can be defined as independent and objective measures which gauge language progress and which are not tied to specific grammatical structures or lexis. Developmental indices are not measures of language proficiency per se, but they can depict an increase in the learner's linguistic ability. Wolfe-Quintero et al. (1998) explain the difference between the two in the following way:

> Language development refers to characteristics of a learner's output that reveal some point or stage along a developmental continuum. This includes developmental measures such as the number of clauses per T-unit, which are assumed to progress in a linear fashion as language use develops (Hunt, 1965). Language proficiency is a broader concept that is related to separating language users into cross-sectional groups based on normal distribution of their language abilities.
>
> (Wolfe-Quintero et al., 1998, pp. 1–2)

The authors list a number of indices which have been proposed in the research on second language acquisition with a view of describing the learner's developmental level in precise terms. These measures usually apply to one of the components of proficiency such as vocabulary or grammar and tap one of three aspects of development: complexity, accuracy and fluency (Housen & Kuiken, 2009; Bulté & Housen, 2012). It has widely been assumed that each of these aspects increases as a learner becomes more proficient, but as Wolfe-Quintero et al. (1998) point out this increase does not have to happen simultaneously.

This book focuses on one particular component of language proficiency: vocabulary. The central role that lexis plays in communicative language ability has long been acknowledged. It is vocabulary that constitutes the essential building blocks of language (Schmitt, Schmitt, & Clapham, 2001). As Wilkins (1972, p. 111) states, "without grammar very little can be conveyed, without vocabulary nothing can be conveyed". For the last 30 odd years, vocabulary has enjoyed a spur of interest in various areas of linguistic inquiry, including psycholinguistics and the study of second language acquisition. New facts about lexis as one of the language systems, and about the functioning of the monolingual and bilingual lexicon, have come to light, such as the interdependence of vocabulary and other linguistic components, the complexity of word knowledge, the intricate organisation of lexical competence and the importance of multi-word expressions in language processing. These discoveries have had their effect on the assessment of second language (L2) vocabulary for both educational and research purposes. Instead of evaluating learners' familiarity with selected lexical items, test designers attempt to estimate the size, depth, organisation and accessibility of learners' lexicons as a whole. Another important trend which has shaped the evaluation of L2 vocabulary in recent years is the new approach to language assessment, i.e. pragmatic language testing discussed previously. As a result of these influences, the issues which have recently come to the foreground in the evaluation of L2 vocabulary are not the lexical competence as such, but the ability to apply this competence in authentic communication, as evidence of larger reading, listening, speaking and writing proficiencies.

The purpose of this book is to review various ways of assessing lexical proficiency. More specifically, three approaches will be scrutinised: the evaluation carried out by means of discrete-point vocabulary tests as well as evaluation of learners' extended production performed by human raters and by a range of developmental indices. The book will compare and juxtapose these three different ways of assessment, highlighting their strong points and their weaknesses. It will also explore to what extent they relate to one another as well as to the models of lexical proficiency which will also be presented and discussed. The focus of the empirical study reported in the volume will be on advanced learners of English,

whose lexical proficiency is fairly developed. It will examine which aspects of lexical command come to the foreground in the evaluation of extended written production of this type of learners and how the three approaches to assessment are suitable for this task.

The volume consists of six chapters and conclusions. The place of vocabulary in the overall linguistic system and various definitions and models of lexical competence and lexical proficiency are presented and discussed in Chapter 1. The chapter examines the components of word knowledge, including the familiarity with associated phraseology, but it also approaches the description of the lexicon from a holistic perspective in terms of its breadth, depth, internal structure and access. Chapter 2 reviews the three approaches to assessing L2 vocabulary and discusses the instruments for eliciting learners' performance applied by each of them. It also presents the most popular discrete-point test items used in education and SLA research. The next chapters of the book are devoted specifically to the assessment of vocabulary based on learners' written and spoken extended production. Chapter 3 discusses the characteristics of performance-based assessment and focuses on the evaluation of vocabulary performed by human raters as part of holistic assessment of language proficiency or specifically as assessment of lexical proficiency. Global and analytic scales used in the most popular tests of language proficiency are examined, in particular the role of the overall lexical proficiency as well as its components in these scales. The descriptors related to vocabulary use for each proficiency band are also scrutinised. Special attention is given to the scales and descriptors proposed by the Common European Framework of Reference, as the standards introduced by this document have become widely recognised and applied across and even beyond Europe. Chapter 4, in turn, presents and analyses various statistical developmental measures which have been proposed in literature for the purpose of assessing learners' use and acquisition of lexis based on samples of their production. A comprehensive review of several types of measures tapping different aspects of lexical quality is conducted. The chapter will also report on most recent attempts at quantifying phraseological proficiency, which is strongly related to lexical proficiency, and which also plays a role in evaluation of text quality.

While Chapters 3 and 4 contain analyses of the two methods of performance assessment based mostly on theoretical considerations, Chapter 5 examines and juxtaposes these two approaches empirically. It reviews most influential studies which investigated how various lexical indices distinguish between learner groups at different levels, or how their results are related to the assessment of lexical proficiency performed by human raters, and to holistic evaluation of an individual's global proficiency level. Chapter 6 reports on a new empirical study which analysed a whole host of lexical and phraseological indices as well as raters' scores produced for 150 essays written by L2 learners of English at upper-intermediate

and advanced levels as well as English native speakers. The results of vocabulary tests administered to the L2 learners will also be presented and discussed. In addition to the quantitative scrutiny of the results, the report will also contain an examination of qualitative data consisting in interviews with the raters. The conclusion pulls together the ideas and issues discussed in the six chapters and sums up the strengths and weakness of each of the three approaches to assessment of vocabulary ability. It also proposes an extended model of lexical proficiency.

The analysis of the three methods of assessing the L2 learner's lexis and the discussion of the results of a multitude of studies in this area—including the one conducted by the author—will shed a new light on lexical proficiency and foster a better understanding of this construct. They will also contribute to a more precise definition of effective vocabulary use, the concept which at the moment is based on intuition, but which has not been adequately operationalised so far. Finally, they will increase awareness of valid and reliable ways to evaluate vocabulary ability. A deeper appreciation of these issues will have important implications for both language instruction and assessment and well as for research on second language acquisition.

1 Lexical Competence and Lexical Proficiency

1.1 Introduction

In the first 50 years of modern linguistics, vocabulary remained at the periphery of researchers' interest and agenda. Lexis was traditionally conceptualised as a large and unstructured collection of individual items stored in memory, with no consequence for the system that language was considered to be. For example, Bloomfield (1933) famously defined the lexicon as "an appendix of the grammar, a list of basic irregularities" (p. 274). Although the main linguistic theories of the time—*structuralism* and *generativism*—recognised that words constituted the building blocks of language, they treated lexis as amorphous raw linguistic material and directed their attention to the structures and rules which held it together. An interest in the lexicon began to appear in the 1970s in the field of generative semantics (e.g. Katz & Fodor, 1963; Jackendoff, 1972) and a decade later in the area of applied linguistics, in particular of second language learning and teaching (Meara, 1980). Due to these influences vocabulary started to be perceived as a complex and organised linguistic module. It also started to play a more important role in linguistic research. Recently, it has featured prominently in the descriptions of language, which emphasise its equal status with other linguistic subsystems (e.g. *pattern grammar*, Hunston & Francis, 2000). It has also been vigorously studied in the field of psycholinguistics, which has explored mental representation, processing and acquisition of vocabulary (e.g. *connectionism*; Rumelhart, McClelland, & PDP Research Group (1986); Seidenberg & McClelland, 1989).

Recent models of language not only recognise the role of lexis in the overall linguistic system, but postulate its inseparability from other subsystems. The *systemic functional approach* advocates the complementarity of grammar and lexis. Its founding father, M. A. K. Halliday, observed that "The grammarian's dream is . . . to turn the whole of linguistic form into grammar, hoping to show that lexis can be defined as 'most delicate grammar' ". He further asserted that "grammar and vocabulary and not different strata [of language]; they are two poles of a single continuum, properly called *lexicogrammar*" (Halliday & Matthiessen, 2014, p. 24). According to Halliday, lexicogrammar forms the stratum of 'wording',

which mediates between the lower stratum of 'sounding' (graphology/phonology) and the higher stratum of 'meaning' (semantics/discourse). It consists of a closed system of meaning-general grammatical structures and open sets of meaning-specific lexis, but also of intermediary stages between these two ends of the cline such as collocations and colligations.

The traditional distinction has likewise been challenged by *cognitive linguistics* which also postulates a continuum approach to the whole range of linguistic components rather than their modularity. Such a view was put forward by Ronald Langacker in his conception of grammatical structure, which he named *cognitive grammar*:

> There is no meaningful distinction between grammar and lexicon. Lexicon, morphology and syntax form a continuum of symbolic structures, which differ along various parameters but can be divided into separate components only arbitrarily.
>
> (Langacker, 1987, p. 3)

According to Langacker, these symbolic structures are associations of phonological and semantic units, i.e. forms and established concepts. Their simplest kind are morphemes; however, basic structures can combine and create increasingly larger symbolic structures, e.g. words and grammatical patterns, all of which function as pre-packaged unitary entities. Thus, lexical and grammatical structures (i.e. words and grammatical patters) differ from each other and other symbolic structures "not in kind, but only in degree of specificity" (p. 58). An offshoot of cognitive linguistics, *construction grammar*, also recognises words as one type from the pool of linguistic units, holistically termed constructions, other types of constructions being morphemes, idioms, phrases, partially lexically filled and fully general grammatical patterns. Goldberg (2003, p. 219) maintains that all constructions are "pairings of form with semantic or discourse function". They create a large hierarchically structured network—construct-i-con—which encapsulates "the totality of our knowledge of language" (p. 219). The two approaches do not deny the existence of the lexicon as a discernible language constituent, but they assert that its distinctiveness from other types of linguistic elements is blurred and its precise delineation impossible (Langacker, 1987, p. 19). The mental representation of words is usage based and emergent (Bybee, 2006), and encompasses information on words' various forms, syntactic patterns in which they occur, their different meanings and contexts of use.

This spur of interest in the lexical aspects of language and the new approaches to the status of lexis in the linguistic system have resulted in the multitude of definitions and models of vocabulary, which have been proposed in order to provide a theoretical framework for empirical studies, ranging from corpus-based explorations through psycholinguistic experiments to neurolinguistic imaging. Each of these attempts centred on lexical properties which suited its research agenda. Even within the more

focused linguistic sub-discipline of Second Language Acquisition (SLA), there has been no agreement on the nature of vocabulary knowledge and on how it should be defined, described, measured and assessed. This lack of consensus has been aptly articulated by Read and Chapelle (2001):

> An observation that emerges from a review of this literature is the ill-defined nature of vocabulary as a construct, in the sense that different authors appear to approach this from different perspectives, making a variety of—often implicit—assumptions about the nature and scope of the lexical dimension of learners' language.
>
> (Read & Chapelle, 2001, p. 1)

This chapter provides an overview of different approaches to and models of vocabulary that have been put forward in the literature on second language acquisition and assessment.

1.2 Preliminary Definitions

The precise definition of the construct of lexical command has far-reaching consequences for its measurement and assessment as well as for the interpretation of the results of these two procedures. However, there has been no agreement even regarding the terminology used to describe the lexical aspects of language. SLA researchers refer to *the lexicon*, *vocabulary knowledge*, *lexical competence* and *lexical proficiency* either employing these labels interchangeably or applying them to different, but poorly defined concepts. In fact, if these terms are ever elaborated on, it is usually through mentioning the component parts of the constructs which they refer to, rather than delineating the constructs themselves. For example, the seminal book on the mental lexicon *Words in the Mind* (Aitchison, 2003) offers only a rudimentary definition of its key term, which is "the word-store in human mind" (p. 4). A more elaborate explication can be found in Schwarz (1995), who writes:

> The mental lexicon is a system in our long term memory (LTM), where all our knowledge about the words of our language(s) is stored.
>
> (Schwarz, 1995, p. 63)

One of the few definitions of lexical competence, which can be found in the literature has been offered by the Common European Framework of Reference. The document specifies lexical competence as the "knowledge of, and ability to use, the vocabulary of a language" (Council of Europe, 2001, p. 110). This delineation may seem simplistic at first glance, but it emphasises two important aspects of vocabulary—knowledge and ability, as well as observable behaviour through which they are manifested—vocabulary use. It implies that the lexical competence does

not only consist of a body of information but it also includes a capability of applying this information to perform a communicative act. In this way the authors of CERF relate the concept of lexical competence to larger models of language and frameworks of communicative competence, which were developed in the 1980s and 1990s.

1.2.1 Communicative Competence, Language Ability and Language Proficiency

For a long time preparing learners for the demands of communication in L2 was conceptualised as developing their knowledge of and about the second language, which was perceived as the only factor enabling learners to use language. This was a reflection of formal approaches to language dominant in linguistics for the first seven decades of the 20th century. Structuralism introduced the distinction between *langue*, an abstract and complex system of linguistic structures which existed independent of its users, and *parole*, concrete instances of the use of *langue* (de Saussure, Bally, Sechehaye, & Riedlinger, 1916). It was *langue* which constituted the focus of linguistic inquiries. A revolution in linguistics brought about Noam Chomsky (1957, 1965) and transformational-generative linguistics shifted the perspective by focusing on *linguistic competence*, i.e. "the speaker-hearer's knowledge of his language" (1965, p. 4), but also discarded *linguistic performance*—defined as "the actual use of language in concrete situations" (p. 4)—as not worthy of scientific scrutiny. Although these major schools of thought were not directly concerned with second language acquisition and use, they exerted their influence on the SLA discipline. Thus, L2 knowledge was placed at the heart of L2 research, teaching and assessment and it was believed to be the driving force of L2 performance. It was not until Dell Hymes's (1972) seminal paper that the direct link between knowledge and performance was challenged. Working in the context of sociolinguistics, Hymes introduced the concept of communicative competence which encompassed two separate components: *knowledge of language* and *ability for use*. He claimed that the actual language performance is a result of a complex interaction between underlying linguistic knowledge as well as more general cognitive and psychological mechanisms constituting ability for use.

Hymes's model of communicative competence caught the attention of researchers in second language acquisition, teaching and testing, and in the '80s and '90s its several adaptations and elaborations were proposed by Widdowson (1979), Canale and Swain (1980), Bialystok and Sherwood-Smith (1985), Taylor (1988), Davies (1989), Bachman (1990) and most recently by Bachman and Palmer (1996). Each of these paradigms recognises the basic distinction introduced by Hymes, but differs in the way of conceptualising these two components and their constituents (see McNamara, 1996 for a detailed review).

One of the most influential accounts of these components and the patterns of interaction between them were proposed by Bachman (1990) and later modified by Bachman and Palmer (1996). In the more recent version, the authors introduce the term *language ability* which consists of *language knowledge* and *strategic competence* and which interacts with topical knowledge, personal characteristics and affective states and further with contextual factors in generating language use.

Bachman and Palmer (1996) describe their model in the following words:

> Language use involves complex and multiple interactions among the various individual characteristics of language user, on the one hand, and between these characteristics and the characteristics of the language use or testing situation, on the other. Because of the complexity of these interactions, we believe that language ability must be considered within the interactional framework of language use. The view of language use we present here thus focuses on the interactions among areas of language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata, on the one hand, and how these interact with the characteristics of the language use situation, or the task, on the other.
>
> (Bachman & Palmer, 1996, p. 62)

Bachman and Palmer further list the constitutes of the two components of language ability, which are presented in Figure 1.1 and Figure 1.2.

- Organisational knowledge
  - Grammatical knowledge
    - Knowledge of vocabulary
    - Knowledge of syntax
    - Knowledge of phonology/graphology
  - Textual knowledge
    - Knowledge of cohesion
    - Knowledge of rhetorical or conversational organisation
- Pragmatic knowledge
  - Functional knowledge
    - Knowledge of ideational functions
    - Knowledge of manipulative functions
    - Knowledge of heuristic functions
    - Knowledge of imaginative functions
  - Sociolinguistic knowledge
    - Knowledge of dialects/varieties
    - Knowledge of registers
    - Knowledge of natural and idiomatic expressions
    - Knowledge of cultural references and figures of speech

Figure 1.1 Areas of Language Knowledge

Source: Bachman and Palmer (1996, p. 68)

Goal setting
(assessing what one is going to do)
Assessment
(taking stock of what is needed, what one has to work with and how well one has done)
Planning
(deciding how to use what one has)

Figure 1.2 Areas of Metacognitive Strategy Use
Source: Bachman and Palmer (1996, p. 71)

The strategic competence is placed at the heart of the whole process of language use in Bachman and Palmer's model. They define it as a set of metacognitive strategies which are executive processes regulating language use by managing the interaction of various components of the process, language knowledge being only one of them.

The interesting new development in this model of language use and language test performance consisted in including both topical knowledge and affective schemata, the latter defined as "affective and emotional correlates of topical knowledge" (p. 65) which determine the language user's affective response to the language use situation and which can influence his or her linguistic reaction to it. Emotional response can have a facilitating or debilitating effect on the user. Yet, as pointed by McNamara (1996, p. 74) the understanding of the precise functioning of the affective schemata in relation to performance is still very crude.

The language ability, as conceptualised by Bachman and Palmer has also been referred to by other researchers as language proficiency. For example, Thomas (1994, p. 330, footnote 1) defines language proficiency "a person's overall competence and ability to perform in L2". A more in-depth elaboration of the concept of language proficiency was provided by Hulstijn (2011):

> *Language proficiency* (LP) is the extent to which an individual possesses the linguistic cognition necessary to function in a given communicative situation, in a given modality (listening, speaking, reading, or writing). Linguistic cognition is the combination of the representation of linguistic information . . . and the ease with which linguistic information can be processed (skill). . . . Linguistic cognition in the phonetic-phonological, morphonological, morphosyntactic, and lexical domains forms the center of LP (*core components*). LP may comprise *peripheral components* of a less-linguistic or non-linguistic

> nature, such as strategic or metacognitive abilities related to performing listening, speaking, reading or writing tasks.
>
> (Hulstijn, 2011, p. 242; emphasis in original)

Clearly, all the notions already discussed: Hymes's (1972) communicative competence, Bachman and Palmer's (1996) language ability and Hulstijn's (2011) language proficiency relate to the same construct which—according to the researchers—consists of two main components termed by the authors as: knowledge of language, language knowledge or linguistic cognition; and, on the other hand, ability for use, strategic competence, or strategic or metacognitive abilities. Yet, while for Hymes the two components seem equally important, Bachman and Palmer place the strategic competence in the centre of their model, and Hulstijn treats strategic or metacognitive abilities as peripheral constituents of the construct. Nevertheless, the researchers agree that this latter element of the construct is not a purely linguistic faculty and includes broader cognitive and affective factors. Its exact nature is not well understood, which was best commented on by McNamara (1996):

> Language *knowledge* is relatively straightforward, and . . . somewhat of a consensus has emerged about what aspects of this knowledge (of grammatical and other formal linguistic rules, sociolinguistic rules, etc.) it is appropriate to consider. *Ability for use*, on the other hand, is more difficult to grasp, because we need to consider here a range of underlying language-relevant but not language-exclusive cognitive and affective factors (including general reasoning power, emotional states and personality factors) which are involved in performance of communicative tasks. Because these factors are not exclusive to the domain of language use they are not the preserve of language specialists, they have therefore been less often discussed in the language field and, consequently, their role in communication is less clearly understood.
>
> (McNamara, 1996, p. 59)

The paradigm presented previously—termed by different authors as communicative competence, or language ability, or language proficiency—represents what McNamara (1996, p. 59) calls *potential* for performance which is available to the language user. McNamara also observes that this potential needs to be distinguished from *actual instances* of language use (in real time), or, in other words, *actual performance*. Bachman and Palmer (1996) emphasise that language use is a result of interaction of the learner's individual characteristics—his or her strategic competence in particular—and the characteristics of a task or setting in which the learner interacts. Housen and Kuiken (2009, p. 461) also make a distinction between language proficiency and language performance. They

remark that the learner's proficiency underlies his or her performance; however, they do not attempt to examine this relation in more detail.

It needs to be noted here that the distinction between knowledge, proficiency and performance highlighted in the models discussed earlier is different from the Chomskian competence/performance dichotomy. According to Chomsky, performance is not a direct reflection of competence solely due to such unsystematic and "grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic)" (Chomsky, 1965, p. 3). The paradigms proposed by Hymes, Bachman and Palmer, and Hulstijn recognise that the linguistic competence is mediated by general cognitive and metacognitive processes before it manifests itself in actual performance.

1.2.2 Cognitive Linguistic Models of Language

The models of language ability and language use discussed previously represent a modular view of the mind and clearly separate linguistic knowledge from the more general cognitive mechanisms and the context of use. Cognitive linguists and construction grammarians challenge this compartmentalised account of language. Instead, they posit that the representation, processing and acquisition of linguistic structures are motivated by general cognitive processes and social interaction (Ellis, Römer, & O'Donnell, 2016, pp. 23–25). Thus, language knowledge, retrieval and learning is not different from the knowledge, retrieval and learning of non-linguistic phenomena such as facts or rules of social behaviour. The cognitive linguistic theories also oppose rigid dichotomies such as knowledge vs. ability, competence vs. performance, linguistic knowledge vs. topical knowledge or semantic meaning vs. pragmatic function. In their place, they propose a description of various linguistic phenomena in terms of a continuum. This view is well exemplified by the following quotation:

> A third dimension of the discreteness issue concerns the propriety of posing sharp distinctions between broad classes of linguistic phenomena, thereby implying that the classes are fundamentally different in character and in large measure separately distributable. The nondiscrete alternative regards these classes as grading into one another along various parameters. They form a continuous spectrum (or field) of possibilities, whose segregation into distinct blocks is necessarily artefactual.
>
> (Langacker, 1987, p. 18)

Thus, the semantic meaning of a particular symbolic unit/construction, for example, of the word *dog—a common four-legged small or*

medium-sized animal—cannot be separated from the encyclopaedic knowledge of its characteristic properties—"a domesticated carnivorous mammal, commonly kept as a pet, that is wonderfully faithful and has a long snout, a wet nose, an acute sense of smell, a barking voice, a wagging tail and no sense of decorum" (Ellis et al., 2016, p. 26). By the same token, the semantic meaning of the symbolic unit/construction *Nice to meet you—I am pleased to make an acquaintance with you*—cannot be separated from its function—an acknowledgment that I am meeting you for the first time in my life and from now I should consider you my acquaintance—as well as the context in its use—a first contact with a person in a social or professional context, usually accompanied by shaking hands. Words and other constructions form cognitive schemas of related concepts called semantic frames (Fillmore, 1977). Frames specify contractions' attributes, functions and typical associations and are used to interpret events and situations in life. Their emergence is based on recurring experiences.

In cognitive linguistic approaches to language, the border line between linguistic knowledge and ability for use (i.e. linguistic competence and proficiency) is also believed to be fuzzy. Bybee (1998, pp. 424–425) links the distinction between these two constructs to the distinction between declarative and procedural knowledge proposed in psychology (Anderson, 1993). Declarative knowledge refers to the memory of facts and information, which can be probed and attested directly, while the procedural knowledge encompasses cognitive routines and is only manifest in performance of a skill. Declarative knowledge can become procedural through repeated use when parts of information become joined in larger behavioural chains, automatised and applied holistically. Linguistic knowledge is usage based and emergent (Bybee, 2006). With recurring exposure and repetition, sequences of cognitive processes become one unit of procedure which links the form, meaning and their triggering context. Langacker (1987, p. 59) observes that automatisation is also a matter of degree. With repeated use, a novel structure becomes gradually entrenched in cognitive organisation to the point of becoming a single unit. This process is progressive and depends on the frequency of the structure's occurrence. Yet, on the whole cognitive linguists maintain that linguistic knowledge is primarily procedural. This view is well expressed by the following quotation:

> Linguistic knowledge is not just propositional or representational knowledge. A large portion of the stored knowledge that makes language possible is procedural knowledge. Stored chunks are procedural chunks, embedded in context not just cognitively and socially, but also embedded physically in the production and comprehension systems along whose paths they run, and also physically in the articulatory gestures and the manual gestures that are coproduced with them.
>
> (Bybee, 1998, p. 434)

Such a view of language renders the sharp distinction between the knowledge of language and language ability—or between linguistic competence and proficiency—irrelevant.

1.2.3 Aspects of Language Proficiency

The customary way of describing language proficiency in applied linguistics over the last 30 years has been through distinguishing their three dimensions: complexity, accuracy and fluency. The triad was first applied by Skehan (1989) and then used by other researchers (Foster & Skehan, 1996; Wolfe-Quintero et al., 1998; Housen & Kuiken, 2009; Bulté & Housen, 2012). The definitions of these three notions proposed by various researchers were aptly summarised by Housen and Kuiken (2009):

> Complexity has thus been commonly characterized as "[t]he extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003, p. 340), accuracy as the ability to produce error-free speech, and fluency as the ability to process the L2 with "native-like rapidity" (Lennon, 1990, p. 390) or "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (R. Ellis, 2003, p. 342).
>
> (Housen & Kuiken, 2009, p. 461)

A careful look at the definitions reveals that the three dimensions are in fact characteristics of linguistic production and contribute to the perception of its quality. A proficient L2 learner is expected to produce a complex, accurate and fluent spoken and written discourse. Therefore, it can be said that the quality of the learner's production is a reflection of his or her L2 proficiency and the three main characteristics of language production represent the dimensions of language proficiency. At the same time, Housen and Kuiken maintain that the three dimensions are frequently perceived by researchers as the main by-products of the psycholinguistic mechanism underlying L2 processing and therefore they are assumed to offer an insight into the representation of and access to L2 knowledge, with complexity and accuracy linked to L2 knowledge representation and fluency related to control over or access to this knowledge system. The authors also point out that according to a large body of research the three dimensions are distinct components of L2 proficiency which can be measured separately (p. 462). However, the precise measurement of these three dimensions raises several issues related to their operationalisation in terms of observable performance characteristics. Accuracy may seem the easiest of the three, as it can be linked directly with the occurrence of errors in L2 production, but even error identification and classification are a challenge in itself (cf. James, 1998). Fluency can be quantified by such observable phenomena as speed of delivery, number and length of pauses or a number of false starts or repetitions. However, all

these characteristics relate to speech and there is far less agreement as to the quantifiable features of fluency in writing (see Wolfe-Quintero et al., 1998 for a discussion). Finally, the measurement of complexity generates most controversies. Researchers agree that its main subcomponents include size, elaborateness, richness and diversity of language use, but there has been many debates around establishing quantifiable characteristics for each of these qualities (cf. Wolfe-Quintero et al., 1998).

Cognitive linguistic accounts of language discussed in this and previous sections are not directly concerned with second language acquisition and do not devote a lot of attention to the perceptual characteristics of language production. However, both Langacker (1987, pp. 35–36) and Bybee (1998, pp. 432–433) observe the predominantly procedural storage of language, which they postulate warrants—and at the same time explains—the conventionality and fluency of language use. Out of an almost limitless stock of possible grammatically accurate chunks of language, users tend to select the ones that are standardised ways of expression in a given context. Prefabricated chunks stored in the memory also enable users to process language rapidly in real time without much stalling and hesitation. This belief points to the fact that while accuracy and fluency are recognised as important features of language production in both modular and cognitive linguistic approaches, the latter puts additional emphasis on conventionality rather than complexity.

1.2.4 Approaches to the Description of Lexical Competence

The modular and cognitive linguistic models of language competence and language proficiency discussed in this section can be very helpful in arriving at precise definitions of the notions of lexical competence and lexical proficiency. In the following sections, various accounts of lexical competence will be presented first, before relating them to the construct of language proficiency.

Two approaches to the description of lexical competence have been dominant in the literature. The first one perceives lexical competence as constructed of the knowledge of individual words. This is a word-centred approach. The other approach is not concerned with individual words but with the lexicon as a whole and its functioning as a system. The researchers taking the former stand list components and degrees of word knowledge; the proponents of the system-centred approach concentrate on dimensions of vocabulary knowledge.

1.3 Word-Centred Approaches to the Description of Lexical Competence

The interest in the lexical aspects of second language acquisition started with explorations of what is involved in knowing a word. In addition, an

observation was made that not all words in one's lexicon can be accessed in the same way.

1.3.1 Components of Word Knowledge

According to the lay view, knowing a word involves knowing its form and its meaning, yet researchers interested in vocabulary recognise that word knowledge is much more complex and includes other components. First attempts at describing the knowledge of a word were made by Cronbach (1942) and Richards (1976). More systematic endeavours in classifying various aspects of lexical command were undertaken by Nation (1990, 2001). In his later publication Nation proposed the following framework of word knowledge (Figure 1.3).

Nation's framework is an exhaustive account of various aspects of knowing a word. Its three main categories (form, meaning and use) are universal in the descriptions of any linguistic unit, be it a morpheme or a grammatical structure (cf. Swan, 1995 for the description of grammatical structures). The knowledge of the form includes the learner's familiarity with both the word's phonetic and orthographic shapes as well as his or her awareness of its morphological constituents, i.e. its stem and affixes. The meaning component incorporates the link between the underlying meaning of a word and its form. Only if this connection exists, we can assume that the learner knows the word (and not just the form or just the candidate meaning). Moreover, the strength of this connection determines the availability of a word for use:

> The strength of the connection between the form and its meaning will determine how readily the learner can retrieve the meaning when seeing or hearing the word form, or retrieve the word form, when wishing to express the meaning.
>
> (Nation, 2001, p. 48)

Figure 1.3 Components of Word Knowledge

Source: Proposed by Nation (2001, p. 27)

In addition to knowing the underlying meaning of the word, the learner needs to be aware of its different senses and the objects or abstract concepts they refer to. Finally, the meaning component of word knowledge involves awareness of the word's semantic relations with other words, i.e. its synonyms, anonyms and co-hyponyms, as well as familiarity with other lexical units from the broader semantic network the word is part of. The knowledge of the use of a word implies awareness of its part of speech and its grammatical properties (e.g. countable/uncountable, transitive/intransitive, gradable/ungradable). It also involves the learner's acquaintance with the grammatical patterns it can occur in (colligations) as well as other lexical words with which it can form syntagmatic relations (collocations). The last element embraces the constrains on the use of the word in context, i.e. awareness of its frequency in language, and its appropriateness in particular registers and styles.

Cronbach's, Richard's and Notion's accounts have laid foundations for how the knowledge of a word is described, and there is a general consensus in more recent publications as to its components. For example, Bogaards (2000, pp. 492–493) lists six aspects of knowing a lexical unit: form, meaning, morphology, syntax, collocates and discourse. Laufer and Goldstein (2004, p. 400) claim that word knowledge is a sum of six interrelated 'subknowledges': knowledge of spoken and written form, morphological knowledge, knowledge of word meaning, grammatical knowledge, connotative and associational knowledge and the knowledge of social or other constraints to be observed in the use of a word. Even though each of these lists categorises the individual components in a slightly different way, in fact they refer to the same aspects of lexical competence.

The emphasis of the framework approach to word knowledge is on enumerating all possible subknowledges that contribute to lexical command. However, it should be noted that not all these components add to the overall control of a word to the same extent. Some aspects appear to be more central than others in the description of what it means to know a word. These aspects are the acquaintance with form of the word (either spoken or written or both), and with its core meaning (Laufer & Goldstein, 2004) and the connection between them. It is impossible to envisage the familiarity with other aspects of word knowledge such as associations or register constraints without the grasp of these two key components. The two traits are essential in communication and lay the foundations for more peripheral aspects of knowledge. This quotation by Laufer and Goldstein is a good exemplification of this observation.

> A student who knows what *advice* means, but does not know that it is used as uncountable noun, and says, *"The mother gave her daughter many advices", will be understood in spite of the grammatical error. On the other hand, a student who knows that *advice* is

> used in singular but confuses its meaning with *advance*, for example, will experience a break in communication.
>
> (Laufer & Goldstein, 2004, pp. 403–404)

It should be duly noted that this remark does justice to lay intuitions about what it means to know a word.

It should also be emphasised that even though the frameworks discussed earlier itemise different traits of word knowledge as separate subknowledges, all these components are in fact interrelated (Schmitt, 2000). For example, there is a connection between the command of spoken and written forms of a word, especially in languages with regular phoneme-to-grapheme conversion rules. The awareness of the frequency of a word is connected with the awareness of its formality—more formal words tend to be less frequent in language. The knowledge of collocational constraints and sense relations of a word is associated with the familiarity with its different meanings (e.g. *dry* as opposed to *wet* collocates with *weather*, but *dry* as opposed to *sweet* collocates with *wine*). The same is true about syntactic properties of a word (e.g. the adjective *mobile* can be used as a modifier only when it means *portable*, but as a verb complement only when it means *able to move on its own*). Thus, even though listed and described separately, all the components of lexical command form an interdependent network of knowledge with form and core meaning occupying the central role in it. The framework approach to word knowledge conceives lexical command as the sum of interrelated subknowledges.

1.3.2 Degrees of Word Knowledge

The different aspects of word knowledge discussed in the previous section accumulate into a fairly substantial bulk of information. It is very unlikely that every speaker of a language has the complete information about all the words in his or her lexicon. Moreover, a full command of all the components of lexical knowledge is not necessary for a word to be operational in one's everyday language use. For example, one may know the spoken form of a word but not its spelling, if one encounters this word—or uses it—only in conversation, or the other way around, if the word appears only in one's reading or writing. Most researchers agree that knowing a word is not an 'all-or-nothing' phenomenon, as is commonly assumed by laymen. The following quotation exemplifies this view:

> We need to rid ourselves of the *knows/doesn't know* view of vocabulary and realize that words will be known to a greater or a lesser degree, at least until they are fully mastered. It is also useful to bear in mind that many (the majority?) of words in even a native speaker lexicon are only partially known, without a complete and confident

> knowledge of their collocational and stylistic subtleties, for example. In fact, partial knowledge may well be the norm outside of a minority of very frequent/very well-known words.
>
> (Schmitt & McCarthy, 1997, p. 108)

One's incomplete knowledge of a word implies either no knowledge or a partial knowledge of its various components. Thus, one can be familiar with different aspects of a word in varying degrees. Bogaards (2000) illustrates this phenomenon in the following way:

> One can have a vague notion, e.g. that *haematin* has something to do with a blood or that a *beech* is some kind of tree. It will be clear that knowing the differences between *arrogant*, *presumptuous*, and *superior* is of another order. In other words, knowing something about the meaning of a lexical unit does not necessarily mean that one knows its meaning nor does it imply that that element has been fully integrated into the semantic network it belongs to or that one has understood all its connotations. Moreover, knowing (something of) one meaning that is associated with some form does not imply knowledge of other meanings that the same form may have, as in the case of *party*.
>
> (Bogaards, 2000, p. 492)

In the same way one may have a partial knowledge of the spoken form of a word, if one can only vaguely remember what sound it starts with, or a full command of this form, if one can pronounce it accurately with a correct stress. Therefore, it should be assumed that at any point in time the different components of lexical knowledge exist at various degrees of mastery, and that partial knowledge of words is the norm rather than an exception.

The most fundamental representation of the degrees of word knowledge is the distinction made between active and passive knowledge of a word or passive and active vocabulary. This distinction is one of the most basic and deep-seated concepts in the literature on second language vocabulary acquisition and teaching; at the same time it seems to be one of the most debated notions. Many researchers point to the lack of agreement on the nature of the passive and active knowledge and emphasise the need for a satisfactory definition of the terms (Melka, 1997; Waring, 1999).

In the discussion of the passive/active vocabulary, first some terminological issues have to be untangled. Passive knowledge of a word implies that one can retrieve its meaning when confronted with its form, and it is put into operation in activities involving reading and listening; active vocabulary, on the other hand, means that one can supply the form of a word, when one needs to express a certain meaning, and it is necessary

for speaking and writing. The skills of reading and listening used to be regarded as passive, since the language user was perceived a passive recipient of a message. In parallel, speaking and writing were considered active skills since the language user was seen as actively involved in formulation of a message. The kinds of lexical command necessary for these skills were termed, by extension, passive and active. However, new results of psycholinguistic research suggested that when involved in reading and listening, language users by no means remain passive, but they are actively involved in processing the language. That is why the passive/active terminology was challenged and replaced with more adequate terms: receptive skills and productive skills. These labels again were extended to the lexical command and thus new terms—receptive and productive vocabulary—started to be used. Yet, in the discussion of word knowledge the connotations of passive/active vs. receptive/productive are not really so relevant as they are when talking about language skills and processing. That is why the terms passive/active and receptive/productive are used interchangeably by researchers (Melka, 1997; Waring, 1999; Nation, 2001). Other terms used less frequently are comprehension/production vocabulary. In each of these cases the same phenomena are being referred to.

There are several facts about passive and active vocabulary that are taken for granted and rarely challenged. One of them is that receptive knowledge precedes active knowledge of a word. Not all words in one's lexicon are available for active use, but those which are can also be used receptively. This implies that passive vocabulary is larger than active vocabulary (Melka, 1997). However, though intuitively appealing, the receptive/productive dichotomy poses many problems in research on second language vocabulary acquisition and testing. There are several ways in which the idea of the passive/active distinction and of the degrees of word knowledge has been elaborated by researchers. The attempts can broadly be divided into four approaches: the two-lexicon approach, the continuum approach, the scale approach and the connection approach.

One way to explain the phenomenon is to understand passive and active vocabulary as two different lexicons. This view is rooted in psycholinguistics. Researchers working in this field are concerned with the representation of language in the mind and with the mental processes involved in language use. Models of the mental lexicon advanced by them have to account for both the storage and retrieval of words. They view the output/input lexicon as containing different information arranged differently for different types of processing. For the purposes of reception the lexicon needs to be organised according to its phonemic/graphemic futures. For the purposes of production it should be organised in semantic fields. The double storage may seem inefficient, but it is the processing constraints not the storage capacity that is most problematic in the human brain, so this model has its appeal (cf. Aitchison, 2003).

A different perspective on this issue is taken by researchers working within the context of applied linguistics. They point out that the passive/active terms are ambiguous and each can refer to several different knowledges and abilities. For example, the term receptive knowledge can imply being able to recognise a word form without any idea of its meaning, or being able to vaguely recall the meaning of a word form, or, being able to give a precise meaning of a word form, or being able to recognise incorrect or inappropriate use of a word in context. The same is true about the productive knowledge. Nation (2001) lists several different knowledges of a word that are subsumed by the term active knowledge (Figure 1.4).

These observations have led applied linguists, most notably Melka (1997), to reject the idea of two separate lexicons and to propose a different way to conceptualise the active/passive distinction. For them the distinction is not really a dichotomy but represents a continuum of word knowledge.

> My proposal is that the distance between R[eceptive vocabulary] and P[roductive vocabulary] should be interpreted as degrees of knowledge of degrees of familiarity. . . . These degrees are numerous, even infinite, and the passage from one degree to the next is imperceptible, because it has to do with barely perceptible degrees of knowledge of a word.
>
> (Melka, 1997, p. 99)

According to this view the lexical command can vary from a total lack of acquaintance with a word to a good familiarity with it. The terms *receptive* and *productive* represent the two ends of the continuum. This conceptualisation can be illustrated as follows (Figure 1.5).

The conceptualisation of word knowledge as varying points along the continuum is very appealing but it poses some problems. First, as Melka

- being able to say it with correct pronunciation and stress
- being able to write it with correct spelling
- being able to construct it using the right word parts in their appropriate forms
- being able to produce the word to express the meaning
- being able to produce the word in different contexts to express the range of meanings
- being able to produce synonyms and opposites
- being able to use the word correctly in an original sentence
- being able to produce words that commonly occur with it
- being able to decide to use or not use the word to suit the degree of formality of the situation

Figure 1.4 Detailed Elements of Active Word Knowledge

Source: Proposed by Nation (2001, p. 28)

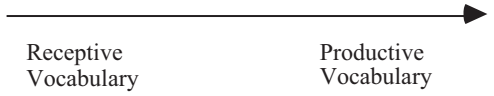

Figure 1.5 A Continuum of Receptive and Productive Word Knowledge

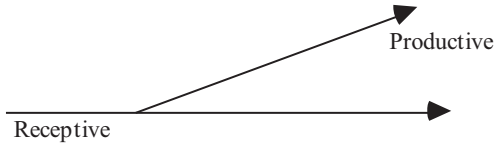

Figure 1.6 A Modified Version of the Receptive/Productive Continuum
Source: Waring (2002)

herself remarks, it is impossible to identify the point on the continuum at which the passage from receptive knowledge to productive knowledge occurs. Second, this view implies that the full receptive command of a word is a prerequisite for the onset of the active knowledge. However, Waring's (1999) results demonstrate that the ability to use a word productively, at least in limited ways, precedes full mastery of its receptive aspects. For example, one may be able to produce a word with its core meaning but not understand the metaphorical use of this word in a text. Thus, productive knowledge does not occur sequentially after receptive, there is a certain overlap between the two. Waring (2002) proposes a modified illustration of the continuum approach, which is presented in Figure 1.6.

Yet, as Waring comments, such a conceptualisation of the passive/active distinction is also unsatisfactory. First of all, it is still unclear at which point of receptive knowledge there is an onset of productive command. Next, this illustration does not allow for the interaction of the passive and active commands. For example, in this model more experience in using a word productively cannot influence its comprehension.

Another way of elaborating on this approach is an assumption that the passive/active distinction does not refer to words as complete units. Instead, it is better to assume receptive/productive command of various aspects of word knowledge. This idea is in fact built into the framework of word knowledge proposed by Nation (1990, 2001). Each of the components listed in his framework can be known receptively or productively. Nation's later model (2001) is presented in Figure 1.7 with specifications of active and passive knowledge.

However, this approach is also not free of problems. By atomising the receptive and productive command and relating it to the components of word knowledge the framework fails to account for how active and

Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognisable in this word?
		P	What parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency. . .)	R	Where, when and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?

Figure 1.7 Passive and Active Word Knowledge
Source: Nation (2001, p. 27)

passive knowledges are interrelated. It seems unrealistic to assume that a language user can have active knowledge of register constraints without having a passive knowledge of the form. The paths and directions of connections between the various passive and active aspects of lexical command are not included in the framework, thus making it an incomplete representation of receptive/productive knowledge.

The continuum perspective on the degrees of word knowledge is taken even further by Henriksen (1999), who argues that lexical competence should be described along three distinct dimensions: (1) a "partial-precise knowledge" dimension, (2) "a depth of knowledge" dimension and (3) a "receptive-productive" dimension. The first dimension refers to the degree of familiarity with the word's meaning/s, which can vary from rough categorisation to the mastery of finer shades of meaning. The second dimension describes how well the item is embedded in the network of sense relations with other words. The third dimension is related to the level of access to a word or the ability to use it. According to Henriksen, dimensions 1 and 2 are knowledge continua and dimension 3 is a control

continuum, thus she clearly differentiates the knowledge of a word from its use. She also claims that the three dimensions are interrelated.

> I hypothesize that depth of knowledge of a lexical item (as defined along dimension 2) is important for precise understanding (as defined along dimension 1). Moreover, rich meaning representation is seen as an important factor for a word to become productive (as defined along dimension 3).
>
> (Henriksen, 1999, p. 314)

The continuum approach to word knowledge is widely accepted in the literature on second language vocabulary acquisition, teaching and testing. Some researchers, however, point to the shortcomings of this perspective. Meara (1996b, 1997) argues that even though it is a useful metaphor to guide thinking about the degrees of lexical command, it contains a serious flaw in the assumption that word knowledge changes in a linear and unremitting fashion.

> Although the continuum idea is a plausible one at first sight, it turns out to be much less satisfactory when examined closely. The main problem with it is that a continuum by definition implies at least one dimension that varies continuously, and it is by no means obvious what this dimension might be in the case of words.
>
> (Meara, 1996b, p. 5)

An alternative to the continuum approach is conceptualising word knowledge as moving through a number of stages on a scale. In fact, scales of lexical command are not a new idea and a few of them pre-date the notion of a continuum (Eichholz & Barbe, 1961; after Waring, 2002; Dale, 1965). The scale of the degrees of word knowledge which has most frequently been referred to in the literature was proposed by Paribakht and Wesche (1993, 1997) and Wesche and Paribakht (1996). It consists of five stages, which are defined as self-report statements, but in the case of the more advanced ones, some evidence is also required. Paribakht and Wesche's scale is listed in Figure 1.8.

I:	I don't remember having seen this word before.
II:	I have seen this word before but I don't know what it means.
III:	I have seen this word before and I think it means _____.
IV:	I know this word. It means _____.
V:	I can use this word in a sentence. E.g.: _____.

Figure 1.8 A Scale of the Degrees of Word Knowledge

Source: Proposed by Paribakht and Wesche (1993, p. 15)

Although Paribakht and Wesche's series of levels is an improvement on the earlier scales, it has some shortcomings. It was devised to capture the first stages in the development of core knowledge of words, for example, their most common meaning. This explains why it concentrates on less advanced levels of lexical command to the detriment of the more advanced ones. Stage 1 reflects no knowledge at all about the word, which means that the effective scale includes only four levels. Stage 2 reflects the ability to recognise the word form, and stages 2 and 3 correspond to two degrees of passive knowledge. Stage 5 is most problematic. The authors' assumption is that it is supposed to capture the more advanced aspects of word knowledge, beyond its meaning, such as the familiarity with grammatical, collocational and register constrains on its use. This stage is interpreted by some researchers to tap the productive knowledge of a word (Melka, 1997; Henriksen, 1999; Waring, 2002), but whether it really performs this function is debatable. The ability to use the word correctly in a sentence when its form is provided is not equivalent to the ability to retrieve it when searching for the way to express a certain meaning. Furthermore, Waring (2002) and Laufer and Goldstein (2004) give examples of sentences which can be produced at stage 5 and which do not necessarily demonstrate a familiarity with the advanced aspects of word knowledge.

> the problem with sentence writing is that, in many cases, a sentence reveals not much more than knowledge of meaning. For example, the unique grammatical feature of *news* being a singular is not revealed in the sentence *Every morning I listen to the news*. Whether a task taker is familiar with the collocation *heavy traffic* is not clear from the sentence *I hate traffic in the morning*. All that such sentences show about the knowledge of *news* and *traffic* is that the student understands the referential meaning of these words rather than their various semantic and grammatical frames.
>
> (Laufer & Goldstein, 2004, p. 403)

In fact, the sentence does not even have to demonstrate the knowledge of the meaning of the word like in a hypothetical response suggested by Waring *I heard the word old in class today*.

Laufer and Goldstein's (2004) hierarchy of strength of word knowledge avoids some of the shortcomings of the earlier scales by concentrating on the link between two central components of lexical command—form and meaning. Their scale is built around four types of tasks which tap different degrees of word knowledge. These four tasks reflect two dichotomous distinctions. The first one depends on which component of word knowledge is given and which has to be retrieved: either a word form is provided and its meaning has to be recovered, or the other way round, a word meaning is indicated and the word form has to be regained.

This distinction is based on the assumption that there is the difference in knowledge between the ability to retrieve the word form for a given meaning and the ability to retrieve the meaning for a given word form. Laufer and Goldstein refer to former as active knowledge and to the latter as passive knowledge. The retrieval can also be of two kinds: either the word knowledge component (a form or its meaning) has to be supplied or it has to be selected from a set of options. This distinction again presupposes that there is a difference between these two forms of retrieval: the former represents recall and the latter recognition. The interaction of these two distinctions allows Laufer and Goldstein to distinguish four degrees of word knowledge which are illustrated in Figure 1.9.

Laufer and Goldstein assume that the four degrees of word knowledge form a hierarchy. The researchers subscribe to the common assumption that since reception precedes production, active knowledge of a word is more advanced than its passive knowledge. At the same time research into human memory indicates that recall needs a stronger memory trace than recognition (Baddeley, 1990); hence it is more advanced. Consequently, Laufer and Goldstein hypothesise that passive recognition is the least advanced and the active recall the most advanced of the four degrees of knowledge. The ranks of the intermediary stages cannot be established based on theoretical stipulations, so Laufer and Goldstein conducted an empirical study to determine this sequence. Their results suggest that the degrees of word knowledge form a hierarchy in the following order: passive recognition, passive recall, active recognition, active recall. They name their model the hierarchy of the strength of knowledge of meaning.

It is important to note that Laufer and Goldstein distinguish active knowledge of a word from its use, and they define the former as "the willingness of the learner to put the knowledge to use" (2004, p. 426). In an earlier paper Laufer (1998) elaborates on this particular distinction. She differentiates between controlled productive knowledge and free productive knowledge. The former is the ability to produce a word in response to a prompt, and the latter is the ability to use a word in uncontrolled production, without elicitation, at one's free will. Laufer justifies her distinction of productive knowledge by the fact that "not all learners

	Recall	Recognition
Active	supply the word form for a given meaning	select the word form for a given meaning
Passive	supply the meaning for a given word form	select the meaning for a given word form

Figure 1.9 Degrees of Word Knowledge

Source: Proposed by Laufer and Goldstein (2004)

who use infrequent vocabulary when forced to do so will also use it when left to their own selection of words" (1998, p. 257).

Laufer and Goldstein's explanation of the passive/active distinction based on the hierarchy of the strength of knowledge is an interesting contribution to the discussion on the degrees of word knowledge. It avoids the shortcomings of the previous scales by explicitly defining the kinds of tasks which tap particular degrees of lexical command. Yet, Laufer and Goldstein's hierarchy also does not fully capture the complexities of word knowledge, nor is it meant to. Their aim is to differentiate the types of lexical command that are amenable to testing and in this respect hierarchy performs its function.

Meara (1997) takes a very different perspective on the degrees of word knowledge. He proposes a simple representation of the mental lexicon as a network of unidirectional connections between words. In order to become part of one's lexical competence a newly encountered word needs to form connections with some other words that already exist in the lexicon. The number of connections a word has with other words in the lexicon determines the degree of knowledge of this word. Poorly known lexical units have few connections with the lexicon. Repetitive encounters with a word are conducive to forming a rich set of links, which taken together corresponds to better knowledge. But since connections are unidirectional, the activation of words can spread only in one direction. Some words can hold both outgoing and incoming connections. These items can spread activation to other words and also be aroused by other words. Yet, some words may only hold outgoing links. These items can only be activated by external stimuli, but not by other units, yet they themselves can spread activation to other words. Meara asserts that the difference between passive and active vocabulary is the result of the types of connections words have with the lexicon.

> the crucial difference between active and passive vocabulary might simply be that active vocabulary items are connected to their parent lexicons by more than one type of connection.
>
> (Meara, 1997, p. 119)

Meara's model accounts for several facts which were left unexplained in the previous approaches. First of all, it makes it possible to isolate the threshold between the passive and active knowledge of a word—it is the moment of creation of the incoming link from the lexicon. The model also implies that there is no linear progression from passive knowledge to active knowledge. Some words can become active as a result of a single exposure, whereas others need several contact situations before they can be used productively. Finally, the model suggests that being active is not a permanent state as it depends on the activation of other words. If a section of the lexicon which does not have

links to a word is activated, the word remains passive. Thus, according to Meara, it seems more appropriate to talk about passive and active states rather than passive and active knowledge. Interestingly, Meara's perspective on the passive/active distinction is often acknowledged but rarely adopted. Nation (2001, p. 25) criticises it by noting that language use is not driven by associations but by meaning, for example when a word is produced without being activated by its associational links, but in response to a picture prompt, which is an external stimulus. However, this observation can only indicate that the lexicon is not a self-contained module but has links with more general knowledge representation structures.

The review of various views on what it means to know a word exposes a lack of a clear distinction between lexical knowledge and lexical use, which has been perpetuated in the literature on second language vocabulary acquisition and has already been signalled by Read and Chapelle (2001, cf. Section 1.1). While the descriptions of the components of word knowledge refer to the type and structure of information about the word, the models of the degrees of word knowledge confound the amount of this information with an ability to access it and put it into use in broadly understood communication (see the discussion in Section 1.5). Few researchers explicitly acknowledge this difference (cf. Laufer, 1998; Henriksen, 1999). So far, Meara's model seems to be the most accurate explanation of various facts concerning passive and active vocabulary and if it continues to be revised and extended it can become a fully fledged model of the mental lexicon.

1.4 Lexicon-Centred Approaches to the Description of Lexical Competence

The accounts of lexical competence discussed previously focus on defining what it means to know a word, either by listing various components of word knowledge, or by specifying its degrees. This perspective assumes that lexical competence is the total sum of the language user's 'knowledges' of individual items. Yet, there are some shortcomings of this approach. First, it renders the representation of lexical competence as consisting essentially in a large number of fine details and therefore extremely dismembered. This implies that the assessment of the learner's lexical competence is practically impossible as it would have to involve assessing his or her knowledge of individual components or degrees of his or her knowledge of every word in his or her lexicon. More importantly, this perspective ignores the results of psycholinguistic research which demonstrate that the mental lexicon is something more than just a collection of words. Aitchison (2003) reviews and summarises the findings of various studies in this area to arrive at the conclusion that the representation of word knowledge in

our mind makes a highly dense and complex network which is organised in many different ways.

> Overall, the mental lexicon—a term which might need to be regarded as a metaphor—is concerned with links, not locations, with cores not peripheries, and with frameworks, not fixed details.
>
> (Aitchison, 2003, p. 248)

One of the first system-oriented descriptive approaches to lexical competence was proposed by Anderson and Freebody (1981) who distinguished between "two aspects of an individual's vocabulary knowledge" (p. 100). They called the first one "breadth" of knowledge, which was used to refer to "the number of words for which the person knows at least some of the significant aspects of meaning" (p. 101). The second dimension was named the quality or "depth" of understanding and the researchers assumed that "a person has a sufficiently deep understanding of a word if it conveys to him or her all the distinctions that would be understood by an ordinary adult under normal circumstances" (p. 101). Anderson and Freebody's two-dimensions framework dealt exclusively with the knowledge of word meanings and disregarded other components of word knowledge. It was developed not as a general model of lexical competence but with the purpose to explain the connection between vocabulary knowledge and reading comprehension. Despite these facts the breath/depth metaphor has flourished in field of SLA. In more recent years, however, the third dimension was added to the framework, which refers to access to lexical information stored in memory. For example, Daller, Milton, and Treffers-Daller (2007, pp. 7–8) propose a model of three-dimensional lexical space where each dimension represents one aspect of lexical competence, i.e. of word knowledge and ability. This model is depicted in Figure 1.10.

The three-dimensional framework has become one of the most widely known and frequently used descriptive models of lexical competence in research on second language acquisition and assessment. Its individual

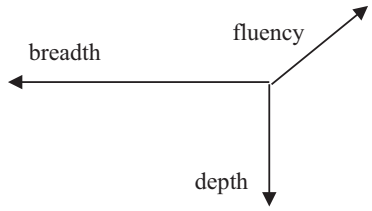

Figure 1.10 The Model of Lexical Space
Source: Daller et al. (2007, p. 8)

aspects may be referred to as breadth or size, depth or quality, and access, accessibility, automaticity or fluency. Its popularity is illustrated by the following quotation:

> In summary, while there is no clear definition for lexical competence, most researchers agree that it comprises breadth of lexical knowledge, depth of lexical knowledge, and access to core lexical items.
>
> (Crossley, Salsbury, McNamara, & Jarvis, 2011a, p. 249)

However, while the concepts of breadth as well as access have been relatively straightforward to conceptualise, the depth component has evoked a lot of controversies regarding its nature (cf. Schmitt, 2014 for a review of various approaches to this concept). Meara (1996a), in particular, observes that while the breadth dimension refers to the lexicon as a whole, the depth aspect is not a feature of the system but of individual words, as described by the word knowledge frameworks discussed in Section 1.3.1. In contrast, he argues that lexical competence can be described by a set of characteristics independent of the properties of word knowledge:

> despite the manifest complexities of the lexicon, lexical competence might be described in terms of a very small number of easily measurable dimensions. These dimensions are not properties attached to individual lexical items: rather they are properties of the lexicon considered as a whole.
>
> (Meara, 1996a, p. 37)

Meara proposes to replace the depth dimension with the concept of organisation of the lexical competence, which specifies to what extent individual items in the lexicon are interconnected with other items, and in result integrated into the system. Meara believes that the number and the type of links a word has with other words also determines the degree of its knowledge (see Section 1.3.2). Poorly known items have few connections with other words, while well-known items are well integrated into the system. Thus, a more complex organisation results in better lexical competence. In a later publication, Meara (2005a) also proposes the third dimension of lexical competence, accessibility, which indicates how fast a word can be accessed and retrieved from the lexicon. As with the organisation, the speed of this process determines the quality of lexical competence. In addition, Meara claims that the three dimensions are interconnected. The density of organisation of the lexicon is related to its size, since it is impossible to have a highly developed system with a small number of individual items. Accessibility depends to a large extent on the organisation of the lexicon—the number of links a word has with other words increases its chances of access and retrieval since it can be activated directly by more items.

The discussion in this section and the previous section (Section 1.3) clearly points to the fact that the notion of lexical competence is conceptualised differently by different researchers. Some authors perceive it as a construct which is more complex than compound knowledge of individual words and stress its design as an organised system with its own characteristics, independent of the features of its constituent elements. Some scholars focus on the purely declarative knowledge of L2 vocabulary, some complement it with procedural knowledge or accessibility and still others include the ability for use in the constructs.

1.5 Lexical Competence vs. Lexical Proficiency

More recently, a definition of lexical competence was proposed by Bulté, Housen, Pierrard and Van Daele (2008). Its importance lies not so much in an elaboration of the concept itself but in making a clear distinction between lexical competence and lexical proficiency and in linking individual components of the former to different aspects of the latter. The researchers stress that lexical competence is a cognitive construct, a representation of vocabulary in the hearer-speaker's mind, which is "not open to direct observation or measurement" (p. 279) and it operates at the theoretical level of linguistic inquiry. Lexical competence underlies a lower-order behavioural construct, operating at the observational level, and Bulté et al. propose the term *lexical proficiency* for the behavioural manifestation of lexical competence. A similar understanding of lexical proficiency—as observable behaviour reflecting the psychological reality of lexical competence—was voiced by Crossley and his colleagues:

> Recent investigations of lexical competence as a property of human factors have investigated human judgments of lexical proficiency.
>
> (Crossley et al., 2011a, p. 244)

Bulté et al. (2008) also provided a comprehensive elaboration of the notion of lexical proficiency. Their model includes the dimensions of both lexical competence and lexical proficiency as well the links between them. According to the researchers, lexical competence consists of two key elements: declarative component containing lexical knowledge, and the procedural component defined as "skill and control over knowledge" (p. 279). The authors further assert that lexical knowledge can be subdivided into three constructs: size, width and depth. These constructs are conceptualised in a manner close to Meara's (1996a) model discussed previously. Size refers to the number of lexical entries in memory, while width and depth both refer to the organisational aspects of the lexicon. These constructs represent the quality or degree of elaboration of lexical knowledge as manifested in relations between entries in the lexicon (width) or between various elements of knowledge within the same

entry (depth). Procedural competence "determines how well learners can access, retrieve and encode/decode relevant lexical information in real time" (p. 164).

The aspects of lexical proficiency proposed by Bulté et al. include lexical diversity, lexical sophistication, lexical complexity, lexical productivity and lexical fluency. They all are distinct observable phenomena that contribute to the perception of lexical proficiency. In Bulté et al.'s model lexical diversity denotes the number of different words that are used by language users in their production and a low level of word repetition. Lexical sophistication refers the "use of semantically specific and/or pragmatically more appropriate words from among a set of related words" (p. 279). The authors define this last feature in a different way than it has generally been accepted in literature (see Chapter 4): most researchers understand lexical sophistication as use of a larger number of less frequent lexical items. While lexical sophistication pertains to vocabulary as a whole, lexical complexity[1] operates at the level of individual words. It refers to the occurrence of words with their more specific, peripheral and less frequent properties in a language user's production. These unusual properties can pertain to semantic, collocational, grammatical or pragmatic components of word knowledge. Lexical productivity designates the number of words used to complete a language task. The related construct of lexical fluency refers to the speed of language production, i.e. the time used to produce a fixed number of running words. Figure 1.11 presents an excerpt of Bulté et al.'s framework depicting the correspondences between the theoretical aspects of lexical competence and the observational aspects of lexical proficiency.

As can be seen in Figure 1.11, each component of lexical competence is reflected in a different aspect of lexical proficiency. These correspondences are marked with thick black lines between the theoretical and observational levels. However, the size of vocabulary knowledge can be observed not only in lexical diversity of language users' production but also in its lexical sophistication and productivity. This implies that the use of more

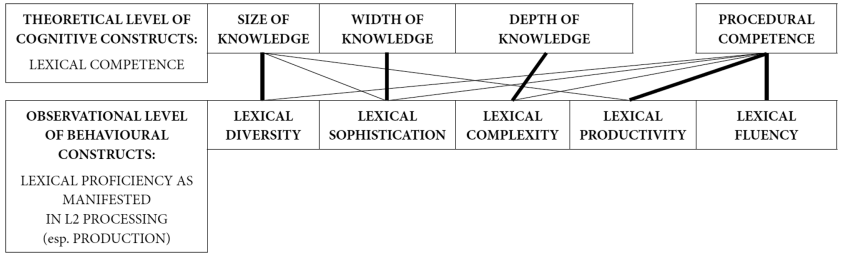

Figure 1.11 A Framework of Lexical Competence and Lexical Proficiency

Source: Bulté et al. (2008, p. 279). Reprinted with permission.

semantically and pragmatically specific words, and a larger number of words in general, requires a larger lexical store. Procedural competence, which can primarily be observed in lexical productivity and fluency, is conceptualised as controlling every aspect of vocabulary knowledge, thus its operation can be also discerned in lexical diversity, sophistication and complexity. Bulté et al. also include the third operational level in their framework. It consists of statistical constructs which serve as measures of different aspects of lexical proficiency and—by extension—of lexical competence. These measures will be discussed in detail in Chapter 4.

Lexical proficiency is sometimes also described through the three dimensions attributed to lexical competence—as discussed in the previous section—i.e. breadth, depth and access. This fact gives support to the observation made in the introduction this chapter that the two constructs are not always defined precisely and the two terms are sometimes used interchangeably. This approach is exemplified by the quotation by Crossley, Salsbury, McNamara and Jarvis:

> Generally speaking, lexical proficiency comprises breadth of knowledge features (i.e., how many words a learner knows), depth of knowledge features (i.e., how well a learner knows a word), and access to core lexical items (i.e., how quickly words can be retrieved or processed.
>
> (Crossley et al., 2011b, p. 182)

A more comprehensive perspective on lexical competence and lexical proficiency was proposed by Chapelle (1994). She based her framework on a highly influential model of communicative language ability proposed by Bachman (1990), the predecessor of the model of language ability proposed by Bachman and Palmer (1996) and discussed in Section 1.2.1. In parallel to Bachman's communicative language ability, Chapelle introduces the notion of vocabulary ability which she defines as "a capacity for language [vocabulary] use in context" (p. 163). She lists three key components of her construct definition: (1) the context of vocabulary use, (2) vocabulary knowledge and fundamental processes and (3) metacognitive strategies of vocabulary use.

The first element of the definition—context—specifies how a situation embedded within a broader cultural setting can constrain lexical choices made by language users. Chapelle refers to Halliday and Hasan's (1989) model of context analysis, which consists of three components: field, tenor and mode, and she briefly summarises these three elements.[2] Field refers to the subject matter of communication, including its topic(s), as well as its setting. Tenor denotes the social relations between communication participants. Finally, mode includes such elements as the channel or medium of communication. The characteristics of context depicted by these three elements determine to a large extent the lexical texture of a communication act.

The second element of Chapelle's definition is most related to vocabulary per se. Chapelle expands it further into: vocabulary size, the knowledge of word characteristics, lexicon organisation and fundamental vocabulary processes. Vocabulary size refers to a number of content words the language user knows. Chapelle stresses that this number cannot be estimated in an absolute sense, but always with reference to particular contexts of vocabulary use. The knowledge of word characteristics covers the components of word knowledge discussed in Section 1.3.1 and includes phonetic, graphemic, morphemic, syntactic, semantic, pragmatic and collocational features of individual words. According to Chapelle, these characteristics are also related to context of word use. The third dimension of vocabulary knowledge, lexical organisation, refers to how individual words are interconnected in the mental lexicon by a variety of links including phonological or semantic associations. This feature of vocabulary knowledge is also dependent on contextual factors which prompt the connections that language users make between individual words. Fundamental vocabulary processes are the fourth dimension of this element of Chapelle's framework, but this component is qualitatively different from the previous three. While the former three dimensions refer to declarative knowledge, this feature is purely procedural as it is associated with lexical access. Chapelle lists such fundamental processes as: concentrating on relevant lexical features in input, encoding phonological and orthographic information into short-term memory, parsing words into morphemes, obtaining structural and semantic properties from the lexicon or integrating the meanings of the words with the emergent semantic representation of the input text, all in operation in receptive language use. These individual processes are closely tied to the three aspects of vocabulary knowledge. For example, accessing relevant meanings in the lexicon during language production depends on the number and quality of connections between words, i.e. lexical organisation. Chapelle notes that the fundamental lexical processes, similarly to the former three dimensions, are also related to context of vocabulary use.

Finally, the third element of the definition of vocabulary ability refers to a set of metacognitive strategies, which—in parallel to Bachman's model—are responsible for assessing the communicative situation, setting goals for communication and then planning and executing language use. Chapelle notes that these strategies are not specific to vocabulary, yet they are involved in controlling vocabulary knowledge and fundamental processes as well as in compensating for gaps in the lexicon or for problems with accessing it. In addition, they are involved in matching the constrains of the context with the choice of appropriate lexical resources.

Chapelle remarks that her framework belongs to the group of interactionalist definitions because "it attributes performance to learners' characteristics (including knowledge, processes and strategies) to contextual factors, as well as interactions between them Messick, 1989, p. 15)"

(p. 163). It is important to note that the context of vocabulary use is external to language users' competences, yet it is included in Chapelle's framework because it determines the use of vocabulary. According to Chapelle "vocabulary will differ qualitatively depending on the context in which it used, and therefore must be specified [described/assessed] with reference to that context" (p. 164).

Chapelle does not use the term *lexical competence* in her definition of vocabulary ability, but it can be stipulated that the concept is represented by the second element of her framework—vocabulary knowledge and fundamental processes. It can be noted that her conceptualisation of this construct is not very different from the lexicon-oriented models discussed previously. All these approaches combine declarative and procedural knowledge by pointing that lexical competence consists not only of the knowledge of and about words, but also of a capacity to access this information. Chapelle, however, reunites the dimension of organisation proposed by Meara (2005a) with the depth aspect proposed by such researchers as Deller, Milton and Treffers-Daller (2007) or Crossley et al. (2010a). The strongest point of Chapelle's definition, however, lies in linking lexical competence with lexical performance. Since lexical competence is an internal characteristic of language learner, not available for direct observation, it can only be described and assessed through an analysis of language use. Therefore, it is essential to take into consideration that language use is also determined by factors other than lexical competence. Consequently, the description and assessment of lexical competence have to account for the interaction of these elements in the final observable product.

In the same way as Bachman and Palmer's (1996) model of language ability, discussed in Section 1.2.1, was linked with the notion of language proficiency proposed by Hultijn (2011), Chapelle's (1994) framework of vocabulary ability can be related to Bulté et al.'s (2008) paradigm of lexical proficiency. The strength of the latter lies in the enumeration of the components of lexical proficiency (which are similar, but not identical with the CAF components of language proficiency presented in Section 1.2.3); the advantage of the former is that it relates lexical proficiency with performance and emphasises the importance of metacognitive strategies and context in vocabulary use.

1.6 Lexical Competence, Lexical Proficiency and Phraseology

The models of lexical competence—both word-centred and system-oriented—discussed in the previous sections were founded on the knowledge of and access to individual words which make up the lexicon. Yet, in the last decades new linguistic and psycholinguistic approaches (cf. Section 1.1) have also postulated the role of phraseology in language

representation, processing and acquisition. Parallel to lexis, phraseology has moved to the central position in most recent models of language (see Gries, 1998 for an overview).

Phraseology is a fuzzy linguistic phenomenon. It is difficult to establish a precise definition of its basic unit, and to mark its boundaries. Granger and Paquot (1998, p. 28) distinguish two main approaches to this task. One is the traditional approach, stemming from the classic Russian research carried out in the middle of the 20th century. It defines phraseological units as word combinations bound by linguistic convention and characterised by special semantic, syntactic and pragmatic properties. Different amalgamations of these characteristics result in different types of phrasemes ranging from idioms, which have a frozen form and an opaque meaning, to collocations, which are formally more flexible, syntactically fully productive and semantically compositional. The other, more recent approach to defining phraseological units is rooted in corpus-based explorations of language. It does not start with predefined linguistic categories, but identifies lexical combinations based on their distribution and frequency in language. It is not concerned with neat classifications of different phrases, and it stimulated the recognition of new types of word combinations which are extremely frequent in language but had not been accounted for before in the description of phraseology. It also disclosed an interesting fact: some phraseological units, which had stayed in the core of the traditional approach, such as idioms, are in fact very rare in comparison with other more widespread types of transparent collocations. Two categories of lexical collocation which are distinguished by the distributional approach are n-grams and collocations. N-grams, also called lexical bundles, are continuous sequences of two, three or more words that show a statistical tendency to co-occur, regardless of their idiomaticity and their structural status. Biber, Johansson, Leech, Conrad and Finegan (1999) assert that lexical bundles are not fixed expressions per se because they do not form single semantic units, and they are often not recognised as such by native speakers; yet, they constitute the basic building blocks of accurate, natural and idiomatic language (pp. 989–991). In most cases, n-grams cut across phrasal and clausal boundaries. They can be composed of the beginning of a main clause followed by the beginning of an embedded clause (e.g., *I don't know why*), or of a noun phrase followed by the preposition that typically introduces its complement (e.g., *a reason for*). Collocations are combinations of two lexical words which co-occur in a statistically significant way in a predefined distance of each other (usually up to five words).

It is generally believed that phraseology is one of the hardest aspects of foreign language learning and poses problems even to advanced L2 users (e.g. Pawley & Syder, 1983; Wray, 2002). As Biber et al. (1999) assert "producing natural, idiomatic English is not just a matter of constructing well-formed sentences, but of using well-tried lexical expressions in

appropriate places" (p. 990). This is exactly what second language users find most challenging: choosing from the pool of possible word combinations those that are idiomatic, or in other words, native-like. In addition, possessing a large collections of such phrases at one's disposal facilities the speed of language processing, i.e. fluency (Pawley & Syder, 1983).

There is no agreement as to how phraseology is represented in the mental lexicon. Many researchers propose that phraseological units are stored in the mental lexicon as single items. This view is illustrated by the following quotation:

> Of all the varied structures and combinations that may be considered 'idiomatic' as opposed to just grammatically acceptable (pure idioms, collocations, standard speech formulas, and much that constitutes metaphor), a great many at best be simply listed in the lexicon—or perhaps more appropriately, [in] the light of what we want to suggest here, the 'phrasicon'.
>
> (Magee & Rundell, 1996, p. 18)

This approach to the representation of lexical combinations in the mental lexicon is gaining its popularity in psycholinguistics and cognitive linguistics (e.g. Wray, 2002; Gries, 1998). For example, according to CEFR, lexical competence consists of two kinds of items: lexical elements which are single word units and fixed expressions as well as grammatical elements, that is items belonging to closed word classes such as quantifiers and conjunctions. In this way the document acknowledges that multiword units are part of the L2 lexicon. However, this approach also has its limitations. It handles fixed and semi-fixed expressions well. It seems reasonable to assume that fixed expressions with a non-compositional meaning, a strong pragmatic function or a non-canonical structure are stored as single units in human memory. However, the status of collocations, which are not only semantically transparent but also very flexible, is debatable. They can be more convincingly conceptualised as strong activation links between individual words in the lexicon (e.g. Hoey's [2005] idea of lexical priming).

In any case, phraseology should be considered as a vital component of lexical competence. Phraseological units are either represented in it on equal footing with individual words, or as collocational links between items. In fact word-centred and lexicon-centred models of lexical competence discussed in the previous sections accounted for the storage of word combinations. The collocation links of a word with other lexical items were included in the components of word knowledge, as well as in the depth, organisation or width dimensions of different frameworks. The use of phrasemes in L2 learners' production can also be considered an important element of lexical proficiency. In fact, Levitzky-Aviad and Laufer (2013) include the use of collocations as one of the elements of

vocabulary use, along with vocabulary variation and richness. It can be thus postulated that formulaicity constitutes yet another dimension of lexical proficiency. This claim is also reflected in the taxonomy of L2 complexity proposed by (Bulté & Housen, 2012, p. 23), where lexical complexity consists of two elements: lexemic complexity and collocational complexity.

1.7 Conclusion

Vocabulary is a vital component of language representation, processing and use. The word-based and system-oriented models discussed in this chapter provide together a fairly exhaustive account of lexical competence. They present it as an abstract and complex cognitive construct, which surpasses the familiarity with individual words and their various formal, grammatical, semantic, pragmatic and collocational characteristics, and which involves functioning of the lexicon as a system with its internal structure and origination. Lexical competence consists of both declarative knowledge of assorted lexical information and procedural knowledge related to the ways of accessing this information. It is part of a larger construct of lexical proficiency, which is an ability to apply both declarative and procedural lexical knowledge in real language use. The use of vocabulary (i.e. lexical performance) is the observable manifestation of lexical proficiency and is determined not only by lexical competence but also by language users' metacognitive strategies. Lexical performance can be characterised by several discernible characteristics such as productivity, diversity, sophistication, elaborateness or fluency, which are construed as components of lexical proficiency. Assessing lexical proficiency involves rating or measuring these components. The following chapters will discuss different methods which have been proposed in order to make such an assessment.

Notes

1. It should be noted that the term *complexity* used by Bulté et al. is not equivalent to the same term in the CAF model of language proficiency discussed in Section 1.2.3. To avoid confusion it will be referred to as *elaborateness* in further discussions.
2. The elements of the model are extensively discussed by Halliday and Hasan (1989).

2 Lexical Assessment Methods

2.1 Introduction

Chapter 1 reviewed issues related to lexical command in order to delineate the complex constructs of lexical competence and lexical proficiency and the relationship between them. As lexical competence is defined as an abstract notion rather than a tangible phenomenon, it cannot be observed, measured and assessed directly. However, it forms the foundation of lexical proficiency, which refers to the ability to apply vocabulary knowledge (both declarative and procedural), and manifests itself in observable lexical behaviour. Thus, the act of assessment requires gathering and evaluating concrete evidence of lexical performance, through which lexical proficiency can be scrutinised. Based on results of such evaluation, inferences may be drawn regarding the underlying lexical competence. This chapter presents the fundamentals of assessment practices and introduces different approaches to the evaluation of lexical proficiency and—by extension—lexical competence. It also describes briefly the most popular discrete-point vocabulary tests used in education and SLA research.

2.2 Definition and Qualities of a Language Test

The concrete evidence of lexical proficiency usually takes the form of a sample or several samples of learners' performance. It can be the learner's naturally occurring linguistic behaviour, but, for the reasons discussed later, this behaviour is usually elicited with the help of one or several instruments, called tests. In everyday use the word ***test*** refers to a series of short questions and problems which a learner has to reply to in a specified amount of time; in the area of language assessment the term refers to any kind of task or a series of tasks—including for example a summary or translation of a text—eliciting the learners' performance. J. B. Carroll, an American psychologist, well-known for his contributions to

psychology, educational linguistics and psychometrics, defined a test in the following way:

> a psychological or educational test is a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual.
>
> (Carroll, 1968, p. 46)

Cronbach defines a test in a very similar manner as

> a systematic procedure for observing a person's behavior and describing it with an aid of a numerical scale or category system.
>
> (Cronbach, 1971, p. 444)

According to Bachman (1990) what distinguishes a test from other types of measurement (for example those based on L2 learners' naturally occurring language performance) is the fact that "it is designed to obtain a specific sample of behaviour", "according to explicit procedures" (pp. 20–21). Bachman further elaborates on the special role a test plays in the assessment process in comparison with other types of gathering evidence about the learner's abilities and knowledge:

> I believe this distinction is an important one, since it reflects the primary justification for the use of language tests and has implications for how we design, develop and use them. If we could count on being able to measure a given aspect of language ability on the basis of *any* sample of language use, however obtained, there would be no need to design language tests. However, it is precisely because any given sample of language will not necessarily enable the test user to make inferences about a given ability that we need language tests. That is the inferences and uses we make of language test scores depend on the sample of language use obtained. Language tests can thus provide the means for more carefully focusing on the specific language abilities that are of interest.
>
> (Bachman, 1990, p. 21)

Thus, a test elicits the L2 learner's performance, which is then observed, measured and evaluated. The result can be a numerical score, a set of scores related to different test components, a mark or a label (e.g. 'very proficient' or 'poor'). This test result is interpreted in order to extrapolate the L2 learner's general or specific language capacities. It is also used to make decisions about the learner, a programme or a teaching method. The whole process of assessment of L2 learners' language abilities can

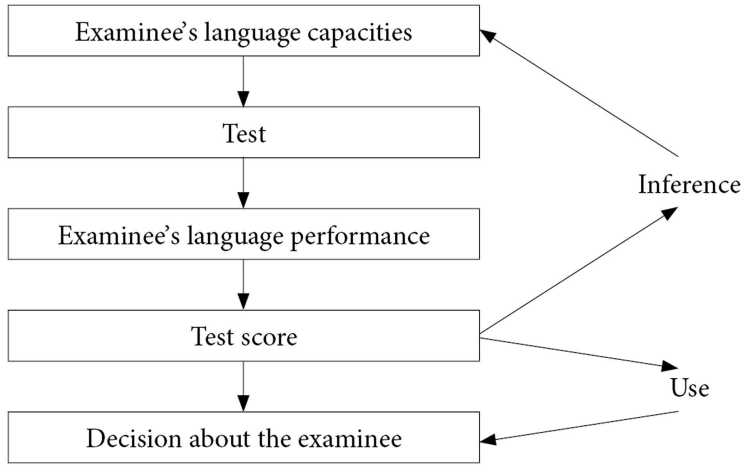

Figure 2.1 The Process of Language Assessment
Source: Adapted from Chapelle and Brindley (2002, p. 268)

be summarised by the diagram in Figure 2.1, adapted from Chapelle and Brindley (2002).

Therefore, a test is crucial to the assessment process because it makes it possible to observe, measure and evaluate what is otherwise an implicit property of the L2 learner. In addition, the choice of a test has an enormous influence of the assessment outcome. In other words, our evaluation of lexical proficiency and a further extrapolation about a learner's lexical competence depend greatly on the test and its characteristics.

In order to be a dependable instrument for assessment purposes a test has to display several characteristics. Bachman and Palmer (1996) list six characteristics of a useful test: reliability, construct validity, authenticity, interactiveness, impact and practicality. The researchers maintain that "These six test qualities all contribute to test usefulness, so they cannot be evaluated independently of each other" (Bachman & Palmer, 1996, p. 38). What is more, they can have a detrimental effect on each other. For example, a truly authentic test may require procedures which are not very practical, or a truly valid test may necessitate tasks which cannot be reliably assessed. Yet, the authors further claim that

> the relative importance of these different qualities will vary from one testing situation to another, so that test usefulness can only be evaluated for specific testing situations. Similarly, the appropriate balance of these qualities cannot be prescribed in the abstract, but can only be determined for a given test. The most important consideration to

> keep in mind is not to ignore any one quality at the expense of others. Rather we need to strive to achieve an appropriate balance.
>
> (Bachman & Palmer, 1996, p. 38)

However, according to Alderson, Clapham and Wall (1995) validity and reliability are "the overarching principles that should govern test design" (p. 6) and these two qualities cannot be compromised in any circumstances. Different tests, which will be presented and discussed in this book, will be evaluated according to these six characteristics reviewed earlier and their usefulness for assessing L2 learners' lexical proficiency and lexical competence will be considered.

2.3 Tasks Assessing Lexical Proficiency

Vocabulary can be assessed for a variety of purposes and these purposes may relate to different aspects of lexical proficiency. For example, a test may be designed to check if students can remember the form and meaning of words introduced in a lesson, a textbook unit or a course. Another possible purpose of a test can be a verification if students have sufficient vocabulary at their disposal to read academic texts in a particular area. Different aspects of lexical proficiency targeted by a test require different kinds of performance to be elicited, measured and evaluated and the required kind of performance, in turn, conditions the choice of a test format, since it is the format of the test (i.e. the task[s] it contains) that determines the kind of performance elicited from the L2 learner.

There exist many different lexical assessment procedures. For example, some tests address more directly the learner's familiarity with lexical items as independent stand-alone units while others as elements embedded in a larger context. Read (2000) proposed a framework of three dimensions which describe the various types of vocabulary tests. This framework is presented in Figure 2.2.

The three dimensions are not dichotomous, but they form three independent continua, each reflecting a degree of a different feature of a test format. For example, vocabulary can be tested in the form of word lists, with learners expected to provide a translation, a synonym or a derivation of an individual item. Such tasks can be placed at the context-independent end of the context-independent/dependent continuum. At the other end of this continuum there is a gapped text which requires the learner to understand its overall message in order to fill out the gaps with appropriate words. Yet, there are tasks—located somewhere in the middle of the scale—which provide the learner with a context, usually in the form of a phrase of a sentence, manipulated in such a way that is does not give a clue about the meaning of a word, but provides only a hint about its part of speech and a collocation.

Discrete A measure of vocabulary knowledge or use as an independent construct	**Embedded** A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective A measure in which specific vocabulary items are the focus of the assessment	**Comprehensive** A measure which takes account of the whole vocabulary content of the input material (reading/listening task) or the test-taker's response writing/speaking tasks
Context-Independent A vocabulary measure in which the test-taker can produce the expected response without referring to any context	**Context-Dependent** A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Figure 2.2 Dimensions of Vocabulary Assessment
Source: Read (2000, p. 9)

Although the three dimensions are independent of one another, the features of an individual test type tend to cluster together at one or the other end of the three continua. At one extreme end one can place discrete, selective and context-independent tests such as traditional word list translation tests and at the other end there are embedded, comprehensive, context-dependent tests, when the learner's lexical proficiency is assessed as a component of a larger linguistic skill such as speaking or writing.

Read's framework summarises several important features discriminating between different types of vocabulary tests, yet it misses a dimension which seems to be crucial in distinguishing different approaches to the assessment of vocabulary, and which does a full justice to the current understanding of the various concepts discussed in Chapter 1. In particular, it does not take into account the models of general language ability and lexical proficiency, in particular the framework proposed by Chapelle (1994). She stresses that lexical performance is a reflection of the interplay between vocabulary knowledge and fundamental processes, context, as well as metacognitive strategies. In fact, some types of vocabulary tests strictly control the context of vocabulary use and therefore make the operation of metacognitive strategies more predictable. Due to such constrains, these assessment instruments give a better insight into L2 learner's lexical competence underlying lexical proficiency and performance. Such tests can target specifically the information about the form, meaning and use restrictions of individual lexical items. The assessment procedures require the learner to recognise or recall this information in isolation from the demands of real language use in communication.

At the same time, the results of these procedures are not indicative of whether the learner can apply the same lexical items without elicitation. Other vocabulary tests are meant to address the learner's more general vocabulary ability. Such instruments aim to assess how the learner can apply the vocabulary he or she has learned in real language use situations, that is for communicative purposes. These procedures engage the learner in language comprehension or production tasks which are identical or similar to the kinds of task he or she may encounter in non-test conditions: communication in and outside the classroom. The main difference between the tests addressing lexical knowledge and vocabulary ability lies in the test characteristics of authenticity and interactiveness mentioned in the previous section. Tests targeting lexical ability demonstrate higher levels of these two properties than assessment procedures aimed at evaluating lexical knowledge. Since, according to Bachman and Palmer (1996), these two test characteristics are relative rather than absolute, this dimension—like the other three proposed by Read—is more of a continuum than a dichotomy, in particular because a total elimination of strategic competence and the learner's personal characteristics from language use situations—even the strictly controlled ones—is not possible. Thus, the observation and evaluation of pure lexical competence is not perfectly viable.

One more important aspect of vocabulary tests is not fully captured by Read's framework. A vocabulary test can set out to observe, measure and evaluate the learner's knowledge of selected individual words, or his or her vocabulary as a whole. On the one hand, an assessment procedure can focus on words which, for example, have recently been studied in class or the learner should know in order to be able to carry out a certain task. On the other hand, one may be interested in estimating the learners' L2 vocabulary as a whole. Obviously, such an estimate is performed through testing the knowledge of individual words, but these lexical items as such are not the focus of assessment, as they are chosen to represent a certain group of words making up the whole lexicon, for example at a certain frequency level or of a certain topic/function. In such a case the test is not intended to demonstrate if the learner knows a particular lexical unit, but rather a group of words which this item is representative of. This dimension is different form Read's selective/comprehensive continuum. For Read the term selective implies that the knowledge of specific items is targeted in a test and a test attempts to elicit these particular items from the L2 learner. A comprehensive test has less control over the choice of words elicited from the learner, since they are determined by a listening/reading text or the learner's spoken or written output, and the learner's command of all words used in the task is assessed.

In conclusion, a vocabulary test can check if the learner is familiar with particular words and is aware of different kinds of information about them, such as their form, meaning and use restrictions. Another test can

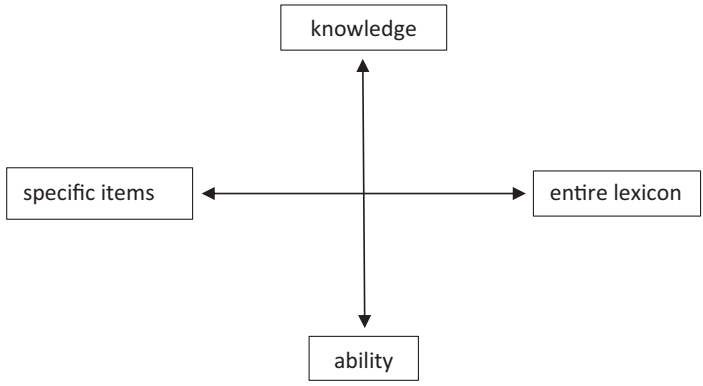

Figure 2.3 An Alternative Framework of Vocabulary Assessment Dimensions

examine if the learner can use these words in communication. Alternatively, a test can appraise the size of the learner's L2 lexicon by focusing on the learner's knowledge of L2 words' meanings. Still another test may assess the learner's skill in applying L2 vocabulary in his or her authentic written or spoken production. Thus, an alternative framework of vocabulary assessment can be proposed, which captures additional dimensions of vocabulary tests not covered in Read's model. This framework is presented in Figure 2.3.

2.4 Task Formats

There are a number of instruments eliciting the L2 learner's performance, which make it possible to assess the learner's lexical proficiency. Different assessment procedures display different properties along the dimensions discussed previously and target different aspects of vocabulary knowledge and ability. The various tasks employed for the observation, measurement and evaluation of the L2 learner's vocabulary proficiency will be presented and discussed in this section.

2.4.1 Discrete-Point Tasks

The formal assessment of the L2 learner's linguistic proficiency, including vocabulary, is probably as long as the history of language teaching. However, language testing—understood both as large-scale educational initiatives and as a separate area of inquiry within applied linguistics—can be traced back to the 1960s (Spolsky, 1995; Barnwell, 1996). Language tests developed at this time were heavily influenced by a new trend

in psychology called psychometrics, which was related to the **indirect** (i.e. inferred) measurement and assessment of various psychological constructs such as knowledge, skills, abilities, personality traits and attitudes. The important tenant of this trend was that assessment procedures should be **objective**, that is independent of the examiner's own beliefs and interpretations. Thus, the testing methods used by test developers elicited the kind of L2 performance which could be scored by a rater without a recourse to his or her personal judgement. By strictly controlling the possible output, indirect tests were better suited for measuring and assessing lexical competence underlying lexical proficiency and performance.

Another important trend in language testing, which also appeared during the 1960s, and which exerted an enormous influence on the format of vocabulary tests was **discrete-point** testing. It was based on the structuralist view of language, maintaining that language consists of different structural elements—phonemes, morphemes, lexical items and syntactic structures, and that each type of element makes up a distinct linguistic system and can be tested adequately in isolation from the other systems. Thus, the assessment of an L2 learner's linguistic mastery should address L2 pronunciation, vocabulary and grammar separately and it should be based on the learner's responses to a number of discrete-point questions, each addressing one specific detail of language such as the meaning of a particular word or the form of a particular grammatical structure. Obviously, not all linguistic points can be included in such an assessment procedure but a test should cover a sufficient representation of such items to make a judgement about the L2 learner's overall linguistic proficiency. Vocabulary is particularly suited for use in discrete-point tests because it forms a set of easily identifiable units, familiarity with whom can be checked by means of questions with one correct answer.

A number of assessment instruments were and continue to be used in objective and discrete-point tests. The recognition aspects of word knowledge can be tested through yes/no, matching and multiple-choice tasks. The **yes/no test** is the simplest assessment procedure which requires the learner to decide if he or she is familiar with the presented items. It targets the most shallow information about L2 words: recognition of the form. The learner does not have to provide any evidence justifying his or her decision. To counterbalance the effects of blind guessing, a number of non-words are included in the list. A special algorithm is used to calculate the final score. It takes into account the real words checked by the learner and the so-called 'false-alarms', that is the ticked non-words, for which the learner is penalised (for more on the algorithms and yes/no vocabulary items see Eyckmans, Velde, Hout, & Boers, 2007). Another version of the yes/no test is the so-called **lexical decision task** in which the learner has to decide if the stimulus word is a real L2 item or a non-existing letter string. The latter always respects phonotactic and spelling rules of the target language. Lexical decision tasks are always

administered through a computer programme and the learner is presented with one word at a time.

Another assessment procedure suitable for testing the recognition aspects of word information are matching and multiple choice. In a **matching** task L2 words can be linked with pictures, L1 equivalents, synonyms, definitions or collocations. A more elaborate format of a matching task presents the learner with a gapped text and a list of words that have to be paired with the blanks. **Multiple-choice** tasks are similar to matching. The learner is presented with prompts in the form of individual words, phrases or (gapped) sentences and for each prompt he or she is required to select a correct answer from four options. The test can also be based on a gapped text, in which case options are provided for each blank separately. The matching and multiple-choice tasks are suitable for evaluating the learner's recognition of almost all aspects of word information: the link between the word's form and meaning, its grammatical behaviour, as well as syntagmatic and paradigmatic relations between individual items in the lexicon and use restrictions. Multiple-choice items are generally preferred by test designers over the matching procedures, as the items in a multiple-choice test are independent.

The next group of objective and discrete-point instruments targets the recall aspect of lexical knowledge. These tasks require the learner to produce the target words or phrases rather than recognise them among available options. In **short-answer** tasks the learner is expected to write L2 words or phrases that correspond to pictures, L1 items, definitions or synonyms. The prompts can also be presented as gapped sentences or a text. The blanks can contain hints in the form of the first letters of the target words or their base forms. The short-answer or gap-filling tasks can address many aspects of word information: form, the link between form and meaning, grammatical behaviour, collocational knowledge and other use restrictions. Other tasks targeting the recall aspects of word knowledge require the learner to **transform** a prompt sentence using a specified word. The learner is expected to write a new sentence or its part. Such items focus on the knowledge of the grammatical behaviour of the target items and their collocational restrictions. Finally, the learner can be requested to **write individual sentences or a text** which contain specified words. Such tasks address several aspects of word knowledge: form, meaning and use restrictions. Yet, since such tests exert little control over the content of the learner's output, the learner's production may not demonstrate well the quality of his or her word knowledge.

2.4.2 Integrative Tasks

The reaction to structuralist, discrete-point trend in language testing was the integrative approach, inspired by the emerging field of pragmatics, which gained its importance in linguistics in the 1970s. Pragmatics

rejected the view of language as a closed system of structural elements and postulated regarding it in context, both linguistic and extra-linguistic. The integrative approach to language assessment is based on the belief that language proficiency is a unified set of interacting abilities that cannot be separated apart and tested adequately in isolation. Integrative tasks are pragmatic in that they cause the learner to process the elements of language in context, and also require him or her to relate language to the extra-linguistic reality, such as thoughts or feelings. The integrative testing advocated the assessment procedures which were set in context and treated language holistically. Four popular techniques in this approach were **cloze tests**, **C-tests**, **dictations** and **partial dictations**. These test items were studied and strongly recommended by researchers in language testing (for a review of research on cloze and C-tests, see Read, 2000, pp. 101–115; for a history of dictation as a testing technique see Stansfield, 1985); yet they gained little popularity, especially as tests of vocabulary knowledge. They give the test developer no or little influence and control over which language elements are being targeted by each item, and thus few of the items can address specifically different aspects of lexical proficiency. This can hardly be taken as criticism, as inseparability of language elements and skills in testing is in fact the main tenant of the integrative approach to language assessment. Thus, the four methods are regarded as tests of overall language proficiency and even though they also assess lexis, their applicability for testing specifically L2 learners' lexical proficiency is limited.

2.4.3 Communicative Tasks

The most recent trend in language testing—communicative language testing—grew in parallel with the communicative approach to language teaching (Brumfit & Johnson, 1979). It maintains that assessment procedures should compel the learner to *use* language *naturally* for genuine communication, or, in other words, to push the learner to put *authentic* language to use within a *context*. This approach to language testing reflects the current understanding of language (Bachman & Palmer, 1996; cf. Chapter 1) by targeting ability rather than knowledge. In communicative language testing, tasks bear resemblance to language use situations in real life and elicit only the kind of L2 performance which the learner is likely to deliver outside the testing conditions, in authentic life circumstances, requiring him or her to engage in interaction in L2. Such tasks involve both comprehension of spoken and written language (listening and reading comprehension) and well as production in these two modes (speaking and writing). The four skills may also be integrated (as in the case of conversation) or mediated (for example when the learner is expected to write a summary of a text). In addition to requiring linguistic and communicative abilities, such tasks may also address extra-linguistic

skills, such as interpreting graphs and figures, or general knowledge. Similarly to the integrative approach, communicative language testing does not support the separation of individual language elements. It also promotes pragmatic tests as it stresses the importance of both linguistic context and external reality—including opinions and feelings—in the evaluation of learner's L2 comprehension and production. What makes the communicative assessment procedures different from cloze tests and dictation is the requirement of the authenticity of the task. Filling out gaps in a text and writing verbatim a text being read aloud are indirect tasks as they do not resemble the real use of L2 in comparison with direct tasks such as listening to a train station announcement or (an excerpt of) an academic lecture, reading a bus schedule or a story, describing a picture orally or writing an argumentative essay.

Vocabulary is a recognised component of general communicative competence (see Chapter 1 for a discussion) and even though communicative assessment procedures target general language proficiency as an integrated construct, lexical proficiency is included in evaluation. The assessment of listening and reading comprehension is usually based on questions pertaining to the information in the text or other spoken or written linguistic input. Alternatively, the learner's understanding of such information can also be evaluated based on his or her performance in a productive task incorporating the information from reading or listening, such as planning a trip after reading a tourist leaflet or writing a summary after listening to a story. Such questions or tasks, however, rarely target the learner's familiarity with individual words or phrases, but they focus on a global understanding of a message or messages. Extensive research in first and second language reading has demonstrated that the learner's lexical proficiency is the single best predictor of reading comprehension both in L1 (Anderson & Freebody, 1981) and L2 (Laufer, 1992; see Grabe, 2008, chapter 13 for a recent review). The same trend, but less documented, is present for listening comprehension (Stæhr, 2008, 2009; van Zeeland & Schmitt, 2013). Nevertheless, despite a strong correlation between vocabulary knowledge and comprehension, reading and listening tasks have never been used as tests of the learner's lexical proficiency. On the other hand, the lexical content of a text has been used to predict and measure its readability or listenability (for more information on readability formulas see Chapter 4).

Productive communicative tasks usually elicit extended spoken and written output from the learner. In addition to checking the productive components of language ability, they also test comprehension and general strategic competence. The learner has to choose an appropriate reaction (strategic competence) to a prompt or other linguistic stimuli (language comprehension)—an interlocutor's comments or questions in an interactive speaking task or texts read or heard in an integrated writing task. Due to these underlying characteristics, the two kinds of performance,

spoken and written extended production in tasks simulating real language use, have become important and valuable data for the assessment of learners' general language ability as well as its components including lexical proficiency.

The assessment of L2 written and spoken production has traditionally been performed by human raters. Yet, under the influence of objective approaches to testing, attempts have been made to reduce the subjective element of evaluation, inherently present in human rating, to a minimum. This has been achieved through the development of rating scales, which specify explicit criteria for assigning scores to samples of learners' written or spoken production, and detailed descriptors of expected performance at each level. In addition, raters often undergo training in the interpretation of the scoring scales and the application of rating criteria. Finally, frequently at least two raters are assigned to evaluate a single sample, and in the case of large discrepancies between them, an additional rater is brought in. The great majority of scales applied in the evaluation of learners' spoken or written production list vocabulary use as one of the assessment criteria. Lexis can be evaluated together with other features of the learner's performance, with the result in the form of a single score allotted by the rater, if a holistic rating scale is employed. It can also be scored separately through one of the analytic scales, each focusing on a different aspect of linguistic and communicative ability, and then the mark for vocabulary is entered into an algorithm together with other marks in order to determine the global score. The simplest algorithms are a sum of the component scores or their average, more complex ones may include different weighing for the component marks in the final score. Different assessment scales are more or less specific about the aspects of vocabulary use which need to be taken into account. The different types of rating scales will be discussed in more detail in Chapter 3.

In the last decades extensive research has explored the possibility of automatic scoring of learners' written (and to a lesser extent also spoken) output. This trend has focused on the assessment of the learner's extended written responses in the form of essays. Automated scoring uses techniques and tools developed within the field of statistical linguistics, stylometry and natural language processing. It produces multiple measures for a single essay, each describing one particular aspect of performance (for example mean sentence length) which can be estimated without human intervention by computer software designed especially for this purpose. Similarly to analytical human scoring, the final score for an essay is the result of an algorithm incorporating these indices, yet this algorithm usually includes more features than in the case of human scoring and it is much more complex. The advantages and disadvantages of this method of evaluation are summarised in the quotation from the Research Report published by Educational Testing Service, the institution

which has developed and administers one of the most popular tests of English as foreign language worldwide called TOEFL (Test of English as a Foreign Language):

> Automated scoring in general can provide value that approximates some advantages of multiple-choice scoring, including fast scoring, constant availability of scoring, lower per unit costs, reduced coordination efforts for human raters, greater score consistency, a higher degree of tractability of score logic for a given response, and the potential for a degree of performance-specific feedback, that is not feasible under operational human scoring. These advantages, in turn, may facilitate allowing some testing programs and learning environments to make greater use of CR [constructed response] items where such items were previously too onerous to support. However, accompanying such potential advantages is a need to evaluate the cost and effort of developing such systems and the potential for vulnerability in scoring unusual or bad-faith responses inappropriately, to validate the use of such systems, and to critically review the construct that is represented in resultant scores.
>
> (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012, p. 2)

Vocabulary is among linguistic features which lend themselves well to an automated analysis and lexical indices usually feature prominently in automated scoring systems. So far only a handful of automated scoring systems have been developed and implemented. An example of such a system is the *e-Rater* scoring engine[1] produced by ETS and used for scoring two written tasks in the TOEFL exam. It includes several indices related to vocabulary use such as average word length, good collocation density or wrong or missing words.

In both methods of assessment—human rating and automated scoring—lexical proficiency is understood as the L2 learner's use of vocabulary in a communicative task. This use is assumed to be the reflection of the learner's underlying vocabulary ability, which includes both the learner's control of different lexical forms, meanings and use restrictions as well as his or her strategic competence to draw on the available stock of words to choose those which are suitable for completing the task successfully.

2.5 Vocabulary Tests

Different tests of vocabulary, which will be discussed in this section, have employed different tasks to elicit samples of test-takers' production in order to observe, measure and evaluate L2 learners' lexical proficiency. In consequence, their results reflect different aspects of this proficiency and by extension address different components of lexical competence.

In the introduction to their book on language testing, Bachman and Palmer (1996) state that one best language test does not exist. The suitability of a test for a given purpose depends on many factors. One of these factors concerns particular uses that a test is supposed to be put to. Vocabulary assessment has been performed within two areas of applied linguistics: foreign language education, and research in second language acquisition. Both domains, even though closely related, have different foci and different agenda. In consequence, vocabulary assessment performs a different function in each of these fields. In foreign language education it is an instrument for making decisions about learners and providing them with feedback on their abilities; in research on second language acquisition, on the other hand, it is a tool to gain a better understanding of the L2 learning process. Since both disciplines have different goals, the language tests they employ have to differ as well. In addition to being well grounded in the theory of language learning and testing, language tests used in education have to be, above all, practical. Thus, they have to be easy to administer and mark and their results have to be easily interpretable by test users. On the other hand, researchers are not concerned with issues of practicality. They usually use their tests with a limited group of people—research subjects—thus their instruments can be longer and more complex. Moreover, the results have to be meaningful only to the researcher himself or herself and a relatively small circle of other researchers in the area. Thus, while there has been some exchange of know-how between the two disciplines, the instruments employed by each field have been different.

Read (2000) points out that both traditions in vocabulary testing have had their own weaknesses. On the one hand, language testing theorists tend to ignore the recent research in second language acquisition, which brings the lexicon to the centre of the second language learning process and they do not give enough importance to the assessment of learners' lexical proficiency. On the other hand, second language acquisition researchers tend not to pay sufficient attention to most recent trends in testing concerning test validation. This section will discuss in detail vocabulary tests developed within both traditions. It will examine critically the purposes behind their design and the uses they have been put to. The strong points and weakness of each test will also be highlighted.

2.5.1 Vocabulary Testing for Educational Purposes

In the area of language education, vocabulary is assessed for a variety of practical purposes. Achievement tests are used to check whether learners have mastered the items that have been focused on within (q section of) a language programme. Diagnostic tests allow teachers to detect gaps in learners' lexical knowledge in order to prepare remedial work. In placement tests it is checked whether learners' vocabulary is sufficient to follow a particular language course. Finally, proficiency tests are used to

establish the level of learners' lexical mastery, which forms part of general language ability. The assessment procedures frequently used in these contexts are discrete-point tasks.

Achievement/progress tests are by definition selective, as they focus on a set of particular words introduced in the classroom. They are frequently prepared by teachers themselves; however, sometimes they are developed by publishers and accompany EFL coursebooks. Tasks employed in such tests usually include matching, multiple choice, writing short answers (e.g. translations or synonyms) and gap-filling.

Placement tests rarely focus on vocabulary, but some specific words or phrases may be targeted by individual questions. However, learners' lexical proficiency is tested indirectly through items addressing learners' knowledge of grammar, their pragmatic competence or their receptive skills. For example, ***Oxford Online Placement Test***[2] and ***Cambridge Placement Test***[3] comprise multiple-choice and gap-filling items which address both form and meaning of grammatical structures and individual words as well as the understanding of larger stretches of spoken and written discourse.

There are few truly **diagnostic tests** available in the area of EFL education. A widely known example of such a test is ***Dialang***,[4] developed by a group of experts in language testing from several European universities working under the auspices of the Council of Europe. Unlike the placement tests already discussed, *Dialang* contains a separate section devoted solely to the assessment of a learner's lexical proficiency. This section consists of 30 multiple-choice and gap-filling questions, most of which are set in the context of a sentence. They test four aspects of lexical knowledge: semantic relations, word combinations, word formation and meaning, and feedback is provided for each of these aspects separately. The final score for this part is expressed by a CEFR level. A very original feature of the programme is a 75-item vocabulary task of the yes/no format given at the beginning of the session for initial placement. Its result is also expressed according to the CEFR levels. Based on the outcome of this test section the difficulty level of the items in the test proper is selected by the programme. This seems to be the only use of the yes/no format in the assessment of vocabulary in the educational context and the only instance in this domain where vocabulary by itself serves as an indicator—even if only preliminary—of the overall language proficiency.

The assessment of vocabulary plays an important role in **proficiency tests**. A series of widely recognised proficiency exams offered by Cambridge English Language Assessment—***Cambridge English Key (KET)***, ***Preliminary (PET)***, ***First (FCE)***, ***Advanced (CAE)*** and ***Proficiency (CPE)***—all assess learners' lexical proficiency through both discrete and embedded tasks. In *KET* and *PET*, designed for lower proficiency levels (A2 and B1 respectively), discrete vocabulary items are part of the Reading and Writing paper of the exam. The tasks employed for checking the learner's

awareness of various aspects of vocabulary knowledge related to form, meaning and use include multiple-choice questions set in individual sentences and texts, short answers in response to a definition and text-based gap-filling. In the *FCE*, *CAE* and *CPE* exams, pitched at B2, C1 and C2 levels respectively, discrete vocabulary questions are part of the Use of English paper. The assessment of various aspects of the learner's lexical knowledge is addressed in each of the first four items of this paper called Multiple-choice cloze (a text-based multiple-choice task), Open cloze (a text-based gap-filling task), Word formation (another text-based gap-filling task with prompts) and Key word transformations (sentence transformation).

In all the five Cambridge exams, vocabulary is also assessed through communicative tasks: reading, listening, writing and speaking; however, in the case of receptive skills, lexical proficiency is not evaluated explicitly but together with other components of linguistic ability, through checking the understanding of written and spoken discourse. On the other hand, vocabulary forms one of the assessment criteria of the productive tasks, involving writing a text (e.g. an essay, a letter, etc.) and engaging in a conversation on a given topic. More about the details of assessing these sections of the exam can be found in Chapter 3.

Another popular proficiency exam is ***Test of English as a Foreign Language (TOEFL)***, developed and administered by the English Testing Service based in the US. It is one of the most widely recognised exams measuring the learner's ability to use English at the university level. It is frequently required of non-native English speakers for admission to English-medium tertiary education programmes. *TOEFL* consists of four sections: Reading, Listening, Speaking and Writing. Its Reading section contains a few items addressing explicitly and directly the candidate's lexical knowledge. This is done by checking the understanding of the meaning of a specific word or a phrase in context. As in all the previous exams mentioned, vocabulary knowledge is also evaluated indirectly through testing reading and listening comprehension. Vocabulary is also part of the assessment of the productive skills. The rubrics used for evaluation of the writing and speaking tasks also include vocabulary as one of the criteria, but lexical knowledge is not evaluated separately. The details of the scoring procedures of these tasks are also presented in Chapter 3. Unlike the Cambridge English exams, *TOEFL* does not contain a section addressing more direct vocabulary knowledge.

Of all the tests discussed, only *Dialang* reports a separate mark referring to the learner's lexical proficiency in the final report. While vocabulary is still being tested as an independent component of language in short achievement or diagnostic tests developed by teachers or institutions for their own purposes, placement tests and large high-stakes proficiency tests such as the Cambridge English exams or *TOEFL* treat vocabulary as part of global linguistic ability and do not contain a separate part addressing specifically lexical proficiency. If vocabulary is tested directly,

it is assessed together with grammar, and the exams do not report a separate mark for lexical ability.

2.5.2 Vocabulary Testing for Research Purposes

Tests designed specifically to evaluate the learner's lexical proficiency have been more common among researchers exploring the process of second language vocabulary acquisition than language teachers and professional test developers. Contrary to the recent trends in foreign language testing, vocabulary researchers have continued to treat the lexicon as an independent component of linguistic competence and have focused on its size, organisation, access and development in the process of second language learning. In studies tackling such issues as EFL students' vocabulary gains after a period of study abroad, or the acquisition of vocabulary through reading, instruments were needed which could evaluate specifically L2 learners' lexical competence and proficiency. Due to the fact that there have been no suitable tests available in the field of language education, researchers have had to develop their own instruments to assess learners' vocabulary. Moreover, these researchers frequently have had very specific research questions in mind, which their instruments need to address precisely, for example how words are learned through particular tasks or how vocabulary ability develops over time. Finally, some researchers took up the development of valid and reliable instruments to assess lexical proficiency as the very aim of their investigations. Examples of the most popular tests used in studies on the L2 lexicon and its acquisition are discussed next.

The most popular tests among researchers interested in second language vocabulary acquisition are the instruments assessing the size of the L2 learner's lexicon. These tests do not target the learner's knowledge of particular lexical items, but attempt or deliver an estimate of how many items altogether are stored in the learner's lexicon at a particular moment in time. All these assessment procedures are constructed on the assumption that the order of acquisition of lexical items generally follows the order of frequency of words in language. That is to say, the learner is more likely to know very frequent words before he or she masters rare items.

One of the first tests of this kind was Meara and Jones's ***Eurocentres Vocabulary Size Test*** (***EVST***; Meara & Jones, 1990). It had a yes/no format, which requested that learners simply decide if they know the presented words, without requiring any evidence. The test consisted of randomly selected ten-item samples from ten consecutive 1000-word frequency bands. In addition, ten non-words accompanied each sample to estimate the learner's willingness to take risks with his or her answers and overstate his or her lexical competence. The test was delivered through a computer programme and it was adaptive: once the criterion of correct responses in a 20-item set of words and non-words corresponding to one frequency band

was met, the programme moved the learner to the next frequency level; if not, the learner was tested with an additional set of words drawn from the same band to produce a more accurate estimate at this level. Based on the answers to the presented sets of words, the estimate of the learner's knowledge of 10,000 most frequent words of English was computed.

Interestingly, the initial version of the test was commissioned by a European chain of language schools called Eurocentres as a placement test. The test was very practical in comparison with the more extensive placement examination used earlier by the schools, as it could be taken individually, took only about ten minutes to complete and the computer programme delivered the score immediately. The validation of this instrument for this purpose produced satisfactory results (Meara & Jones, 1988). However, the test did not gain a wide popularity in education. The critics insisted that it was too far-fetched to assign students to levels based on one single-format language test which captured a fairly superficial knowledge of one component of general linguistic proficiency (Read, 2000).

Meara continued to work on the original version of the vocabulary size test. A few years later a pen-and-pencil version of the test was made publicly available checking the test-taker's knowledge of the first and second 5000 most frequent words (Meara, 1992, 1994).[5] The booklets contained a battery of checklists, each consisting of 40 words and 20 non-words corresponding to a particular frequency level. Over a decade later computerised versions of the tests were released: ***X-Lex*** for the 1–5K frequency bands (Meara & Milton, 2003) and ***Y-Lex*** for 6K–10K levels (Meara, 2006). Both presented learners with 120 words—20 from each target level and 20 non-words. Most recently, the ***V-Yes/No*** test has been made available to complete by learners online (Meara & Miralpeix, 2015). It presents the test-takers with the total of 200 words and non-words and provides the estimates for 1–10k frequency bands. All these releases of the tests of vocabulary size reflect Meara and his colleagues' striving to improve the validity, reliability and practicality of the instrument, by choosing more solid reference for frequency bands, improving the format of non-words and finally working on the statistical formula which uses the positive answers to non-words (false alarms) to correct the final score (Read, 2000, p. 128).

The main criticism levelled against the multiple versions of the vocabulary size tests in this format is that they address the most superficial layer of lexical competence—passive recognition of written word forms. Yet, Meara (1990, 1996a) refutes this argument by admitting that although on surface it is this aspect of vocabulary which is being evaluated by the test, in practice passive vocabulary stays in a proportional relation to the learner's active knowledge. He admits, however, that the exact relation between the passive and active vocabulary has not been determined. In spite of the criticism, the tests were applied in research on second

language vocabulary acquisition (e.g. Milton & Meara, 1995; Milton, Wade, & Hopkins, 2010).

One more development in this area which needs to be mentioned is the aural version of vocabulary size test. Milton and Hopkins (2006) designed the ***A-Lex*** test which was an equivalent of the *X-Lex* test mentioned previously, but words were presented to learners in the spoken rather than written form. This test addresses the recognition of the aural word forms and was considered to be a better instrument to investigate the relationship between learners' lexicons and their aural skills (Milton & Hopkins, 2006; Milton et al., 2010). Unfortunately, the test has not been widely available.

At approximately the same time another instrument to measure EFL learners' vocabulary size was proposed by Nation (1983) and Nation (1990). Like in the case of *Eurocentres Vocabulary Size Test*, the ***Vocabulary Levels Test*** was originally devised for educational purposes—as a diagnostic tool for teachers to help them decide where learners need to be offered help with vocabulary learning. Yet, in the absence of other tools it was soon adapted by researchers for the purpose of measuring the vocabulary size of EFL learners. Nation chose the matching format for his test, which required test-takers to match words and their definitions. According to Nation (1990, p. 261), the format has advantages of being quick for test-takers to solve and for teachers to mark, provides evidence that the learner indeed knows the words and reduces the chances of guessing. The test includes words belonging to four frequency levels: 2000, 3000, 5000 and 10,000. In addition, it assesses learners' knowledge of lexical items belonging to the University Word List—an inventory of non-technical academic vocabulary which is indispensable in order to study in English at the tertiary level (Xue & Nation, 1984). Each of the five sections contains 36 words, grouped in six sets of six words, and 18 definitions, three for each set. The words in one set represent the same part of speech so as the definitions would not provide additional clues. Yet, the target items are unrelated in meaning in order to capture learners' rough idea of the words' senses rather than fine semantic distinctions. The definitions are written using words belonging to lower levels. In 1993, Norbert Schmitt revised slightly the original version of the *Vocabulary Levels Test* and added three equivalent versions of the same instrument (versions B, C and D; Schmitt et al., 2001).

As the test was originally developed for educational purposes, Nation recommended that scores for each of its sections be reported separately as they are more informative to the teacher than the total score as far as the gaps in learners' English vocabulary are concerned. Laufer (1998, p. 269), however, proposed the method of converting the raw scores into an estimate of the learner's vocabulary size. She extrapolated the scores for the tested levels to the missing levels. Next, the scores for the 1000–5000 bands were added, multiplied by 5000 and divided by 108 (18

items per level for 6 levels, 1, 2, 3, 4, 5 and UWL). Laufer did not include higher bands in her calculations.

Even though the Vocabulary Levels Test quickly gained popularity for measuring the size of vocabulary knowledge for research purposes, for a long time almost no attempts were undertaken to run its thorough validation study, with two notable exceptions: Read (1988) and Beglar and Hunt (1999). Schmitt et al. (2001) undertook a large-scale validation analysis, which involved a sample of 801 learners of English and employed a range of quantitative and qualitative methods. As a result, the original four versions were revised and combined into two extended versions of the test (Schmitt, 2000; Schmitt et al., 2001). The new versions have the same format as the original test, but their four sections covering different frequency bands contain ten sets of six words instead of six sets—as in the original test—and the Academic Vocabulary List-based level (Coxhead, 2000), which replaced the earlier UWL section, contained two sets more.

The different versions of the *VLT* address learners' receptive knowledge of written vocabulary. In 1995, Laufer and Nation proposed the active version of this test, whose aim was to estimate the size of the learner's productive lexicon. Parallel to the passive test, it also consists of five sections targeting words at four frequency levels (2K, 3K, 5K and 10K) plus the words from the University Word List, and each section also includes 18 target words. However, instead of a matching task, the active version is based on the C-test format. Learners are exposed to sentences with the target words missing and they are expected to produce these items based on the context. To ensure that test-takers will retrieve the right word, and not some other semantically related item, the word beginning (two to six first letters) is provided. Laufer and Nation (1995), Laufer (1998) and Laufer and Nation (1999) provide some evidence supporting the validity of the test, yet a thorough validation study has not been conducted. Read (2000, pp. 125–126) and Schmitt (2010, pp. 203–204) discuss some threads to the validity of the test arising from the C-test format and the selection of target words. Yet, the active version of the VLT has been used by researchers along its passive version to measure the growth of learners L2 lexicons and the factors that may influence this process (Laufer & Nation, 1995; Waring, 1997; Laufer, 1998; Laufer & Paribakht, 1998; and more recently Yamamot, 2011).

Nation and Beglar (2007) proposed a new instrument addressing the breadth of lexical competence. Unlike the *Vocabulary Levels Test*, designed as a diagnostic tool for language teachers, the ***Vocabulary Size Test (VST)*** was developed as "a proficiency measure used to determine how much vocabulary learners know" (p. 10). The researchers list intended uses of the test, which include charting the growth of learners' vocabulary or comparing the vocabularies of native and non-native speakers. The test consists of 140 items and it samples ten words from each 1000-level of the

1400 most frequent word families (Nation, 2006). Unlike the *VLT* it has no separate section devoted to academic vocabulary, but academic words are spread across different levels. The test has a multiple-choice format and it requires a test-taker to choose the best definition of the target word among four options. The task is more demanding than in the case of the *VLT* because the correct answer and the three distractors share elements of meaning. Thus the learner needs to have a fairly developed grasp of the word's meaning in order to choose the appropriate definition. The definitions were written with a restricted choice of words of higher frequency than the target words. The target items are presented in a sentence, yet the context is neutral and gives no clue of the word's meaning; it only provides information on its part of speech. The estimate of vocabulary size is computed from the test score in a straightforward way. The final score needs to be multiplied by 100 to get a learner's total vocabulary size (in the range of 14,000 most frequent word families). Since there is no need to extrapolate the scores for missing 1000-word bands, the estimate is much more accurate than in the case of the *VLT*. Beglar (2010) and Gyllstad (2012) conducted validation studies of the *VST*. Their results demonstrated adequate reliability of the test. They also confirmed the expectation that test-takers should get higher scores in higher frequency bands and lower scores in lower-frequency bands. Yet, both studies also found that some items needed revision. The *VST* has been gaining popularity among researchers and it has been used in studies into vocabulary proficiency as the primary instrument measuring vocabulary size. For example, Uden, Schmitt and Schmitt (2014) employed it to estimate the lexicon of four ESL students enrolled in a graded reading programme in the UK. The development of *VST* is ***Phrasal Vocabulary Size Test, BNC Version*** (1–5K)[6] (Martinez & Schmitt, 2012).

A very different approach to measuring the size of a learner's vocabulary size was adopted in a computer programme called ***Lex_30***[7] (Fitzpatrick, 2006; Meara, 2009, Chapter 3). Similarly to the productive version of the Vocabulary Levels Test, the programme aims to assess the breadth of a learner's productive lexicon. The instrument has a format of a word association test. A learner is requested to provide up to four words that come to their mind in response to each of 30 carefully selected target words. The programme does not check the validity of the associations produced by the learner, but analyses the number of responses and their frequency in language. Only low-frequency responses are awarded a point. The final score can theoretically range from 0 (no low-frequency associates) to 120 (all associations are low-frequency words). However, even native speakers get a score of around 60, because their associations usually include both high-frequency and low-frequency items. Unfortunately, the score cannot be translated into an estimate of the number of words a learner can use. However, the test can be used to compare individual learners in terms of the breadth of their vocabulary. Although

Meara (2009, Chapter 4) demonstrates the concurrent validity of this instrument, he admits that the programme is still in its experimental phase and its results should be interpreted with caution. Meara also remarks that his instrument measures the productive recall ability, rather than productive use ability.

In addition to tests assessing the breadth of learners' lexicons, researchers need instruments which would allow them to gauge the depth of L2 word store. However, while the size of the mental lexicon is a relatively intuitive concept, it is much more difficult to define what exactly is the depth of lexical competence and how it should be measured. Gyllstad (2013) and Schmitt (2014) review different approaches to conceptualising and operationalising the depth component of the lexicon, and the consequence they have for the assessment of vocabulary knowledge. Two approaches can be distinguished in tests targeting the quality of the L2 lexicon (Schmitt, 2010, p. 216). The first one, the component approach, is based on the compositional models of word knowledge such as the one proposed by Nation (1990) and discussed in Chapter 1. Instruments within this approach are constructed with an assumption that the knowledge of a word consists of several components and each of these components can be assessed and measured separately.

An **oral interview** is the best method of eliciting rich information from learners, which could attest to their familiarity with multiple aspects of knowledge of individual items (Read, 2000). A researcher prepares a selection of words and a series of open-ended questions aimed at eliciting various aspects of word information about each item from a learner. Read also describes a **written version** of this procedure involving three steps. First, learners are requested to self-assess their familiarity with an item; next they are presented with three tasks which involve (1) writing two sentences with the target word and its two collocates, (2) providing three collocates of the target word and (3) providing three derivatives of the target word in context; and finally, they are to write an explanation of the item's meaning in their own words. Both instruments make it possible for a researcher to assess the learner's knowledge of the target item's senses, collocations, derivations and grammatical behaviour, with the oral interview giving a better opportunity of detailed probing into the learner's lexical proficiency. However, as Read himself admits, both procedures are very impractical. They are very time consuming and only a small selection of words can be assessed in this way. Furthermore, consistent scoring is problematic. Finally, even these demanding procedures do not elicit all possible aspects of word knowledge. Schmitt (2010) remarks that some elements of word knowledge, such as a learner's intuitions about a word's frequency are difficult to be elicited and measured adequately.

A solution to the challenge of addressing several components of word knowledge in one instrument is to focus on a single element and

to quantify a learner's knowledge of this element only. Schmitt (2014, p. 942) remarks that in fact all vocabulary size tests measure depth of knowledge at the same time because "[s]ize by definition is the number of lexical items known to some criterion level of mastery. But the criterion will always be some measure of depth". This is visible in the difference between yes/no, matching and multiple-choice formats discussed in the previous section. All these tests address a different component of vocabulary knowledge—the recognition of the written (aural) form, some familiarity with its meaning or a firm grasp of the word's semantic intricacies. Yet, these assessment procedures fail to differentiate between different levels of word knowledge for individual items and this is what makes them different from the true tests targeting the depth of lexical competence.

Such an approach was adapted by Read (1998) in his ***Word Association Test***.[8] It tackles a learner's familiarity with non-technical academic vocabulary and consists of 40 target adjectives. Each of the adjectives is followed by two groups of four words and a test-taker's task is to choose among them four words which are associated in some way with the target word. One group of options represents paradigmatic associations with the target word (synonyms or adjectives representing one aspect of meaning of the target item), the other group represents analytic relations (collocates). To diminish the effect of guessing, the number of correct associates belonging to each group is not stable. This instrument, in fact, makes it possible to elicit a lot of information about a word: its different senses, its collocations and paradigmatic meaning relations. In addition, unlike vocabulary size tests, it treats the knowledge of a word as gradable. Each item can receive 0, 1, 2, 3 or 4 points (depending on the number of associations a learner chooses correctly) and the score reflects the quality of a learner's knowledge of this item. The validity of the test was established by Read (1998). Yet, Schmitt (2010, p. 227) voices his reservations against interpreting intermediate scores. He maintains that it is not clear to what extent they represent lucky guesses rather than true knowledge. In spite of this criticism, the test has been used by researchers to gauge the quality of L2 learners' vocabulary knowledge. For example, Qian (2002) used it to investigate the roles of breadth and depth of vocabulary knowledge in reading comprehension in academic settings. The *Word Association Test* checks a learner's familiarity with academic adjectives; thus, unlike the tests of vocabulary size, it is not suitable for all learners irrespective of their level. Yet, this format was adapted by other researchers, who developed other tests matching their learners' profiles more closely. One example is Schoonen and Verhallen (2008), who used the word association task to investigate the depth of vocabulary knowledge of children aged 9–12 speaking Dutch as an L1 and L2.

The second approach to measuring the depth of L2 word store is based the developmental scale discussed in Chapter 1. The example of

this approach to measuring the depth of vocabulary knowledge is the ***Vocabulary Knowledge Scale*** proposed by Paribakht and Wesche (1993) and Paribakht and Wesche (1997). It has a format of a self-report and requires the learner to assess their familiarity with a word on a 5-level scale. Test-takers choose the category which in their opinion represents best their knowledge of a word. For the first two categories, the learner's response is taken at its face value, but for the next three stages, they are requested to provide evidence to support their self-evaluation. The scoring of learners' responses is fairly complex and it is reproduced in Figure 2.4.

Wesche and Paribakht (1996) admit that their test does not target the depth of the whole L2 lexicon but the quality of a learner's knowledge of individual words chosen for evaluation. They used the *VKS* as an instrument to trace gains in vocabulary knowledge in learners exposed to two different methods of teaching vocabulary. The test proved sensitive enough to pick up changes in the quality of the knowledge of the target words over a period of one semester. Yet, both Read (2000) and Schmitt (2010) list a number of problems inherent in the instrument. In particular, there seems to be a gap between categories I–IV on the one hand, which all address the form, and the basic meaning of the word, and category V on the other, which requires an extensive knowledge of several aspects of word information including its grammatical behaviour, collocations or stylistic constrains. In addition, test-takers may provide sentences which are too neutral to give sufficient evidence of the learners' familiarity with these different aspects of lexical competence.

The limitations of the tests tackling the depth of vocabulary knowledge is that the very concept of the quality of word information is multifaceted and it cannot be captured in one assessment procedure and one instrument. Thus, a thorough evaluation of a learner's familiarity with

Self-Report Categories	Possible Scores
I I don't remember having seen this word before.	1
II I have seen this word before but I don't know what it means.	2
III I have seen this word before and I think it means _____.	3 or 2
IV I know this word. It means _____.	3 or 2
V I can use this word in a sentence. E.g.: _____.	5 or 4 or 3 or 2
Meaning of scores	
1—The word is not familiar at all.	
2—The word is familiar but its meaning is not known.	
3—A correct synonym or translation is given.	
4—The word is used with semantic appropriateness in a sentence.	
5—The word is used with semantic appropriateness and grammatical accuracy in a sentence.	

Figure 2.4 The Scoring Scheme in the *Vocabulary Knowledge Scale*

Source: Paribakht & Wesche (1997, p. 181)

multiple aspects of word information is not feasible. In addition, the various instruments proposed by researchers do not report on the quality of the learner's lexicon but the degree of familiarity with individual lexical items. Unlike the tests of vocabulary size, which make it possible to draw conclusions about the volume of a learner's lexicon based on his other familiarity with its sample, vocabulary depth tests have never made claims of being representative of the entire vocabulary. Meara and Wolter (2004) maintain that the juxtaposition of breadth and depth of vocabulary knowledge in research literature is rather unfortunate because the size is the property of the lexicon and the depth is the property of individual words (see Chapter 1). They propose that instead of tackling the quality of a learner's lexical information about selected words, assessment of vocabulary knowledge should address the organisation and complexity of the whole lexicon. Wilks, Meara and Wolter (2005) propose an instrument which is designed exactly for this purpose: a computer programme called ***V_Quint***.[9] It also has a word association format, but instead of choosing a number of clearly defined (syntagmatic and paradigmatic) associations for the target words, a learner has to find a link of any type between any two items in a set of five randomly selected high-frequency words. The programme does not check if the response is valid, but compares the number of links identified by a test-taker with the probability of finding a pair of associated words within a randomly selected set of five words. It then extrapolates from this data to estimate the number of associational connections in this core vocabulary. The result is a score that varies from 0 to 10,000. Meara admits that his approach is still an ongoing research into vocabulary measures, and *V_Quint* is an exploratory and experimental test, rather than a definitive instrument.

2.6 Conclusion

Lexical competence, as part of larger constructs of language competence and communicative language ability, has attracted attention from both SLA researchers and language education professionals. As an abstract concept and a property of human cognition, lexical competence cannot be observed directly, and its characteristics can only be inferred from concrete human behaviour. Tests are instruments designed to elicit samples of human verbal performance, which allow language users to exhibit their lexical proficiency. As a behavioural construct, lexical proficiency can be studied directly, and thus described, measured and assessed.

Many types of tests have been proposed and developed in order to elicit most relevant verbal behaviours, which will facilitate most accurate extrapolations concerning the breadth, depth and accessibility of L2 users' lexicons. Such tests can target either specific lexical items or the lexicon as a whole. They may also address more directly the specific dimensions of lexical competence. Some other tests focus on a broader

construct of vocabulary ability which involves L2 users' capacity to use lexis in communication. The remaining chapters will focus on this last type of test and discuss ways of assessing and measuring L2 users' general lexical proficiency in naturalistic language production.

Notes

1. www.ets.org/erater/about (accessed 23 December 2018)
2. www.oxfordenglishtesting.com/defaultmr.aspx?id=3048 (accessed 23 December 2018)
3. www.cambridgeenglish.org/find-a-centre/exam-centres/support-for-centres/placing-students-in-the-right-exam/ (accessed 23 December 2018)
4. https://dialangweb.lancaster.ac.uk/ (accessed 23 December 2018)
5. Most of the resources mentioned in this paragraph are available form www.lognostics.co.uk/ (accessed 23 December 2018)
6. www.lextutor.ca/tests/levels/recognition/phrasal/ (accessed 23 December 2018)
7. www.lognostics.co.uk/tools/ (accessed 23 December 2018)
8. www.lextutor.ca/tests/associates/ (accessed 23 December 2018)
9. www.lognostics.co.uk/tools/ (accessed 23 December 2018)

3 Performance-Based Assessment of Lexical Proficiency

3.1 Introduction

Many SLA researchers and education professionals point out that the most adequate method of assessing L2 learners' language proficiency is evaluation of their performance in authentic tasks. These tasks generally involve writing a text, or delivering a monologue or a dialogue, in response to a specific prompt. Assessing learners' knowledge or ability based on their extended spoken and written production has a long tradition in education (see Spolsky, 1995 for a short history of the origins of examinations in Europe and the United States). Yet, throughout the previous century the process of evaluation of L2 learners' production has changed considerably. Until the middle of the 20th century the assessment of writing was influenced by the form-oriented approach (Behizadeh & Engelhard, 2011) concentrating on linguistic accuracy and demonstration of linguistic proficiency. The 1950s and 1960s witnessed a loss of popularity of writing and speaking tasks, as they were replaced by discrete-point instruments measuring various aspects of L2 production in a more objective fashion. It was only with the onset of the communicative approach to language teaching in the 1970s that test items requiring extended production regained their important role in language assessment.

3.2 Performance Assessment

Eliciting a sample of writing from a learner for assessment purposes is frequently referred to as a **direct writing test**. It can be defined as a test that requires from a candidate to produce at least one piece of extended, continuous and structured text in response to a task involving a set of instructions or a prompt. Direct writing tests are considered to instantiate performance assessment, which is a cover-all term for the evaluation of the learner's use of language in authentic (or semi-authentic) tasks. McNamara (1996) defines performance assessment in the following way:

> A defining characteristic [of performance assessment] is that actual performances of relevant tasks are required of candidates, rather

> than more abstract demonstration of knowledge, often by means of pencil-and-paper tests.
>
> (McNamara, 1996, p. 6)

He observes that L2 performance assessment has drawn heavily on practices in non-language fields; however, it has a special status in the second language context:

> Second language performance assessment is distinguished from performance assessment in other contexts because of the simultaneous role of language as a medium or vehicle of performance, and as a potential target of assessment itself.
>
> (McNamara, 1996, p. 8)

McNamara explains that, on the one hand, a performance test involves a second language as a *medium* of performance and the performance of the *task* itself remains the *target* of assessment; on the other hand, the purpose of a task is to elicit a language sample so that *second language proficiency* may be assessed (pp. 43–44). In other words, unlike in other contexts, both task performance and language performance are being evaluated simultaneously. He also points out that performance tests vary in the relative importance attributed to task and language performance. In the performance tests in the 'strong' sense the focus is on the performance of the task. The test task closely resembles a real-world activity and performance is evaluated based on real-world criteria of its successful execution. Adequate second language proficiency is a necessary but not sufficient condition for a successful completion of the task; other non-linguistic factors play an equally important role, such as background knowledge, creativity, professional competence (in the cases of language for special purposes assessment) or general communication skills. Furthermore, some linguistic deficiencies can be compensated for by non-linguistic factors facilitating successful communication. This idea is well illustrated by the quotation by Jones:

> With regard to second language performance testing it must be kept in mind that language is only one of several factors being evaluated. The overall criterion is the successful completion of a task in which the use of language is essential. A performance test is more than a proficiency test of communicative competence in that it is related to some kind of performance task. It is entirely possible for some examinees to compensate for low language proficiency by astuteness in other areas. For example, certain personality traits can assist examinees in scoring high on interpersonal tasks, even though their proficiency in the language may be substandard. On the other hand, examinees who demonstrate high general language proficiency may not score well on a performance because of deficiencies in other areas.
>
> (Jones, 1985, p. 20, cited after McNamara, 1996)

This approach to performance assessment is characteristic of what McNamara calls 'work sample' tradition of language testing, whose aim is purely pragmatic and involves a selection of candidates with sufficient language proficiency to function successfully in the second language, usually in an academic or vocational context.

In the performance assessment in the 'weak' sense, the learner's capacity to perform the task is not the focus of assessment. Instead, the focus is on language proficiency, that is the quality of language used to carry out the assigned task, and on what it reveals about the candidate's linguistic ability. The task may resemble or simulate a real-world activity, but the requirement of authenticity is less prominent, as the main purpose of the task is to engage a candidate in an act of communication in order to elicit a sufficient sample of language performance for evaluation. The criteria of assessment concentrate on the quality of language, yet the non-linguistic factors related to successful completion of the task do not stay without their influence on the outcome of assessment. Nevertheless, as McNamara observes, even task-related aspects of performance such as the overall fulfilment of the task are evaluated through the lens of language. Such approach is characteristic of what McNamara recognises as a cognitive and psycholinguistic tradition in performance assessment. It has its roots in traditional testing of the first half of the 20th century, as well as in discrete-point testing of the 1950s and 1960s, which both aimed to make judgements about the learner's L2 ability and his or her underlying L2 knowledge. Yet, this approach is also shaped by the theories of communicative competence, which recognised the role of pragmatic and sociocultural competences in second language competence (cf. Chapter 1).

McNamara admits that "This dichotomy is a conceptual one, and is presented as a way of clarifying issues in actual tests; pure examples of either types will be difficult to find" (p. 43). He further observes that most language performance tests represent the 'weak' sense of the term, in that they focus on language performance rather than task performance. Yet, he remarks that "such tests may still be distinguished as relatively stronger or weaker, and even different parts of a single test may be distinguished in this way" (p. 45). This implies that even though L2 assessment instruments aim at making an extrapolation concerning the candidate's linguistic ability, the results are also influenced by non-linguistic and extra-linguistic criteria related to successful completion of the task.

A direct writing (and spoken) test is a good example of this phenomenon. Writing a text is a complex cognitive process embedded in a social and cultural context and involves many extra-linguistic abilities which interact with purely linguistic components to generate the final product (see Weigle, 2002 for a review). They include topical knowledge, familiarity with writing conventions or an awareness of the interlocutor's or the reader's expectations. An effective piece of writing (or an effective

spoken communication) depends on a good mastery of linguistic skills but also of all the other extra-linguistic aspects of production. Moreover, the linguistic and extra-linguistic aspects are frequently interwoven. For example, good topical knowledge is related to an extensive use of specialised vocabulary, an awareness of the reader may in contrast lead to choosing less complex grammatical structures or words. This interdependence constitutes a potential problem in the exploitation of L2 learner extended production for assessment purposes and it needs to be carefully accounted for in the assessment procedure. This issue was discussed by Weigle in the following quotation:

> In a language test (and I am still considering writing tests to be a subset of language tests for the purposes of this discussion), we are primarily interested in language ability, not the other components of language use that are involved in actual communication. Nevertheless, we need to think about these components when we are designing tests, so that we can specify as explicitly as possible the role that they play in the successful completion of the test tasks. In some cases they may be included in the definition of the ability we are interested in testing, whereas in others, we may want to reduce their effect on test takers' performance and thus on their test scores.
>
> (Weigle, 2002, p. 45)

The following sections of this chapter will review a number of tests or test tasks which elicit candidates' written production, and will attempt to classify them as 'weaker' or 'stronger' performance tests. They will also point to the consequences of each approach for vocabulary assessment. Speaking tasks will also be mentioned, as they share some similarities with direct writing tasks in terms of the assessment process, but since speaking items are not the focus of this book, they will not be discussed in detail.

3.3 The Process of Writing Assessment

The process of assessing performance is much more complex than scoring the kinds of tests discussed in the previous chapter. In the case of indirect tests, the learner's linguistic behaviour is strictly regulated by the applied instrument. The correct answers are determined in advance and marking the test is usually a fairly mechanical activity, which can often be performed by a computer. In a performance test, the instrument eliciting the learner's behaviour is a task or a prompt, and the assessment requires a set of criteria related to performance on the task as well as one or more raters who will interpret the criteria in relation to the assessed sample of performance. Thus, the process of assessment depends not only on the instrument but also on the criteria and the raters' application of these

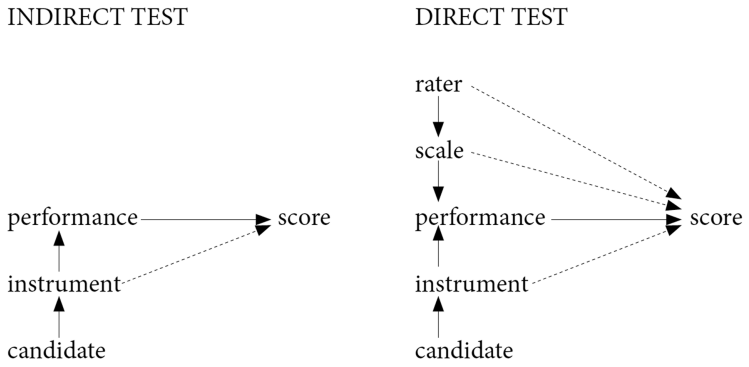

Figure 3.1 Elements of the Assessment Process in Indirect and Performance Tests
Source: Adapted from McNamara (1996, p. 9)

criteria. The difference between these two different types of assessment is illustrated in Figure 3.1 representing the indirect test (on the left), and the direct test (on the right). The different elements of the performance assessment process interact with one another and they jointly contribute to its outcome.

Evaluation of lexical proficiency is seldom a sole purpose of performance assessment. Vocabulary is usually an embedded construct in performance-based evaluation of L2 learners' more general writing or speaking skills or their global language proficiency. Thus, most of the observations concerning performance-based assessment of vocabulary come from a scrutiny of procedures applied in high-stakes language proficiency exams, or from research which focuses on the assessment of overall written or spoken performance. Only a few studies, which will be discussed in Section 3.7, examined how written production can be examined to assess specifically L2 lexical proficiency. The following sections will discuss in more detail the individual elements of the process of performance assessment.

3.3.1 Instrument

The instrument used in performance assessment is a task or a prompt that elicits from the learner an extended, continuous and structured text. In the simplest case, it can take the form of a question or a statement which the candidate should respond to in an expository essay. The topic usually touches on a problem which is relevant to the learner and which does not require background knowledge. The prompt can also be the first or the last sentence of a narrative essay or include non-verbal material, for

example a comic strip or a picture. Finally, writing tasks can also involve a response to written and/or aural stimulus text. Usually the instruction provides an expected length of the text in running words. Alternatively, the length of the learner's response is regulated by space provided on a sheet of paper designated for the candidate's text. Writing exams generally have a time limit, which is sufficient to let test-takers create an outline or even a draft of their text, if they wish to. Examinees are usually not allowed to use reference materials such as dictionaries or grammars.

Writing tasks tasks can be made more authentic by eliciting genres used in real life, outside the language classroom (e.g. a letter, a report or a review) and providing the candidate with information about the context and the audience of the text. It should be noted that some of the genres appearing in an exam are not very likely to be produced by the learner in real life (e.g. a review for a magazine), yet their aim is to simulate a situation in which writing a text is set in a particular sociocultural context, which makes the writing task more authentic as an act of communication. Thus, writing items within proficiency exams place varying degrees of emphasis on the authenticity of the task understood as nesting this task in a real-life situation with real-life target readers and requiring the production of a real-life genre. According to McNamara's classification, in performance assessment in the 'weak' sense, less attention is given to the sociocultural context of writing. However, if the exam gives more importance to the candidate's demonstration of successful communication—as in performance assessment in the 'strong' or 'stronger' sense—writing tasks include rich contextual information and real-life genres. This contrast between 'stronger' and 'weaker' writing items is best illustrated by two writing items featuring in the *TOEFL iBT* exam. Its integrated writing task simulates a real-life situation, frequent in academic settings, where the candidate is expected to write a paper demonstrating that he or she is familiar with, understands and can critically appraise concepts and facts encountered in their lectures and reading assignments. This task does not only involve writing skills, but writing ability has to be integrated with listening and reading comprehension and several non-linguistic (academic) skills, such as finding and understanding relationships between ideas and facts, forming a critical opinion about them and selecting them appropriately in order to support one's own argument. All these abilities and skills are the focus of evaluation. On the other hand, the *TOEFL*'s independent task is less authentic in terms of the reality outside the L2 classroom and taps more exclusively the ability to compose a structured, coherent and linguistically apt text.

One important point needs to be made about the direct writing (and speaking) tasks in the contexts of assessment of lexical proficiency. In high-stakes exams or research projects aiming at the assessment of L2 learners' vocabulary, the lexical characteristics of learners' texts are analysed and assessed from a holistic perspective as the totality of vocabulary

resources demonstrated by the learners. Unlike discrete-point items, such tasks are not constructed in order to elicit the use of specific words and expressions as the target of assessment.

3.3.2 Raters

The traditional and most common method of evaluating L2 learners' production has been human ratings. This approach has been applied to the assessment of the overall performance as well as the assessment of individual aspects of learners' output, including vocabulary; however, the latter are rarely rated alone and they usually form part of global evaluation. Such assessment involves a rater's judgement of the effectiveness and quality of a learner's overall performance and/or its different aspects in relation to the assigned task and the expected standard; and even though it may be founded on several specific criteria, the final decision is based on a perception rather than quantifiable evidence. That is why it is frequently referred to as **subjective assessment** (Alderson et al., 1995, p. 109).

The term *subjective* may seem pejorative in the context of evaluation, in particular if its results are used for high-stakes decisions or research. And indeed the evaluation of writing in the first and the second language carried out in the 19th and in the beginning of the 20th century truly deserved this epithet. Raters were left to their own devices when scoring test-takers' texts, and no care was taken to ensure the reliability and dependability of their judgements (Spolsky, 1995). However, the growing interest in psychometrics and objective language testing, which started in the first decades of 20th century and culminated in the 1950s and 1960s, left an important mark on performance assessment. It brought about the realisation that in order to warrant the reliability and dependability of assessment of learner's performance, it has to be based on explicit standards in the form of scales, descriptors and performance exemplars, which are available to and followed by all raters involved in marking a test. Thus, current principles of performance assessment require that raters do not rely on their own opinions and intuitions, but they are guided by very specific assessment criteria developed, adapted or adopted for a particular assessment situation. These criteria are contained in an assessment rubric and raters usually undergo rigorous training in the interpretation and application of the rubric to ensure a high level of inter- and intra-rater reliability, and in the case of high-stakes exams involving thousands of candidates, such as Cambridge English exams, *TOEFL* or *IELTS*, their performance as raters is monitored by supervisors throughout the whole period of marking. Finally, one exam script is usually evaluated independently by at least two raters, and the final result is the average of their scores. However, in the case of large discrepancies between them, a third rater is called in.

3.3.3 Scales

There are two kinds of rubrics used for the assessment of the L2 learner's production: holistic (global) and analytic scales. Holistic rubrics consist of one scale including several levels or bands, each containing a description of expected linguistic behaviour at this level in terms of manifold aspects of performance. The learner's performance at each growing rank is supposed to represent a higher standard, as manifested by a better quality of each of the aspects of their performance. A rater has to decide which descriptor an assessed piece of learner writing is closest to, and assign one holistic score to the text. Analytic rubrics, on the other hand, comprise several scales, each pertaining to one aspect of performance. Each of these scales includes descriptors for several levels of one individual aspect of performance. The result of the application of an analytic rubric is either a profile of several different scores relating to different aspects of writing or one mark which is a weighted combination of the scores for the separate components.

Each of these scales has its strong and weak points. **Holistic rubrics** are first of all more practical as they are easier and quicker to apply. White (1984) also claims that they are more valid as they resemble a reader's personal perception of—and authentic reaction to—a text. However, thier main weakness is the assumption that all aspects of writing grow in parallel. This implies that, for example, a better organisation of a text will be matched by more complex syntax, more sophisticated vocabulary and fewer errors, which rarely is the case in L2 learners. Thus, holistic scales mask the learner's particular strengths and weaknesses and are less useful for diagnostic purposes. Two learners can receive the same score for very different types of performance. Another weakness of holistic scales which has been observed is that the scores derived from them often correlate with more superficial features of writing such as handwriting or text length (Weigle, 2002, p. 114).

Analytic scales are just the opposite of global rubrics in terms of their advantages and disadvantages. They are more time consuming, as the rater has to evaluate the same text from several angles. Such assessment, based only on one component at a time, is also less natural and more demanding for a rater as it does not correspond to a real-life manner of reading and responding to a text. On the other hand, since each aspect of writing is evaluated independently on a separate scale, the assessment procedure results in a profile of scores reflecting better the learner's strong and weak points. At the same time, a set of several scores characterising the learner's writing ability may be difficult to interpret, and can make it difficult to make a decision about the candidate, for example for the purposes of selection and admission to a study programme, recruitment for a job or including a subject in a research project. Analytic scores scores can always be converted to a composite rating, but then the same problem of

identical grades awarded for very different types of performance occurs, as in the case of holistic assessment. Such composite scores, however, have been shown to be more reliable than holistic scores, as they are based on several components weighted independently (Hamp-Lyons, 1991a). Since analytic scales have a higher inter-rater reliability, and they are regarded as more valid, they are more frequently used in high-stakes language testing situations, in spite of the fact that they are more difficult and demanding to apply.

Weigle recapitulates McNamara's (1996) claims that

> the scale used in assessing performance tasks such as writing tests represents, implicitly or explicitly, the theoretical basis upon which the test is founded: that is it embodies the test (or scale) developer's notion of what skills or abilities are being measured by the test.
>
> (Weigle, 2002, p. 109)

Thus, a scrutiny of scales employed in assessment of direct writing tasks by second language acquisition researchers and language testing agencies can reveal how they perceive the role of vocabulary use in successful performance, and to what extent performance tests measure the learner's lexical proficiency and underlying lexical competence. In addition, rating scales demonstrate what specific criteria are used in assessment of vocabulary in writing production and disclose how the construct of lexical proficiency is conjectured and operationalised for the purpose of assessment.

3.4 Vocabulary in Writing Assessment Scales in Education

The specific criteria used in the assessment of performance in almost all direct writing tests, irrespective of the type of a scale, can be classified into two broad categories: (1) content and its delivery and (2) language use.[1] Individual scales vary in the number of specific criteria in each of these categories and their exact definitions. The most frequently used include: content, organisation, cohesion, vocabulary, language control and mechanics. Vocabulary use is almost always explicitly mentioned in rating rubrics either as a separate component or as part of a broader class referring to language use. Thus, an item requiring written production from candidates can be an example of an embedded vocabulary test according to Read's (2000) taxonomy discussed in Chapter 2. The number of subscales included in an analytic rating rubric depends on the purpose of a test. If the focus of assessment is the acquisition of specific language elements and skills, for example in general foreign language instruction, then the scale can contain two or three subscales referring to different aspects of language use, including vocabulary. If, on the other hand, a writing

test is to certify the learner's ability to write texts in a foreign language for academic or professional purposes, then more emphasis will be placed on content and its delivery, with more subscales for various components of this category, and fewer for specific language features (Weigle, 2002, p. 123).

Both holistic and analytic rubrics include scaled descriptors which specify the features of candidates' typical performance at each rating level. The descriptors may either be rooted in the theoretical consideration of the assessed construct, or be based on empirical examination of operationally rated samples of performance and their textual qualities. The theoretical approach is advocated by Bachman and Palmer (1996) as being more adequate for making inferences about the learner's writing and language ability, yet at the same time it can contain rather vague distinctions referring to levels of mastery of these abilities such as 'adequate', 'skilful' or 'weak', which are hard to interpret and necessitate extensive rater training (Weigle, 2002, p. 123). The empirical approach, on the other hand, is more suitable for predicting future performance on comparable tasks. It contains reference to these characteristics of the performance which have been shown to discriminate well between texts at different levels. The choice of the type of descriptors depends mainly on the purpose of assessment, but the two approaches can also be mixed and applied in one rating scale.

3.4.1 Holistic Scales

A prime example of holistic scales used to evaluate a learner's performance are the rubrics used in the *TOEFL* exam. The exam in fact contains two different holistic scales developed for the two writing items featuring in the exam: the integrated and the independent tasks. Both scales include six bands (0–5), and each band contains a descriptor incorporating several assessment criteria, which jointly form a profile of a text at this level. These criteria include content, organisation, coherence and language use for both the independent and integrated tasks; however, not all of them are mentioned in the descriptors for separate levels. In addition, the individual aspects of writing do not have the same importance in the two scales. In the case of the integrated task more prominence is given to the content and less to the remaining aspects of writing, language use in particular. In the rubric for the independent task, the role of these criteria is more balanced. An example of the descriptors and criteria present in the two rubrics is given in Figure 3.2, which contains specifications for Band 3.

The independent rubric includes explicit reference to vocabulary in the descriptors of performance at different levels. The aspects of vocabulary use which are taken into account include: word choice appropriate to convey meaning, range and accuracy. It is worth noting that the criteria

Score	*TOEFL Scoring Guide (Independent)*	*TOEFL Scoring Guide (Integrated)*
3	An essay at this level is marked by one or more of the following: • Addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • Displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • May display accurate, but limited range of syntactic structures and vocabulary	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: • Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. • The response may omit one major key point made in the lecture. • Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. • Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.

Figure 3.2 An Excerpt From the *TOEFL* Independent and Integrated Writing Rubrics[2]

Score	*Vocabulary (and Structure)*
5	Displays consistent facility in the use of language, demonstrating syntactic variety, *appropriate word choice and idiomaticity*, though it may have *minor lexical* or grammatical *errors*
4	Displays facility in the use of language, demonstrating syntactic variety and *range of vocabulary*, though it will probably have *occasional noticeable minor errors in* structure, *word form or use of idiomatic language* that do not interfere with meaning
3	May demonstrate *inconsistent facility in* sentence formation and *word choice* that may result in lack of clarity and occasionally obscure meaning May display *accurate but limited range of* syntactic structures and *vocabulary*
2	A noticeably *inappropriate choice of words or word forms* An accumulation of errors in sentence structure and/or usage
1	Serious and frequent errors in sentence structure or usage
0	(The descriptor does not make reference to vocabulary use)

Figure 3.3 Descriptors Related to Vocabulary Use in the *TOEFL* Independent Writing Rubrics

concern both individual words and multi-word units. Figure 3.3 includes all descriptor fragments which relate to lexis.

The integrated rubric does not make explicit reference to lexis, yet vocabulary is present together with grammar in the more general criterion of language use. The aspect of language use that is mentioned consistently through the scale is accuracy; other criteria relating to appropriate selection and range are not included in the descriptors.

In recent years automatic scoring of *TOEFL* writing tasks has been introduced to assist human raters. An automated scoring technology called *e-Rater*® is used along human ratings to score both the independent and integrated writing tasks. Using Natural Language Processing techniques, the system automatically computes over 60 indices which relate to various features of learners' production such as the number of discourse elements or of pronoun errors. In the more recent version of the programme (version 11.1), the indices were grouped into 11 scoring features which relate to specific characteristics of the essay. Nine of these features assess general writing quality, while the other two address the content of the essay and are modelled for specific prompts. One of the general writing features, lexical complexity, refers to the learner's use of vocabulary and consists of two indices: the average word length and sophistication of word choice (both of these indices will be discussed in detail in Chapter 4). In addition, the prompt-specific features evaluate the learner's use of topic-related vocabulary (Quinlan, Higgins, & Wolff, 2009; Ramineni et al., 2012). According to ETS "[u]sing both human judgment for content and meaning with automated scoring for linguistic features ensures consistent, quality scores" (Understanding Your TOEFL iBT® Test Scores, n.d.[3]). However, the official documents related to the exam do not make it clear how the automatic scoring and human judgement interact to produce the final grade.

3.4.2 Analytic Scales

Analytic scales produce a more meaningful score for the performance-based assessment of L2 lexical proficiency because vocabulary use is evaluated on a separate scale or at least it forms part of a more focused component of language use. One of the most frequently quoted examples of an analytic rubric, which was used widely in high-stakes ESL exams in the United States, is the scale developed by Jacobs, Zingraf, Wormuth, Hartfiel and Hughey (1981). It consists of five components: content, organisation, vocabulary, language use and mechanics. Each of these elements is divided into four bands including scaled descriptors and a number of points that can be awarded within this level to a composition. The scores for each component can be combined into a composite score of the maximum of 100 points, but the weight of each component in the final result is not equal. The maximum score for content is 30 points

and it contributes most to the total; language use (referring exclusively to grammar) can receive the maximum of 25 points, and the component of organisation and vocabulary contribute 20 points each to the final score. Mechanics are given marginal importance with the maximum of 5 points. Figure 3.4 reproduces Jacobs et al.'s rubric for the component of vocabulary.

The descriptors used in the scale refer to the following aspects of vocabulary use: range, word choice, word usage, word form, register and meaning. Table 3.1 presents how these aspects are described at each level.

The scale can be related to the model of language proficiency proposed by Housen and Kuiken (2009), which postulates three components of proficiency: complexity, accuracy and fluency. Complexity is represented in the scale by the criterion of the range of vocabulary. The wider and more sophisticated vocabulary of a text, the more linguistically complex it appears to be. Accuracy is tapped by the number of lexical errors in word choice, usage and form and by the effect they have on the reader,

VOCABULARY	20–18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
	17–14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*
	13–10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*
	9–7	VERY POOR: essential translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate

Figure 3.4 The Analytic Rubric for the Component of Vocabulary

Source: Jacobs et al. (1981)

Table 3.1 The Criteria for the Assessment of Vocabulary use in Jacobs et al.'s (1981) Analytic Rubric

Criterion	*EXCELLENT TO VERY GOOD*	*GOOD TO AVERAGE*	*FAIR TO POOR*	*VERY POOR*
range	sophisticated	adequate	limited	little knowledge of English vocabulary, idioms, word forms
word/idiom choice	effective	occasional errors	frequent errors	
word/idiom usage	effective	occasional errors	frequent errors	
word/idiom form	mastery	occasional errors	frequent errors	
register	appropriate			
meaning		*meaning not obscured*	*meaning confused or obscured*	

that is the degree to which they obscure the meaning of the message conveyed in the text. Fluency, however, is not addressed either in the descriptors for vocabulary, or the descriptors for other components in the rubric. At the same time, the scale implicitly links with the model of lexical competence, discussed in Chapter 1. The criterion of range taps the size of the lexicon under the assumption that the larger the vocabulary that learners have at their disposal, the larger range of its use they display in their writing. The depth, which is not the quality of the lexicon as a whole but of its individual items (Meara, 1996a), is represented by word choice, word form and word usage. Word form indicates a familiarity with the formal aspects of words used by learners in their writing, such as inflectional and derivational morphemes. It needs to be noted, however, that spelling, which also relates to the formal aspect of word knowledge, is singled out in the descriptors for the component of mechanics, rather than vocabulary, in Jacobs et al.'s scale. Word choice is the most comprehensive aspect of vocabulary use featuring in the rubric. On the one hand, it gauges the meaning component of word knowledge, that is if right words were chosen to express the intended meaning. On the other hand, it taps the contextual restrictions on word use, that is if appropriate words were chosen for the task or topic. Appropriacy for the register also relates to the criterion of word use. The scale implicitly assumes that the fewer errors learners make in different categories, the deeper their L2 lexical knowledge can be. The scale does not address the access/automaticity dimension of the L2 lexicon. On the other hand, the reference to idioms in the descriptors shows that the scale assesses the lexicon in its broader sense, including multi-word expressions.

A similar range of criteria for the evaluation of learners' use of L2 vocabulary in their written production was adopted by the exams developed and administered by Cambridge English Language Assessment. They are widely popular exams of English as a Foreign Language, in particular in Europe but also further afield, in Asia and Africa. The exams, already mentioned in the previous chapter, use an analytic rubric for the evaluation of their written sections by trained raters. The rubric was developed with explicit reference to the Common European Framework of Reference for Languages and it is applied across all the spectrum of Cambridge English exams. The overall writing scale consists of four subscales: content, communicative achievement, organisation and language, each assessed separately. The last component is defined in the following way in the exam handbooks: "Language focuses on vocabulary and grammar. This includes the range of language as well as how accurate it is" (e.g. CPE Handbook, p. 26). This means that vocabulary is not evaluated on a separate scale, but it is collapsed with grammar in the overall category referring to language use. Its descriptors form a grid showing progression from A2 to the firm C2 level. The descriptors for level B2 for the component of Language are reproduced in Figure 3.5.

CEFR level	Language
B2	Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good degree of control. Errors do not impede communication.

Figure 3.5 Descriptors for the Component of Language at the B2 Level
Source: FCE Handbook (p. 33)[4]

However, for each of the exams a more specific rubric is designed, which guides raters to grade learners' production into bands from 0 to 5. The descriptor relevant to the CEFR level targeted by the exam is placed in the middle of the scale and assigned 3 points. The descriptors for the levels at both ends of the scale are taken from the adjacent CEFR levels and assigned points 5 and 1. The intermediate levels 4 and 2 do not have performance descriptors.

The aspects of vocabulary use which are explicitly indicated to in the descriptors include: range, frequency, appropriacy and errors, but not all these criteria are referred to at each level. Table 3.2 demonstrates how these criteria are tackled in the descriptors for each level. The columns on the right present the number of points corresponding to each descriptor in individual Cambridge English exams.

The handbooks for individual exams also give precise definitions of the criteria used in the descriptors. The definition relevant to the criterion of appropriacy of vocabulary is copied in Figure 3.6.

As in the case of Jacobs et al. (1981), an attempt can be made to link the subscale and the vocabulary-related descriptors to Housen and Kuiken's (2009) model of language proficiency. The Cambridge English scale uses two criteria that gauge the concept of complexity: range and frequency, the latter tapping the use of sophisticated (and thus infrequent) words. The dimension of accuracy is addressed by the criteria of appropriacy[6] and errors. The dimension of fluency is also not addressed in the scale except for one descriptor at the highest level of performance. Moreover, it is not explained either in the glossary or the descriptor itself what exactly is understood by this term.

The descriptors can also be linked to the model of lexical competence; however in this instance this is not as easy and clear-cut as in the case of Jacobs et al.'s scale. The dimension of breadth is represented by range and frequency, again under the assumption that the use of a wider range of vocabulary and of sophisticated, infrequent words is a reflection of a larger lexicon. It is noteworthy that the criterion of range does not apply solely to individual words but also multi-word expressions. The depth of

Table 3.2 The Criteria for the Assessment of Vocabulary Use in Cambridge English Exams' Assessment Rubrics

	RANGE	*FREQUENCY*	*APPROPRIACY*	*OTHER*	*ERRORS*	*CPE*	*CAE*	*FCE*	*PET*
	Uses a wide range of vocabulary,	including less common lexis,	(with precision and style)	with fluency.	Any inaccuracies occur only as slips.	5			
						4			
C2	Uses a range of vocabulary,	including less common lexis,	precisely and	effectively.	Errors, if present, are related to less common words, or occur as slips.	3	5		
						2	4		
C1	Uses a range of vocabulary,	including less common lexis,	appropriately.		Occasional errors may be present but do not impede communication.	1	3	5	
						0	2	4	
B2	Uses a range of everyday vocabulary	(with occasional inappropriate use of less common lexis)	appropriately, with occasional inappropriate use of less common lexis.		Errors do not impede communication.		1	3	5
							0	2	4
B1	Uses everyday vocabulary, (while occasionally overusing certain lexis)		generally appropriately.		While errors are noticeable, meaning can still be determined.			1	3
								0	2
A2	Uses basic vocabulary		reasonably appropriately.		Errors may impede meaning at times.				1
									0

Appropriacy of vocabulary: the use of words and phrases that fit the context of the given task. For example, in I'm very sensible to noise, the word sensible is inappropriate as the word should be sensitive. Another example would be Today's big snow makes getting around the city difficult. The phrase getting around is well suited to this situation. However, big snow is inappropriate as big and snow are not used together. Heavy snow would be appropriate.

Figure 3.6 The Definition of the Criterion of "Appropriacy of Vocabulary" Used in the Descriptors

Source: CPE Handbook (p. 26)[5]

No. of Points	Comment
FCE	
2	There is a range of everyday, topic-specific vocabulary, which is used appropriately (*creates new types of clothes; Some people claim; extremely high; is more important than*). Simple grammatical forms are used with a good degree of control, although the use of verbs in the 3rd person is not consistent. There are attempts to express ideas using a range of grammatical forms, passives and modals for example, but these are less successful (*people, who can't afford it, should not be in the society; the fashion industry guide the people to be in a good appearance; It's something which was created to help people what to wear*). Errors are noticeable but meaning can still be determined.

Figure 3.7 Examples of Raters' Comments for the Component of Language

Source: FCE Handbook (p. 43)

vocabulary is tapped by appropriacy and errors, yet the descriptors do not make it specific which aspects of word knowledge these two criteria can pertain to. The definition in the glossary gives examples of meaning and collocation in reference to inappropriate language use, but it is implied that this category can also apply to other components of word knowledge.

In addition to the rubric and scaled descriptors, the exam documentation also lists examples of several exams scripts together with their analytic scores assigned by trained raters and their justification of the scores. They show how raters make use of the scale and its descriptors. Most of the comments include evaluation of vocabulary use. An example of such a comment is presented in Figure 3.7.

Although the overall scale is analytic, each subscale includes several different rating criteria (e.g. range, errors, appropriacy) combined in a

(CPE) A wide range of vocabulary, including less common lexis, is used effectively and with fluency (*'no-frills' airlines, de-humanising, perched on the mountain sides, social interaction, engage in long conversations, strolling*). [C2+] However, there are examples of incorrectly chosen words (*ache, overweight*) and in a few places vocabulary is repetitive (*views*). Occasional errors do not impede communication. [B2]

(FCE) A range of everyday vocabulary is used appropriately, [B2] and although there are some errors (*fasilities; all senses' gratification stuff*) [B2] there is also some good use of less common lexis (*started his spiritual journey*). [C1]

Figure 3.8 Examples of Raters' Comments Related to Vocabulary Use

Source: CPE Handbook (p. 39) and FCE Handbook (p. 37), respectively; levels inserted by the author

single set of scaled descriptors. The assessment of vocabulary includes several aspects of use analysed in Table 3.2. Since they are assessed on a single scale, these different aspects are assumed to develop in parallel. Yet, this is not always the case, as evidenced by raters' comments listed in Figure 3.8.

Obviously, it is unrealistic to expect that an analytic scale could include separate subscales for every aspect of every criterion employed in the assessment. First of all, such a scoring procedure would be highly impractical as its application would require a lot of time and effort. In addition, a large number of subscales can negatively influence the reliability of assessment, as it may be difficult to define each assessment criterion in separation from the others (Hamp-Lyons, 1991b). Finally, a learner's knowledge of various aspects of lexis is not a goal of the assessment in writing tests which aim to evaluate a learner's general writing skill. Yet, this shows that the rating rubrics used in the assessment of writing are not satisfactory instruments to evaluate specifically a learner's lexical proficiency.

One more specification concerning the assessment of writing production in Cambridge English exams is worth mentioning. The test specifications laid out in the exam handbooks assert that although each writing task contains a guideline on expected length (expressed in an approximate number of running words in a produced text), test-takers are not penalised for writing shorter or longer texts. However, exam developers caution that over-length and under-length scripts are likely to contain other problems affecting the score such as irrelevant information, poor organisation or inadequate range of language.

Another take on specific criteria used in the evaluation of vocabulary use is presented in the analytic scale developed by Weir (1990) for the assessment of writing in another high-stakes exam: *Test of English for Educational Purposes*, offered by University of Reading. His rubric

D. Adequacy of vocabulary for purpose.
0. Vocabulary inadequate even for the most basic parts of the intended communication.
1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

Figure 3.9 The Analytic Rubric for the Component of Vocabulary From the Test of English for Educational Purposes

Source: Weir (1990)

contains altogether seven different subscales, including one devoted to vocabulary, which is reproduced in Figure 3.9. Its heading and its scaled descriptors demonstrate that the main criterion in the evaluation of a learner's use of vocabulary in this test is its adequacy for the task. If this criterion is linked to the theoretical underpinnings discussed in Chapter 1, it becomes evident that the scale taps into the strategic competence related to vocabulary ability (Chapelle, 1994) rather than the narrow construct of lexical competence. Learners are judged not on how many different words they employed in their texts and how sophisticated these items are, but if they could choose from their lexical resources the ones which are sufficient and suitable for the task. The assessment of ability rather than knowledge becomes the pivot of this subscale. The other aspect of vocabulary use mentioned in the descriptors relates to accuracy and thus taps the depth of lexical knowledge.

3.5 Analytic Scales for the Assessment of Vocabulary in Education

As shown in the previous section, the analytic scales used for the assessment of writing in various high-stakes proficiency exams include vocabulary as one of the components of evaluation. The scaled descriptors referring to vocabulary mention several aspects as vocabulary use; however, all these lexical features are rated jointly on a single scale which blurs learners' varying profiles as far as their lexical proficiency is concerned. While this is not detrimental to the assessment of the writing skill in a foreign language or of general language proficiency, such an approach may be insufficient for more focused performance-based evaluation of the learner's lexical proficiency or in research on second language vocabulary acquisition. Therefore, in some situations vocabulary alone

needs to be evaluated analytically on several subscales tapping different aspects of lexical proficiency.

Such analytic scales referring to vocabulary are proposed by the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR; Council of Europe, 2001). The document contains altogether 56 scales referring to language knowledge and use. These scales are not meant to be an analytic assessment rubric, but they offer scaled descriptors of various aspects of learners' communicative and linguistic behaviour at different proficiency levels, and they can form a base for the development of assessment procedures for various situations. The Overall Language Proficiency is divided into three broad categories: Communicative Activities, Communicative Strategies and Communicative Language Competences. Linguistic competences, one of three constituents of Communicative Language Competences, are described along two dimensions: range and control, and these two criteria are also applied to the description of vocabulary. Figures 3.10 and 3.11 reproduce the two vocabulary scales.

The CEFR levels as well as its various illustrative scales have gained a wide popularity around the world. Language courses, teaching materials and examinations are set against its standards. Yet, they have also attracted some criticism from specialists in language testing and second language acquisition concerning the validity of the scale system. The

VOCABULARY RANGE	
C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.
C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.
B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his everyday life such as family, hobbies and interests, work, travel, and current events.
A2	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.
	Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.
A1	Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.

Figure 3.10 The CEFR Rubric for Vocabulary Range

Source: Council of Europe (2001, p. 112)

VOCABULARY CONTROL	
C2	Consistently correct and appropriate use of vocabulary.
C1	Occasional minor slips, but no significant vocabulary errors.
B2	Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.
B1	Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.
A2	Can control a narrow repertoire dealing with concrete everyday needs.
A1	No descriptor available.

Figure 3.11 The CEFR Rubric for Vocabulary Control
Source: Council of Europe (2001, p. 112)

CEFR rubrics were created by pooling together a large number of descriptors from scales used in various high-stakes exams all over the world and subjecting them to a careful scrutiny and revision. (For more detailed information on the procedure see Council of Europe, 2001, Appendix A and B; Hulstijn, Alderson, & Schoonen, 2010; Wisniewski, 2013.) Unlike the rubrics used in the high-stake examinations discussed in the previous section, which constantly undergo a rigorous validation procedure against empirical data,[7] the CEFR scales have not been empirically validated against the reality of learner language (Alderson, 2007; Hulstijn, 2007; Wisniewski, 2013). The critics also point out that the scales were not derived from theoretical considerations and were not linked to a particular model of language proficiency and second language acquisition. That is why, in spite of their wide popularity, the CEFR rubrics have to be treated with caution.

The shortcomings of the CEFR scales are also visible in the two rubrics referring to vocabulary. Admittedly, the two scales can be linked to two dimensions of language proficiency—complexity (range) and accuracy (control). Yet, the individual features of vocabulary knowledge and use mentioned in the scaled descriptors cannot be consistently matched with the model of lexical competence and vocabulary ability. The Vocabulary Range scale mentions connotative levels of meaning at the C2 level, but does not address this and other aspects of word meaning in the remaining bands. Only the descriptors for levels C1, B2 and B1 refer to strategies of using vocabulary and it is particularly interesting to see the inconsistency in the treatment of circumlocution across the bands. It is presented as a positive strategy at the C1 level, but it is presented as a tactic revealing gaps in vocabulary at the B2 and B1 levels. One variable in the descriptors does not address range, but taps the fluency component of language proficiency as well as the access/automaticity dimension of lexical

competence: "little obvious searching for expressions" (C1), "lexical gaps can still cause hesitation" (B2), and this aspect is referred to only at two levels: C1 and B2. The Vocabulary Control descriptors include reference solely to correct uses of vocabulary at the C2 and A2 levels, solely to errors at the C1 level and both to correct and incorrect uses in the B1 and B2 bands. The descriptors contain no information on the types of lexical errors except for word choice errors (referring to the meaning aspects of word knowledge), which are mentioned only at the B2 level: "incorrect word choice does occur without hindering communication" (B2). Levels B1 and A2 address simultaneously accuracy and range of vocabulary use.

The problems with the CEFR scales referring to vocabulary can render them invalid and unreliable instruments for tracing the development of learners' lexical proficiency through performance-based assessment. L2 learners' vocabulary use profiles do not reflect various components of lexical proficiency and thus present a somewhat imprecise picture of learners' lexical competence and skills. The limited adequacy of the scale for the assessment of learners' vocabulary was also demonstrated in empirical research validating the scale against learner data (Wisniewski, 2017).

3.6 Vocabulary Assessment Scales for Research Purposes

All the rubrics discussed so far were developed for the assessment of L2 learners' performance and linguistic abilities for real-life educational purposes such as certification or admission to study programmes. The CEFR framework has an even larger range of practical applications including curriculum development or materials writing. The vocabulary rating scales discussed in this section were created with a different agenda. They were devised as part of research related to performance-based assessment of vocabulary and the primary aim for their development was validation of various statistical measures of lexical quality of texts proposed in literature. A host of such automatic gauges—discussed in detail in Chapter 4 of this book—have recently been explored in studies of automated scoring of texts as well as in research on second language vocabulary acquisition. However, the validity of the statistical measures has been usually taken for granted and few attempts have been made to link them to an adequate definition of the construct of lexical proficiency as reflected in writing, and to the construct of lexical competence. Jarvis (2013) expresses this deficiency in the following words:

> More worrisome, however, are measures that have been developed prior to or in the absence of an adequate theoretical construct definition, as well as measures that are used in ways that are incompatible with or reflect a poor understanding of the construct definition (assuming that there is a construct definition in the first place).
>
> (Jarvis, 2013, p. 14)

Three studies described next address this concern. They undertake to propose definitions of relevant lexical constructs, describe them in terms of observable characteristics and then juxtapose human ratings of these features with their statistical measures. In view of the complexity of these projects, their results will be discussed in various sections of this book. The construct definitions and assessment rubrics will be described later in this section, research on validating one of these scales against human perceptions will be presented in Section 3.7.2 of this chapter and the comparison of human ratings with the automated scores will be discussed in Chapter 5 (Section 5.3.3).

Crossley et al. (2011a) developed a lexical proficiency assessment rubric which was to guide raters in holistic evaluation focusing solely on lexical aspects of L2 essays. The rubric was adapted from the scoring instruments published by two national organisations dealing with foreign language teaching and testing in the US and had been previously employed in high-stake examinations. For the new lexical proficiency assessment rubric, the scores and corresponding descriptions of lexical aspects of language proficiency from the existing scales were combined and revised. The resulting set of evaluating criteria defined lexical proficiency as "skillful language use that is accurate and fluent and is characterized by the appropriate use of conceptual categories, coherence, and lexical-semantic connections" (Crossley et al., 2011a, p. 569). The relative mastery of these lexical competences was divided into five proficiency bands (1–5) with relevant descriptors available for each band. An excerpt for the scale is reproduced in Figure 3.12.

The originality of this scale lies not only in the fact that it focuses solely on the use of vocabulary in L2 written production, but also in its introduction of new criteria of assessment, which have not so far been used in assessment rubrics: conceptual categories and lexical-semantic connections. Unfortunately, these new criteria are not adequately defined either in the rubric itself or in the paper presenting it. The scale's descriptors also include the lexical features which frequently appear in rating rubrics such as word choice, accuracy or coherence. The criterion of

> **Score of 3:** A sample in this category demonstrates adequate lexical mastery, although it will have lapses in quality. The sample demonstrates some appropriate uses of conceptual categories (including abstract concepts, but mostly concrete concepts), coherence between words, lexical-semantic connections, and lexical diversity. Overall, the sample uses generally appropriate and precise vocabulary, but demonstrates an inconsistent mastery.

Figure 3.12 An Excerpt From a Holistic Rubric for the Assessment of Lexical Proficiency

Source: Crossley et al. (2011a, p. 569)

Table 3.3 The Criteria in Crossley et al.'s (2011a) Lexical Proficiency Assessment Rubric

Criterion/band	*5*	*4*	*3*	*2*	*1*
mastery	clear and consistent	reasonably consistent	adequate; inconsistent	developing	little
skill	skilful use of vocabulary with ease and fluency	fluent			very little facility in the use of language
accuracy	few minor errors; accurate vocabulary	occasional errors or lapses in lexical quality; accurate	lapses in quality	lexical problems serious enough that the meaning is somewhat obscured	frequent lexical problems so serious that the meaning is often obscured
conceptual categories (both concrete and abstract)	effective use of appropriate conceptual categories; apt vocabulary	appropriate use of conceptual categories, appropriate and precise vocabulary	some appropriate uses (including abstract concepts but mostly concrete concepts); generally appropriate and precise vocabulary	weak (?) vocabulary and inappropriate word choice; depends on concrete words	incorrect word choices
coherence between words	clear	+	+	discourse generally connected	very little coherence
lexical-semantic connections	+	+	+	ONE OR MORE problems	TWO OR MORE problems
diversity (range)	+ (enough to explain complex matters in detail); varied vocabulary	+ allows for the discussion of complex matters but not always	+	lacks variety	limited vocabulary; discourse relies heavily on the use and repetition of memorized phrases

lexical diversity is a different term referring to lexical range. The criteria used in the rubric and their descriptors are analysed in detail in Table 3.3.

The scrutiny of the scale reveals its many weaknesses which reflect the lack of theoretical grounding and a rather pragmatic approach to its creation. The new and original criteria are not only undefined, but they lack adequate descriptors. This is especially visible in the case of lexical-semantic connections, which, according to the rubric, can either be present in a rated text or there may be some problem related to them. The criterion of conceptual categories assumes that concrete concepts are mastered before abstract ones, which assumption has little grounding in research on second language acquisition. Otherwise, this category seems to refer to appropriate use of vocabulary and correct and incorrect word choices, an aspect which is frequently included in rating scales. The fact that the scale was assembled from other instruments can be clear in the repetitiveness of the descriptors. Levels 5 and 4 include reference to accuracy at the beginning and the end of the descriptor. The same problem concerns the criterion of lexical diversity in the descriptor for Level 1. Particularly confusing is the reference to appropriate use of conceptual categories and of vocabulary in general, which appear in the descriptor for Level 4.

In a follow-up study by Crossley, Salsbury and McNamara (2013), this holistic rubric was juxtaposed with a corresponding analytic scale including eight components: basic category use, word specificity and abstractness, word concreteness, semantic co-referentiality, collocational accuracy, sense relations, word sense frequency, word frequency and type/token ratio. These components were assumed to represent four subcategories of lexical proficiency: conceptual knowledge, lexical associations, lexical frequency and lexical diversity. Unfortunately, the researchers do not motivate their choice of analytic components beyond the fact that these are "lexical features of theoretical interest in lexical proficiency research" (p. 110) and that human judgements of these analytic lexical features represent operationalisations of lexical constructs measured by indices evaluated in their study. Each of the components was briefly described in the rubric and it was to be evaluated on a six-point scale with specific descriptors for each band unavailable. Fragments of the scale are reproduced in Figure 3.13.

The two rubrics were employed by three trained raters for the evaluation of 240 texts written by native speakers and learners of English studying at an American university. The holistic and analytic ratings were juxtaposed with several statistical measures of lexical quality computed for each text. The details of these analyses and their outcome are discussed in Section 5.3.3. One result of this study, which is relevant for the current discussion is the inter-rater correlations for each assessment component which are reproduced in Table 3.4. The weighted Pearson correlation indicates that the reliability of the ratings for different categories

. . .
2. Lexical Associations
2.1 *Semantic Co-referentiality* The words in the sample are semantically related to one another, but not necessarily related morphologically. For instance, *cat* and *mouse* are more semantically related than *dog* and *mouse*.
. . .
2.3 *Sense Relations* The words in the sample could have multiple senses making the meaning more ambiguous. Senses refer to the number of meanings a word can have. For instance, the word *class* has at least six related senses (socio-economic class, a body of students, a course of study, a collection of things sharing similar attributes, a sports league ranked by quality, and elegance in dress or behavior) as compared to the word *grape* which has one sense.
. . .

Figure 3.13 Excerpts From an Analytic Rubric for the Assessment of Lexical Proficiency

Source: Crossley, Salsubury and McNamara (2013, p. 132)

Table 3.4 Weighted Pearson Correlations Between Raters

Item	*r*
Basic categories	.486
Word specificity	.542
Word concreteness	.771
Semantic co-referentiality	.733
Collocations	.924
Sense relations	.627
Sense frequency	.657
Word frequency	.769
Lexical diversity	.825
Holistic score	.921

Source: Crossley, Salsubury and McNamara (2013, p. 114)

varied considerably. In spite of the training, raters agreed on some categories (collocation and lexical diversity) and on the holistic judgement of the lexical proficiency, but showed a considerable disagreement in the evaluation of some other features (basic categories or word specificity).

The two studies represent an important trend in research on performance-based assessment of vocabulary. This research calls for defining the underlying constructs in terms of observable, objective and quantifiable phenomena which could be shown to discriminate between

learners representing different levels of ability. Such concrete evidence would provide empirical validation of the definitions, scales and descriptors used for the evaluation of vocabulary. Yet, the individual assessment criteria proposed by Crossley et al. (2011a) and Crossley, Salsubury and McNamara (2013), both in their holistic scale and especially in their analytic lexical proficiency assessment rubric, are questionable. In spite of the researchers' claim of their theoretical relevance, their selection for the analysis was in fact motivated not so much by theoretical considerations related to lexical proficiency or lexical competence, but the need to match available lexical measures. Some of the assessment components, like word frequency, were fairly self-explanatory, some others lacked both construct and ecological validity, which was evidenced by low correlations between the raters evaluating these features. For example, the criterion of *word specificity*, produced the weighted Pearson correlation coefficient of 0.54 among the raters. It can be argued that while specificity is an important characteristic of individual words, reflecting indirectly both size and depth of the mental lexicon, it is doubtful if raters take it into account in a straightforward way when evaluating the vocabulary of a text. It seems more likely that raters judge if the words used by the learner are specific enough or abstract enough in a given context rather than in general and objective terms, as indicated by Chapelle (1994) in her framework. The occurrence of both highly specific words and very general abstract words in an L2 learner's production, i.e. words which are hyponyms or hyperonyms of basic-level categories, are a proof of an advanced lexical proficiency.

Another attempt at defining the assessed lexical construct, distinguishing its observable and quantifiable characteristics and validating them against human judgements and automatic measures was made by Jarvis (2013), yet he focuses only on one component of lexical proficiency: lexical diversity. Lexical diversity is another term for vocabulary range and this aspect is frequently mentioned as an important component of vocabulary use. It is also usually included in performance assessment rubrics.

Jarvis begins his scrutiny by introducing two notions: repetition and redundancy as concepts which are similar, by referring to the same phenomenon, but at the same time very different, by taking a different perspective at the same phenomenon. Jarvis defines the two notions in the following way:

> Critically, repetition in its purest sense is an objective phenomenon, whereas redundancy is fundamentally subjective—not in the sense of being a matter of personal taste and thus varying from one individual to the next, but in the sense of being grounded in human perception. At its most basic level, redundancy involves the perception of excessive or unnecessary repetition.
>
> (Jarvis, 2013, p. 20)

In other words, repetition is a naturally occurring and neutral phenomenon whereas redundancy is an evaluative quality of writing. Some degree of repetition is necessary for purely grammatical and pragmatic reasons. Too much of repetition, however, leads to redundancy and this is a detrimental quality of a text, which has its roots in human tacit consensus and implicit rules about how much repetition is desirable and acceptable and how much of it is excessive. Since these rules are implicit, they are grounded in human perception and they can be influenced by a variety of factors.

Jarvis defines lexical diversity as the opposite of redundancy. This implies that lexical diversity is not an objective phenomenon defined by a mere rate of using many different words in a text, but it is a subjective and evaluative quality of a text which is determined by human perception. Yet, according to Jarvis, defining lexical diversity as a phenomenon related to human sensitivity and experience does not preclude its independent measurement. He explains it in the following way:

> The notion that redundancy [and by extension lexical diversity] is a subjective (perception-based) construct does not necessarily mean that it cannot be measured objectively, but it does mean that it cannot be measured accurately through objective means until the researcher fully understands all of the factors that affect the way it is perceived, and not until the researcher also understands how to apply proper weights to each of those factors.
>
> (Jarvis, 2013, p. 20)

Following this line of reasoning, Jarvis proposes six observable properties of a text which together form the construct of lexical diversity: variability, volume, evenness, rarity, dispersion and disparity. Next, each of these dimensions of lexical diversity is defined briefly.

Variability is the opposite of word repetition and it is the rate with which a text makes use of new words instead of repeating the ones used before. Volume denotes text length. This dimension refers to the widely known and naturally occurring fact that the same ratio of repetition of words is likely to be perceived differently in a short or long text. Evenness denotes the ratio of repetition of individual lexical items. The perception of redundancy is likely to depend on whether the spread of recurrence is fairly even across individual items. Rarity refers to the use of less common and less frequent words in a text. Texts which employ more sophisticated vocabulary may be perceived as being more lexically diverse. Including this property as a component of lexical diversity is a very new approach. As will become evident in Chapter 4 and 5, lexical sophistication has always been defined as a complementary but separate quality of lexical richness. Dispersion refers to the proximity of repeated words in a text. The closer they are together, the more likely the text is likely to

be perceived as redundant. Disparity involves the degree of similarity or differentiation between words in a text. This property can operate on the formal and semantic levels. Two words can be different types but they can be either derivationally very close (e.g. perceive and perception) or semantically related (e.g. perceive and see). Both formal and semantic differentiation of lexical items in a text may contribute to the perception of text redundancy.

The six observable components distinguished by Jarvis did not prompt him to create an analytic rating scale. Nevertheless, the researcher investigated their contribution to the raters' perception of lexical diversity in a text by manipulating the relevant text characteristics and recording judges' reactions. The details of this study and its results are discussed in Section 3.7.2.

3.7 Extraneous Variables in the Assessment Process

As discussed in the previous chapter, the ultimate aim of any assessment, including performance-based assessment, is to draw inferences about the learner's ability, however it is defined. A score, which is the product of the assessment process, is taken to reflect the candidate's level of the targeted proficiency. Yet, it has long been acknowledged by assessment specialists that a score on any test is also influenced by other factors than the test-taker's ability. The extraneous variables in the assessment process can be divided into three categories: test method facets, personal attributes and random factors (Bachman, 1990, p. 165). Bachman states that since a test's aim is to measure the learner's ability as reliably as possible, "[a] major concern in the design and development of language tests, . . . , is to minimize the effects of test method, personal attributes that are not part of language ability, and random factors on test performance" (p. 166). However, since their influence cannot be eliminated altogether "the interpretation and use of language test scores must be appropriately tempered by our estimates of the extent to which these scores reflect the factors other than the language abilities we want to measure" (p. 166). Since both test method facets and personal attributes are fairly systematic, they can be predicted and controlled better than random factors.

Performance assessment procedures are much more complex than discrete-point testing, so the influence of a number of undesirable factors on the final scores can be more significant. Test method facets which have an effect on the assessment result relate to all three elements of the evaluation process, i.e. the task, the raters and the scale, as well as to their interaction. There has been a considerable body of research which investigated the potential areas of influence of these variables on the outcomes of performance-based assessment. This section will discuss a selection of such studies, with special attention given to the performance-based assessment of vocabulary. Unfortunately, since vocabulary is rarely the

main aim of such evaluation procedures, there are few studies devoted solely to this problem.

3.7.1 Influence of the Tasks

Weigle (2002) discusses multiple characteristics of a writing task which can have their influence on L2 learners' performance and their scores. These include: topic (general or requiring content knowledge), genre and rhetorical mode, cognitive demands as well as expected length, time allowed or the mode of transcription (handwritten or typed). Empirical studies such as Bouwer, Béguin, Sanders and van den Bergh (2015) and In'nami and Koizumi (2016) indeed demonstrate that writing abilities are to a large extend task-specific and not stable across tasks, but also that raters' judgements are not balanced across tasks. Read (2000, p. 198) also claims that "[it] is reasonable to expect that tasks vary in the demands that they make on the learners' vocabulary resources", and discusses the influence of such factors as familiarity with the topic or essay rhetorical mode. Yet, the study conducted by Park (2015) did not confirm that such effects are indeed reflected in raters' scores. She analysed judgements of two texts written by 78 Korean EFL students of different proficiency levels varying from beginner to advanced. The students wrote one narrative and one argumentative essay. The essays were scored by eight raters: four English native speakers of English and four Korean EFL teachers. The judges used a holistic scale and an analytic scale—including the component referring to vocabulary use—with a one-week interval between the ratings. The researcher ran a series of paired t-tests comparing the effects of rater background, task and type of scale on assessment outcomes. The results demonstrated that the task had an effect on only one set of scores: narrative essays were rated higher than argumentative essays, but this difference was statistically significant only for native-speaking judges and only their holistic scores. This effect was not significant for native speakers' composite analytic scores and for holistic and composite scores awarded by Korean judges. The effects of the task on vocabulary scores were not statistically significant for both groups of raters. Park's results may indicate that although vocabulary used by learners in their written output can differ by task, raters can balance the task's effects in their judgements of vocabulary use.

3.7.2 Influence of the Scales

The rubrics used to evaluate written production have an important impact on the final outcome of this process. As already discussed in the previous sections, scales define the construct that is being measured. They also contain an explicit statement of the criteria according to which the text should be evaluated. One of the problems which is analysed is the different ways in which holistic and analytic scales influence scores. Composite

scores resulting from analytic scales have been demonstrated to be more reliable than global ratings based on holistic rubrics. However, several researchers have pointed out that this is not clear how raters in fact come to make their final decisions. One possibility is that they in fact first make a global judgement of the quality of an evaluated text and then adjust the component scores to fit the overall ratings (Weigle, 2002, p. 120; Lumley, 2002). In fact, research shows that analytic scores for individual components do not differ widely from holistic judgements.

For example, Lee, Gentile and Kantor (2009) compared holistic and analytic ratings of 930 essays written in response to two *TOEFL* prompts.[8] Each essay was evaluated by two trained raters both holistically and analytically using specified rubrics. The final scores for the essays were the averages of two scores awarded by the raters. The analytic scales included six components: development, organisation, vocabulary, sentence variety/construction, grammar/usage and mechanics. The results, which were provided separately for each prompt, demonstrated high and moderate correlations between analytic ratings (from $r=0.89$ to $r=0.66$, $p<0.01$) and well as between individual analytic judgements and holistic scores (from $r=0.90$ to $r=0.72$, $p<0.01$). The component which showed the highest correlation with the holistic ratings was vocabulary ($r=0.90$ for Prompt 1 and $r=0.88$ for Prompt 2). Lee et al. (2009) also observed that all the ratings showed strong or moderate positive correlations with the total number of words in the essays (from $r=0.90$ to $r=0.60$, $p<0.01$), indicating that all ratings, in particular the holistic judgements and the scores for development and vocabulary, were influenced by essay length. The researchers hypothesised that the holistic and analytic scores could be correlated with each other through the variable of the essay length. That is why they decided to re-analyse the relationships between the scores by calculating partial correlations which eliminated its effect. The resulting correlations were lower, but still significant and ranged between $r=0.15$ to $r=0.69$, $p<0.01$ for analytic scores and $r=0.24$ to $r=0.55$, $p<0.01$ between analytic judgements and global scores. Again vocabulary produced the highest correlation with the holistic score for Prompt 1 ($r=0.50$, $p<0.01$) and was among the highest correlations for Prompt 2 ($r=0.44$, $p< 0.01$).

In an earlier study Astika (1993) examined the relationships between analytic scores awarded by two raters to 210 compositions written by overseas students at an American university. The raters used Jacobs et al.'s (1981) rubric, discussed in Section 3.4.2, which comprised five subscales, including one for vocabulary. The composite score for each essay was calculated by summing the raters' judgements for the individual components, taking into account their different weight. Similar to Lee et al. (2009), Astika also found significant—albeit slightly lower—correlations between individual components, ranging from $r=0.77$ to $r=0.44$, $p<0.01$. Additionally, the researcher performed a stepwise regression analysis in order to determine the amount of variance in the composite score

contributed by each component. The results indicated that the component which was the best predictor of the final score was vocabulary, accounting for 83.75% of its variance.

The two studies demonstrated that even though the analytic scales are considered more valid, as they take into account the non-parallel growth of various components of writing ability, in practice the profiles they produce, as well the composite scores they generate do not effectively differ from holistic ratings. The studies also pointed to the importance of lexical proficiency in general writing ability.

Another important point raised in research on performance-based assessment is how raters make use of scales in the rating process, how they interpret scaled descriptors included in the rubric and how they arrive at their final decisions. Two recent studies analysed exactly this problem through the use of think-aloud protocols.

Lumley (2002) and Lumley (2005) examined the rating process of four raters involved in scoring 24 texts (two writing tasks produced by 12 candidates) written as part of the *Special Test of English Proficiency (step)*, a high-stakes test administered to Australian immigrants for visa-related purposes. All the raters were trained and had a two-year-long experience in scoring *step* writing papers. In the study, they followed the same analytic scale used for the evaluation of the exam scripts. The rubric consisted of these components: (1) Task Fulfilment and Appropriacy (TFA), (2) Conventions of Presentation, (3) Cohesion and Organisation, and (4) Grammatical Control, with six levels for each component, accompanied by scaled descriptors. The scale did not have a separate category related to lexis, but vocabulary use was included in the descriptors for the TFA subscale. During the individual scoring session the raters were expected to voice all their thoughts and explain the reason for awarding a particular score. The think-aloud protocols were recorded and transcribed. Subsequently, the researcher coded the raters' comments using altogether 174 codes grouped into six categories. The codes were used to analyse (1) the steps the raters followed when ratings texts, and (2) the raters' interpretation of the descriptors included in the scale.

Lumley's examination of the coded protocols led him to an initial observation that raters followed the same sequence of rating a text, with only some small deviations. The sequence consisted of three main steps: the initial reading of a text and forming an opinion about it, rating the four scoring categories and then confirming or revising the scores. In relation to the attention given to the descriptors included in the scales, the researcher concluded that the overwhelming majority of comments related to features explicitly mentioned in the scale and that the comments referred to all elements included in the subscales. Yet, the quantitative analyses showed that not all these elements were awarded equal attention. In the TFA category, 32.7% comments concerned the relevance of the content and only 12.2% related to vocabulary use.

Lumley also conducted a qualitative analysis of raters' comments to gain an insight into how they interpret and apply the descriptors. The results led him to write that

> we may claim that although there appears to be some evidence that the raters understand the rating category contents similarly in general terms, there is also evidence that they sometimes apply the contents of the scale in quite different ways. They appear to differ in the emphases they give to the various components of the scale descriptors. Rather than offering descriptions of the texts, the role of the scale wordings seems to be more one of providing justifications on which the raters can hang their scoring decisions.
>
> (Lumley, 2002, p. 266)

In relation to vocabulary as one of the scoring criteria, he concluded that it does not play a major role in the evaluation of written production and it may also be subsumed under a more general category referring to the clarity of meaning.

One of Lumley's findings which showed that raters' comments on the whole tended to relate to the criteria mentioned in the scales was contradicted by the results obtained by May (2006) in a small exploratory study. She analysed retrospective think-aloud protocols produced by two trained and experienced raters, who evaluated learners' spoken production in a high-stakes English for Academic Purposes exam offered by an Australian university. Twelve candidates participated in a paired structured discussion task and their performance was assessed on an analytic scale including five components and scaled descriptors for five bands. The raters first watched a set of six paired discussions recorded earlier, rated them and then watched them again, this time stopping the tape and commenting on the learners' performance and the rating process. These comments were recorded, transcribed and then coded by the researcher. The quantitative analysis of the coded segments revealed that 30% of the comments did not relate to the aspects of performance covered by the scale. They were more general reflections on the rating process or specific comments on realisation of the task by the candidates. In addition, in the comments which did refer to the criteria included in the scales, additional attributes of these criteria, not included in the descriptors, were brought up:

> It was interesting to note that raters appeared to have "fleshed out" the criteria in the band descriptors with features that were not explicitly mentioned in the band descriptors, but which from the content and context of their comments, raters clearly regarded as salient to the categories in the rating scales.
>
> (May, 2006, p. 42)

The adequacy of different scales for evaluation of the L2 learner's performance is also scrutinised by analysing the scales' empirical validity. This involves linking individual descriptors included in the rubrics with observable features of texts and then examining if the variation in these observable characteristics affect the scores awarded to the texts. Two studies, described next in this section, did just that. They examined raters' attention to the features included in scale descriptors as well their sensitivity to lexical aspects of texts by manipulating the lexical features of the assessed texts and comparing the scores awarded to them.

Fritz and Ruegg (2013) examined raters' response to three aspects of learners' use of vocabulary in their written production: accuracy, range and sophistication. The researchers defined these features as determinants of the lexical quality of a text:

> In terms of lexis, a good quality essay can be defined as containing the following characteristics: a variety of different words, a selection of both low-frequency and topic-appropriate words, a high percentage of content words, and no or very few lexical errors (Read, 2000, p. 200). An essay with these lexical qualities would be expected to receive a high lexical score on any writing assessment that uses an analytic scale to measure lexis.
>
> (Fritz & Ruegg, 2013, p. 174)

In order to examine this assumption, the researchers chose a 137-word text written by an L2 learner in response to an exam prompt and then rewrote it several times so as each time each of the three lexical features represented a different level of ability. Only the essay's 32 content words were manipulated with every rewriting. Each of the resulting set of 27 texts represented a unique combination of the three lexical aspects at three different levels; for example a rewritten text could be characterised by high accuracy, medium sophistication and low range. Next, the manipulated texts were handwritten to look like regular exam scripts and added to authentic scripts written in response to the same prompt during a proficiency examination at a Japanese university. They were scored by trained and experienced rates who followed an analytic scale normally used in this exam. The scale contained four scoring components and five bands (0–4) with one of the components referring to vocabulary.

The differences in scores between essays representing low, medium and high levels of each of the three manipulated lexical aspects were examined separately with an AVOVA test. Only lexical accuracy produced a statistically significant result, indicating that essays with higher lexical accuracy received on average higher scores. The eta-square obtained for this variable in the analysis was 0.111, testifying that 11.1% of the variance in scores could be attributed to lexical accuracy. The statistical analysis of the other two lexical aspects did not produce

significant results, which indicates that the mean scores for the essays with high, medium and low levels of vocabulary range and sophistication did not differ. Unfortunately, the authors could not examine how the interaction of the three variables affected the scores due to insufficient amount of data.

Jarvis (2013) used a similar methodology involving manipulating lexical content of texts and examining the influence of this manipulation on raters' perceptions and scores. The aim of his study was the validation of one of the statistical measures of lexical quality of a text, lexical diversity, discussed in detail in Chapter 4. That is why only some findings of Jarvis's study will be described here, and the full account of his research will be provided in Section 5.3.3.

Jarvis starts with defining the concept of lexical diversity theoretically and distinguishing its six observable components: variability, volume, evenness, rarity, dispersion and disparity. The details of this reflection were discussed in Section 3.6. In an attempt to justify the proposed properties of lexical diversity, Jarvis designed two simple tasks in which human judges were asked to compare short text samples and evaluate their diversity. The samples' individual properties relating to lexical diversity were manipulated in order to examine their influence on human perception. The working definition of lexical diversity presented to the raters was very simple: "the variety of word use that can be found in a person's speech or writing" (p. 26). Task 1 consisted of six pairs of sentences which differed on only one property. For the obvious reason of very limited length of the samples, the property of dispersion was not included in the analysis, instead formal and semantic disparity were treated separately. The sentence pairs were presented to 130 judges, including 21 non-native speakers of English, who were to decide which sentence of a pair contained more varied vocabulary. The results demonstrated that two properties, variability and volume, showed strong effects on the participants' judgements, whereas semantic disparity, rarity and evenness demonstrated moderate effects. In Task 2, in turn, 38 human judges were presented with six paragraphs containing a narrative of the same events. One paragraph represented a typical and natural text produced by a native speaker, and the remaining five were modification of this baseline text, each time with one property being manipulated. This time the judges were asked to rank the six paragraphs from the most to the least diverse. The results demonstrated that the paragraph with a high number of rare words (high rarity) was judged to be most lexically diverse followed by the paragraphs with a high number of tokens (high volume) and a high number of types (high variability). The two remaining manipulated paragraphs contained lower levels of evenness and disparity than the baseline text, but interestingly even they were classified as more diverse than the reference paragraph. Jarvis interprets the overall results of these tasks as evidence that his six proposed properties are

indeed components of the construct of lexical diversity, with variability, volume and rarity being more prominent in the construct than evenness, dispersion and disparity, and that all six properties influence the human perception of a text's lexical variation.

3.7.3 Influence of the Raters

The influence of rater variables on the scores is conceptualised as rater reliability. If raters' assessment behaviour was truly objective and unaffected by factors other than the measured ability, their scores given to the same sample of learner writing would always be identical. The discrepancies in ratings given by different judges, or the same judge on different occasions, indicate that other factors also have an effect the raters' assessment. McNamara (1996) claims that two factors are responsible for the lack of perfect agreement between raters, even after training: their bias (that is their severity or leniency) and randomness (error). A rater's bias can relate to the overall performance, but can also apply solely to a particular aspect of performance. For example, Eckes (2008) and Eckes (2012) demonstrated that raters tend to be more severe in the assessment of the components that they perceive as important, and more lenient in the rating of these aspects of writing which they view as less vital. Interestingly, McNamara's (1996) results indicate that raters can be consistently more severe on one aspect of writing (grammatical accuracy), while their perception of the importance of this component in the overall rating can be low. Raters may also be more severe in the assessment of particular test items or candidates.

Research on rater assessment category bias is performed with the application of a complex statistical analysis called many-facet Rash measurement (MFRM). Until now such studies including vocabulary as one of the scoring components have not produced meaningful results as far as rater severity/leniency patterns towards lexis is concerned. For example, Wigglesworth (1993) found that when assessing oral interviews in an Australian high-stakes proficiency exam for immigrants, some judges marked the category referring to vocabulary consistently more harshly, whereas others scored more leniently on this component. Schaefer (2008) demonstrated that 40 native-English-speaking raters in Japan showed consistent bias patterns in their assessment behaviour: those who rated the components of Content and/or Organisation more severely, at the same time rated Language Use and/or Mechanics more leniently. Yet, the researcher found no such effect for Style and Quality of Expression (which referred to vocabulary use) or Fluency. In his analysis of 64 experienced raters involved in scoring the writing part of the *Test of German as a Foreign Language* (*TEstDaF*), Eckes (2008) found that the judges could be divided into six groups, based on their declared perception of the relative importance of various scoring criteria. Only one of

these groups, named by Eckes 'the Syntax group' due to their preference towards the criteria related to Linguistic Realisation, viewed vocabulary as a criterion of high importance in the assessment of learners' written production. Another group, labelled 'the Fluency group' due to their bias towards the criteria connected with global impression, perceived vocabulary use as a criterion of low significance. The remaining four groups (Correctness, Structure, Non-fluency and Non-argumentation) treated vocabulary as moderately important. In a follow-up study, Eckes (2012) examined the actual rating bias of 18 of the raters participating in his 2008 study. The raters' judgements concerning each scoring criterion for all the essays were analysed in order to establish the patterns of bias of individual raters towards nine assessment categories. Overall, Eckes's results indicated that raters demonstrated different patterns of severity/leniency bias towards vocabulary use. However, it was impossible to find more general patterns which would allow to link the perceived and real biases towards lexical proficiency with other marking criteria and more general patterns of behaviour of different types of raters.

3.8 Conclusion

This chapter examined the assessment of L2 performance by human raters in the light of information that their scores offer on learners' lexical proficiency and their underlying lexical competence. It presented the main criteria used in the evaluation of lexical proficiency. It also discussed components of the assessment process: the task, the scale and the rater, and reviewed most important studies referring to the influence of these facets on scores. The discussion revealed that although writing assessment has been subjected to an in-depth theoretical and empirical scrutiny, still relatively little is known about the contribution of various aspects of lexical proficiency to general writing ability, as well as about the applicability of human ratings of candidates' written performance for making inferences about learners' lexical competence.

Notes

1. Only scales designed for the assessment of learners at very low levels (such as Cambridge *KET*) or tests at the 'strong' extreme of McNamara's (1996) strong–weak continuum of performance assessment do not contain the language use component.
2. www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf (accessed 22 December 2018)
3. www.ets.org/toefl/ibt/scores/understand (accessed 22 December 2018)
4. www.cambridgeenglish.org/first (accessed 22 December 2018)
5. www.cambridgeenglish.org/proficiency (accessed 22 December 2018)
6. The term appropriacy is ambiguous in the context of language assessment. It is frequently used to indicate the right choice of linguistic resources to match the register of a text or speech. As evidenced by the definition provided by the

exam developers and reproduced in Figure 3.6, in Cambridge English scales this term refers rather to the choice of words with the right meaning to express the intended message and correct collocates.

7. See for example *TOEFL Research Topics* (www.ets.org/toefl/research/topics/scoring) for a list of studies exploring validity and reliability of *TOEFL* writing and speaking items (accessed 22 December 2018).
8. They were prompts from an earlier version of the *TOEFL* exam, which correspond to the independent task in the most recent version.

4 Statistical Measures of Lexical Proficiency

4.1 Introduction

The previous chapter discussed performance-based assessment of lexical proficiency carried out by human raters with the help of various assessment rubrics. The last section of that chapter (Section 3.7) pointed out that in spite of meticulous procedures aiming at ensuring objectivity of such assessment—such as development of rating scales and rater training—the scores awarded by human judges are influenced by several extraneous variables and thus not fully objective. In addition, the final scores, even in analytic evaluation, can mask learners' complex profiles related to different dimensions of lexical proficiency. Finally, the assessment performed by trained judges is expensive and time consuming. That is why for years researchers have been engaged in a search for methods which would provide the means for the assessment of quality of writing, including lexical proficiency, in objective and quantifiable terms. Different mathematical formulas in the form of simple frequencies, ratios or complex indices have been proposed to capture the lexical properties of a piece of writing or a longer stretch of spoken language. The lexical measures have been related to different dimensions of lexical proficiency and an observed variation in their values was meant to reflect differences in learners' lexical competence.

Wolfe-Quintero et al. (1998), Linnarud (1986) and Crossley and his colleagues (various dates, see later in this chapter) offer a comprehensive review of research in this area. All these publications present and discuss many different formulas pertaining specifically to the lexical component of the learner's linguistic capacity. Wolfe-Quintero et al. group them into measures which relate to the three aspects of general language proficiency: complexity, accuracy and fluency. Bulté et al. (2008), on the other hand, match various measures with specific dimensions of lexical proficiency and, by extension, lexical competence. Their taxonomy is presented in Figure 4.1.

In this chapter, the two classifications will be merged to facilitate a systematic presentation of a host of different gauges capturing the use of

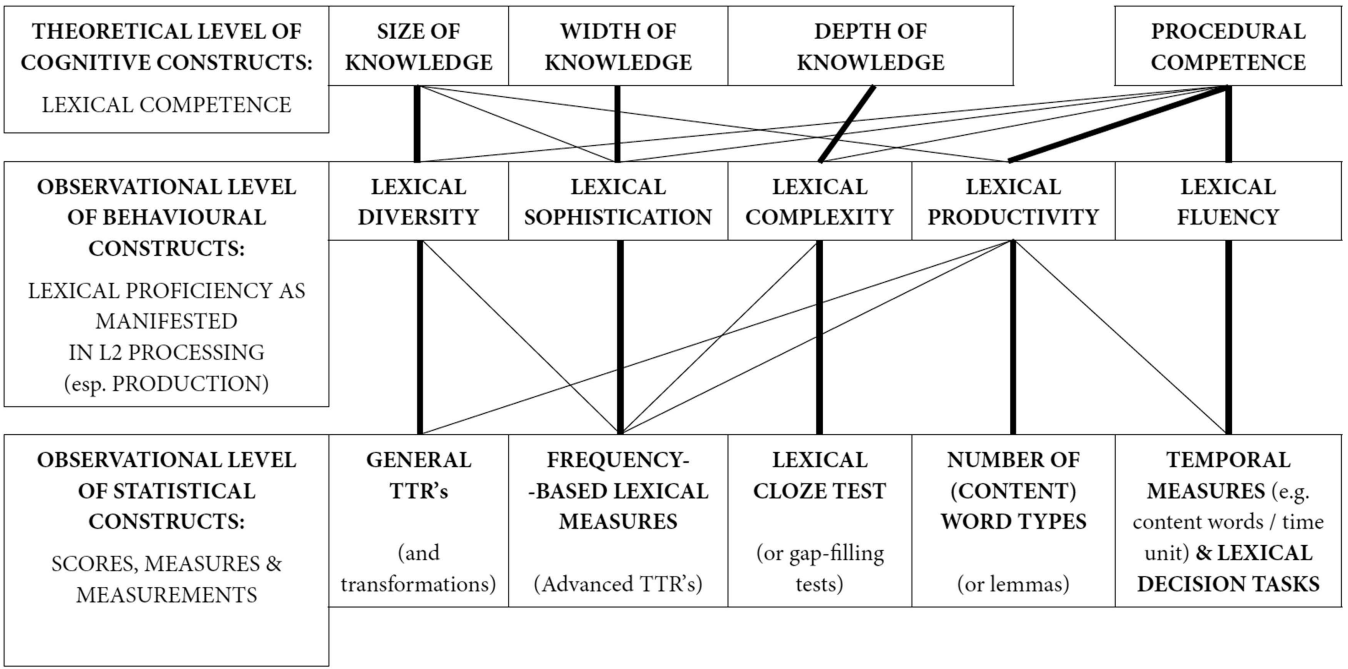

Figure 4.1 Lexical Measures Matching Different Dimensions of Lexical Proficiency

Source: Bulté et al. (2008, p. 279). Reprinted with permission.

vocabulary in natural language production, which have been proposed in the literature over the last 100 years.

4.2 Lexical Measures of Fluency and Measures of Lexical Productivity

Fluency as such does not relate to lexical proficiency itself but it applies to the overall linguistic ability. It depicts the ease and speed with which a learner can process the second language for both production and comprehension. Fluency refers to "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (Ellis, 2003, p. 342). However, writing presented for evaluation is usually edited, so the latter characteristic of the performance is not visible to the assessor. The speed of writing is also inscrutable if all that the rater has at his or her disposal is the final product. An indirect indication of the rate of writing is the number of words the learner produced in a set period of time. Thus, the method of measuring fluency in a piece of writing is by gaging its length. It is based on the assumption that a more proficient L2 user has not only more words in his or her lexicon but he or she can also access them quicker and in consequence produce a longer text. If the time for completing a writing task is predefined, the differences in length of compositions produced by two writers can reflect differences in their fluency and thus proficiency.

The simplest way of quantifying fluency is by counting the number of words in a text. This measure can easily be extracted for a written sample, particularly with the use of a computer. Yet, on scrutiny even this simple index contains some pitfalls related to the precise definition of a word. There is no one generally accepted convention developed for handling contractions or hyphenated words and each computer programme treats then differently, which has its influence on the final count (Gajek, 2006; Anthony, 2014).

Lexical productivity is a similar concept to fluency, but instead of focusing on the temporal aspects of L2 production, it targets its yield, i.e. it indicates how many (content) word types the learner produced to complete a task. In addition to tapping lexical access, this measure also reflects the size of lexical competence, because the learner who produces more content types also needs to have more words in his or her lexicon. That is why Lu (2012b) also considers it the crudest index of lexical variation. This index—expressed by the total frequency of (content) word types in a text—can be computed with such text-processing software as concordances, applying the so-called stop list with all function words to be excluded from calculations. Lu's (2012a) *Lexical Complexity Analyzer*[1] produces this gauge fully automatically.

It has to be noted, however, that the number of running words or of (content) word types in a sample of L2 production are very dependent

on the nature of an assignment. Thus, they can only serve to demonstrate differences between learners engaged in the same or comparable task in the same or comparable setting. They can also be used to measure the learner's development over a period of time if identical conditions are secured. But even then the counts have to be interpreted with caution because the differences in the produced yield can also be influenced by learners' varying attitudes to the task or their motivational levels, not to mention a lot of other extraneous variables such as familiarity with the topic.

4.3 Measures of Lexical Accuracy

Accuracy was defined by Housen and Kuiken (2009, p. 461) as "the ability to produce error-free speech", but this concept also applies to writing. It may be important to note that, whereas in speech errors can not only be a result of inadequate or incomplete knowledge but also of high processing demands, writing is more likely to exhibit competence rather than performance errors. This is due to the fact that learners usually have time to edit and proofread their texts before they submit them for assessment.

Errors are usually considered in relation to grammar and most accuracy measures focus on grammatical correctness of a text. Yet, learners can also make mistakes in their use of L2 vocabulary and this fact can also be reflected in general accuracy measures such as the frequency of errors in a text, or, conversely, the number of error-free clauses/T-units/sentences in a piece of writing. Lexical errors are also included in ratio-based measures of accuracy such as the number of errors per T-unit/clause/sentence or the number of error-free clauses/T-units/sentences per the total number of these units. However, some researchers focused specifically on lexical errors and calculated the number of lexical errors per total number of lexical words or clauses (Linnarud, 1986; Hyltenstam, 1992; Bardovi-Harlig & Bofman, 1989), or the number of verb lexical errors per the total number of verbs (Engber, 1995) or finally the number of semantic errors per the total number of errors (Arthur, 1979). More complicated formulas pertaining to lexical accuracy were also proposed. Arnaud (1984) introduced the lexical richness index which is calculated with the following formula:

$$\text{Lexical richness} = (\text{lexical word types} + \text{rare word types}) - (2 \times \text{erroneous lexical words})$$

and can only be applied to texts of equal length. According to Arnaud (1984), this gauge discriminates well between texts written by native and non-native speakers, yet this conclusion was not tested statistically on a large sample.

Engber (1995) proposed a lexical accuracy index which is calculated as follows:

$$\text{LexAccIndex} = \frac{\text{lexical word types} - \text{lexical errors}}{\text{lexical word tokens}}$$

Measures of lexical accuracy have never been claimed to represent the L2 learner's overall lexical proficiency. That is why they are usually used with other gauges referring to other characteristics of word use. For example, the two indices mentioned earlier also capture lexical diversity by juxtaposing lexical types and tokens in one formula (Arnaud's formula does not do it explicitly, but since it can only be applied to compare texts of the same length, the number of types taps diversity). Arnaud's formula, in addition, includes lexical sophistication in the equation. Lexical diversity and sophistication will be discussed in the following sections.

Lexical accuracy indices are less dependent on the task and its setting, thus allowing comparisons between texts written in different conditions. Yet, they are much more difficult to implement as errors in general and lexical errors in particular are fuzzy categories. One particular problem is making a distinction between lexical, grammatical or pragmatic errors. If a learner uses the word *kid* in a formal text, it can testify that he or she does not know the pragmatic constrains on the use of this particular item, or it can indicate that he or she has problems in recognising or applying an appropriate style in a particular situation. In the former case this could be interpreted as a lexical error, in the latter as a pragmatic error. Furthermore, some researchers include spelling mistakes in the inventory of lexical errors (e.g. Ander & Yıldırım, 2010), some others assume that spelling is a separate linguistic system (the same as pronunciation) and should be considered separately (e.g. Ramineni et al., 2012). Even the very decision if a particular use of a word is an error or not can be arbitrary (cf. Lewandowska-Tomaszczyk, Leńko-Szymańska, & McEnery, 2000; Leńko-Szymańska, 2002), as there are different language standards or the learner can use a word creatively. This measure is also difficult to automate unless samples of essays were first tagged with lexical errors which is not only a subjective but also very time-consuming task.

4.4 Measures of Lexical Complexity

The metrics which have been most widely used and researched, and which show the greatest promise for the assessment purposes relate to lexical complexity. As mentioned in Chapter 1, complexity is the most controversial dimension of proficiency, and it has inspired researchers to propose a whole host of different methods of its measurement. They capture the range and the quality of words that a text is made of, irrespective of the

fact if they are used correctly or not and of how many of them are used altogether.

4.4.1 Lexical Diversity (Variation)

The concept of lexical diversity, also referred to as lexical variation, is based on the assumption that a text of a better quality is characterised by less repetitiveness and, in consequence, more variety in the choice of lexis. Certain repetitiveness of words cannot be completely avoided. Grammatical words have to occur in obligatory contexts and the learner has no or little freedom in their selection. Lexical words also need to be repeated to build a cohesion of a text (Linnarud, 1986, pp. 8–12; McCarthy & Jarvis, 2010). Yet, it is assumed that a more competent writer has at his or her disposal a larger pool of items to choose from and he or she is more flexible in expressing intended meanings, thus his or her writing should be more lexically diverse. At the same time, Linnarud (1986, p. 58) observes that lexical diversity can also be a result of a greater creativity in approaching the topic. Also, van Hout and Vermeer (2007) suggest that the kind of language activity a learner engages in affects the variation in his or her vocabulary choice.

The basic formula for calculating lexical diversity is as follows:

$$\text{Lexical diversity (variation)} = \frac{\text{number of types (different items)}}{\text{number of tokens (all items)}} \times 100\%$$

This formula can also be rendered by two alternatives:

$$\text{Lexical diversity (variation)} = \frac{\text{number of lexical types (different lexical items)}}{\text{number of lexical tokens (all lexical items)}} \times 100\%$$

$$\text{Lexical diversity (variation)} = \frac{\text{number of lemmas (different base forms)}}{\text{number of tokens (all items)}} \times 100\%$$

The former alternative has an advantage of disregarding grammatical words which are highly repetitive in any type of writing and speech and thus significantly lower the score. The latter transformation, on the other hand, is particularly suitable for highly inflectional languages such as French or Polish for which scores can otherwise be inflated due to a multitude of grammatical forms of individual lemmas, treated as separate types by the basic formula (Treffers-Daller, 2013).

The measure of lexical diversity is frequently referred to as the type/token ratio (TTR) in literature. It can be calculated easily with the help of computer programmes dedicated to text analysis. Concordancing packages report type/token ratio in their descriptive statistics of a text. The ratio based exclusively on lexical words is more complicated to calculate, as it requires a list of stop words which need to be excluded from the analysis. The formula based on lemmas requires lemmatising the text first.

In spite of its popularity, as well as its theoretical and intuitive appeal, the application of this index for assessing samples of learners' written production is problematic. Johnson (1944), who is credited with introducing this measure (Treffers-Daller, 2013, p. 80) observed that TTR was highly sensitive to text length, with its values diminishing with an increase in the number of running words in a piece of writing. In other words, the longer a text, the more repetitive its vocabulary tends to be. Thus, this measure cannot be applied for a comparison of texts of considerably different lengths. In order to eliminate the effect of text length, a number of mathematical and computational permutations of the basic formula were proposed in the literature. Johnson (1944) himself introduced the Mean Segmental Type/Token Ratio (MSTTR) as a solution to the problem. It involved dividing the text into smaller segments of 100 words (or other number of tokens), computing TTR for each full segment and calculating the mean value. Such an approach, however, presented some serious drawbacks. While being relatively suitable for documents over 1000 words long, it continued producing unreliable results for shorter pieces of writing as it discarded the sizeable proportion of a text not fitting a full 100 word segment. Reducing the length of a segment, on the other hand, led to inflated results and reduced the capability of the index to distinguish between texts of different lexical quality (McCarthy & Jarvis, 2010).

Other permutations of the formula also sought to diminish the influence of the length:

V = number of types,	N = number of tokens
Guiraud (1954)	Root TTR = $V/\sqrt{N}$
Carroll (1964)	Corrected TTR = $V/\sqrt{2N}$
Herdan (1960)	Log TTR_1 = V / log N
Herdan (1966)	Log TTR_2 = log V / log N
Maas (1972)	$a^2 = \frac{\log N - \log V}{\log_2 N}$

Still other alternatives adjusted for frequency of occurrence of types, for example Yule's K, which measures the likelihood of two types, chosen at

random from the text, being the same (Yule, 1944). A detailed discussion of these measures can be found in Tweedie and Baayen (1998), who observe that all the different permutations of the basic formula remain influenced by text length.

In more recent years even more complex formulas of lexical diversity were proposed in the literature. Malvern and Richards (1997) introduced a new measure, called D, which they further developed in Malvern and Richards (2002) and Malvern, Richards, Chipere and Durán (2004). It is a complex algorithm which takes the following form:

$$\mathrm{TTR} = \frac{\mathrm{D}}{\mathrm{N}}\left[\left(1 + 2\frac{\mathrm{N}}{\mathrm{D}}\right)^{\frac{1}{2}} - 1\right]$$

where D is a parameter whose value can be manipulated in order to create a theoretical curve that fits best the random-sampling TTR curve for the text. It is the parameter D which is proposed by Malvern et al. (2004) as the index of lexical diversity. Its computations involve several stages. First, a hundred random samples of 35 tokens are drawn from the text and mean TTR for these samples is calculated. The same procedure is repeated 15 times for samples of 36–50 words long. The mean values for the samples of various sizes are then plotted in a graph and they create a curve for the text. The next step involves finding a theoretical curve which matches the random-sampling TTR curve. This is done by adjusting the coefficient D in the algorithm, so as it produces the best-fitting theoretical curve. D's predicted values range from 10 to 100.

The calculation of D is very difficult to do manually so a dedicated programme was created for this purpose. It is called *vocd* (McKee, Malvern, & Richards, 2000) and is freely available in the Computerized Language Analysis (CLAN) suite of programmes.[2] Two more computer programmes can be used to compute the D index for a text: *D_Tools* v.2.0 developed by Lognostics[3] and *Coh-Metrix* (Graesser et al., 2004). The D index is sometimes referred to as vocd or vocd-D (McCarthy & Jarvis, 2007; McCarthy & Jarvis, 2010). It has been employed in a number of studies from various disciplines (see Malvern et al., 2004 for an extensive discussion of the use of this index in research on second language acquisition).

However, McCarthy and Jarvis (2007) questioned the validity of the D-index by demonstrating that it turned out not to eliminate the influence of text length completely. The researchers proposed a different rendering of a hypergeometric distribution function corresponding to D, which they called HD-D. It consists of calculating, for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 tokens drawn from the text. The probabilities for all lexical types in the text are then added together, and the sum is used as an index

of lexical diversity of the text (McCarthy & Jarvis, 2010, p. 383). Vocd-D and HD-D are two mathematical representations of the same mathematical concept and that is why it is not surprising that their values correlate very strongly with each other r=0.971 (McCarthy & Jarvis, 2007). However, as the creators of the HD-D index claim:

> the random sampling and curve-fitting procedures used by *vocd* introduce a certain degree of noise (or imprecision) into the . . . conversion, which results in final LD indices (i.e. D scores) that are not fully precise.
>
> (McCarthy & Jarvis, 2007, p. 464)

HD-D is also too complex to be computed manually. The calculation can be done with a programme called *Gramulator* 6.0 (McCarthy, Watanabe, & Lamkin, 2011).[4] This programme also reports the Maas index.

McCarthy (2005) and McCarthy and Jarvis (2010) introduce yet another algorithm to represent lexical diversity of a text, which they call the Measure of Textual Lexical Diversity (MTLD). They point out that both vocd-D and HD-D are based on random sampling from a text. Such a procedure does not take into account the fact that texts possess a certain structure and that humans process texts sequentially. Yet, the researchers do not claim that non-sequential computational processing of a text does not have its merits. That is why their index involves processing a text in both fashions. Similarly to the Mean Segmental TTR, the calculation is performed on sequential segments of a text, but the segments do not have a specified size. Instead, the researchers took an alternative approach. MTLD divides the text into succeeding segments whose TTR reaches or drops below the value of 0.72, and then calculates the average number of tokens in these segments. The researchers chose this value based on their earlier tests which had demonstrated that while TTR ratio tended to fluctuate at the beginning of a text, it always stabilised somewhere around the values of 0.72–0.75. This threshold was then chosen as the stop point for each segment—called a factor by the researchers—after which the calculation of TTR for the following segment started anew. In other words MTLD processes the text sequentially calculating TTR for a string of subsequent words until it reaches the critical value. At this point a new string starts to be processed. The remaining words in the last string which has not reached the critical value are counted as a partial factor. This procedure is exemplified using a famous quotation from Shakespeare (Figure 4.2; see McCarthy & Jarvis, 2010, p. 384 for another example).

The first factor in the quotation in Figure 4.2 is unusually short as it contains repetitions of items in a short span of words. Generally, factors are much longer and their values stay in the range of 60–80 words. The number of tokens is divided by the number of factors, including

To (TTR=1.00) be, (TTR=1.00) or (TTR=1.00) not (TTR=1.00) to (TTR=0.800) be (TTR=0.67) [Factor 1]—that (TTR=1.00) is (TTR=1.00) the (TTR=1.00) question (TTR=1.00).

Figure 4.2 Example of Computing the MTLD Index

the fraction representing the last segment. For example, for a text of 340 words and the factor count of 4.404, the MTLD value is 77.203 (McCarthy & Jarvis, 2010, p. 384). Then, the processing starts again in the reverse direction (from the last to the first word of the text). The two MTLD values computed for both directions are averaged to produce the final MTLD index for the text. According to Crossley, Salsbury and McNamara (2010b, p. 64) the MTLD index does not vary as a function of text length for texts whose size is within the 100–2000 word range. Similarly to the previous two indices of lexical diversity, MTLD has to be calculated by a computer programme. It is available from *Coh-Metrix* and *Gramulator* 6.0.

The algorithms developed to capture lexical diversity have been extensively analysed, compared and validated against one another based on written and oral production data collected from native speakers and L2 learners at different proficiency levels (Vermeer, 2000; Jarvis, 2002; Malvern et al., 2004; McCarthy & Jarvis, 2007; van Hout & Vermeer, 2007; Daller & Xue, 2007 McCarthy & Jarvis, 2010; Lu, 2012b; Treffers-Daller, 2013; deBoer, 2014). Their applicability was also tested for languages other than English, for example Dutch (Vermeer, 2000; Vermeer, 2004) and French (Tidball & Treffers-Daller, 2007; Treffers-Daller, 2013). The results of these studies lead to equivocal conclusions. For example, Meara and Bell (2001) point out that indices of lexical diversity are intrinsic measures of lexical richness, i.e. they evaluate the vocabulary of a text in relation to the text itself, ignoring the status of individual words in language. They give an example of three sentences:

Example 1: The man saw the woman.
Example 2: The bishop observed the actress.
Example 3: The magistrate sentenced the burglar.

The sentences consist of the same number of types and tokens and thus they yield the same value for the lexical diversity index, no matter which algorithm is applied. However, as Meara and Bell observe, intuitively these three sentences differ in terms of their lexical quality. While the first sentence can easily be produced by a fairly novice learner of English, sentences 2 and 3 are much more likely to be credited to more advanced

learners, due to the more sophisticated vocabulary they contain. Thus, it seems that lexical diversity by itself is not nuanced enough to present a full picture of a text's lexical quality and other measures are needed to accurately describe the L2 learner's lexical proficiency. Such a claim was confirmed by Vermeer (2000), who compared several indices of lexical variation and concluded that none of them produced sufficiently reliable and valid results, in particular for later stages of vocabulary acquisition (from 3000 words on). He suggested that measures based on the degree of difficulty of the words used (lexical sophistication) are more effective gauges of lexical complexity. Daller and Xue (2007), on the other hand, came to a contrary conclusion about two indices: D and Guiraud Index (see Chapter 5 for a detailed discussion of this study). McCarthy and Jarvis (2010), in turn, conducted a validation study of four indices, all capturing lexical diversity: Maas, vocd-D, HD-D and MTLD. The analysis of the results led the researchers to conclude that the four indices may not assess exactly the same latent trait and may be able to capture different information related to lexical diversity (p. 391). Yet, in order to confirm this hypothesis, the authors call for more research in this area.

4.4.2 Lexical Sophistication

As mentioned in the previous section, an observation concerning the lexical texture of a text is that use of sophisticated words contributes to the quality of writing. In other words a good text includes a number of advanced words. Sophisticated words do not necessarily have to name refined concepts. For example, the words *rain* and *precipitation* denote approximately the same referent, which is a commonplace weather phenomenon. Yet, the word *precipitation* is regarded as more advanced.

Lexical sophistication is expressed by the formula:

$$\text{Lexical sophistication} = \frac{\text{number of sophisticated tokens}}{\text{number of tokens}} \times 100\%$$

Yet, the same formula can be applied to types rather than tokens, thus reducing the effect of repetitiveness of both simple and advanced items:

$$\text{Lexical sophistication} = \frac{\text{number of sophisticated types}}{\text{number of types}} \times 100\%$$

The main factor determining a word's sophistication is its frequency in language. The less frequent a word, the more sophisticated it is considered to be. The information on frequency of individual words is drawn from corpora which are large enough to be considered representative of language as a whole, or of a certain variety of language (e.g. written

or spoken modes, specialised registers). Special text analysis software (a concordancer) generates a list of all the types in a corpus with the numbers of their occurrences. The frequency of a word is calculated with the following formula:

$$\text{Frequency of a word} = \frac{\text{number of occurrences in the corpus}}{\text{number of tokens in the copus}} \times 1 \text{ million}$$

For example, the word *rain* has the frequency of 62.14 occurrences per 1 million words and the word *precipitation* occurs 2.24 times per 1 million words, according to the British National Corpus (The BNC Consortium, 2007).

Milton (2007) criticised the operationalisation of sophisticated vocabulary based solely on frequency. He pointed out that some words are introduced early in L2 instruction, as they belong to the thematic fields particularly relevant to L2 learners (e.g. cutlery items) and they are learned at the beginning stages even though they are not among the very frequent words in the target language. That is why some researchers identified sophisticated words as items absent from an inventory compiled for the purposes of L2 instruction and syllabus development, such as Thorndike's *Teacher's Word Book* (1921) or West's General Service List (1953). However, in the compilation of such lists, the corpus-based information on frequency or at least the compilers' intuitions about frequency, usually remain one of the important classifying criteria. Still another approach to defining basic and sophisticated vocabulary is asking for teachers' opinions (Daller, van Hout, & Treffers-Daller, 2003), but such judgements are also influenced by human perceptions of word frequency.

The choice of a threshold between basic and sophisticated words is not trivial, as it affects the values of the lexical sophistication metrics. This problem was summarised by Tidball and Treffers-Daller (2007) in the following quotation:

> Clearly, the effectiveness of this measure depends entirely on a valid operationalization of the basic vocabulary. If the choice of words in the basic vocabulary is flawed for some reason [the metric] cannot measure informants' vocabulary richness correctly.
>
> (Tidball & Treffers-Daller, 2007, pp. 136–137)

So far different researchers have applied different approaches to defining basic and advanced vocabulary and they have used different thresholds between simple and sophisticated words. Because of the lack of an agreement on what is a sophisticated word, the results of investigations using lexical sophistication indices are not directly comparable.

One of the more widely used indices of lexical sophistication is the so-called Lexical Frequency Profile (LFP) proposed by Laufer and Nation

in 1995. It reports proportions of words in a text belonging to the first 1000 most frequent words in English, second 1000 most frequent words of English, Academic Word List (Coxhead, 2000) and words not found in any of the previous lists. However, because an index consisting of four proportions is difficult to interpret and compare across texts, Laufer (1995, 1998) also introduced what she called Condensed Lexical Profile (CLFP) which is just the proportion of items with frequency beyond the 2000 most frequent items. More recently the LFP has been based on a number of different frequency lists compiled by researchers for their own projects from different corpora. The index can be obtained for a text in an automated way with the help of software designed especially for this purpose—Paul Nation's *Range* (n.d.), Lawrence Anthony's *AntWordProfiler* (n.d.) and Tom Cobb's *Vocabprofile*.[5] They all include ready-made lists of words at different frequency levels but *Range* and *AntWordProfiler* also allow users to base their analysis on their own inventories.

The classic lists provided by these programmes, first complied by Nation and distributed with *Range*, are in fact lists of word families; thus, they contain the words' base forms together with all their inflectional and derivational forms, as in the example drawn from the 1000-word list (Figure 4.3).

Such an approach has been criticised from both developmental and frequency perspective (Vermeer, 2004; Witalisz, 2007; Lindqvist, Bardel, & Gudmundson, 2011, p. 227; Lindqvist et al., 2013, p. 114). First of all, the fact that an L2 learner knows the base form of a word,

ACCEPT
- ACCEPTABILITY
- ACCEPTABLE
- ACCEPTABLY
- ACCEPTANCE
- ACCEPTANCES
- ACCEPTED
- ACCEPTING
- ACCEPTOR
- ACCEPTORS
- ACCEPTS
- UNACCEPTABILITY
- UNACCEPTABLE
- UNACCEPTABLY

Figure 4.3 A Word Family From the 1000 Frequency Band Available in *Range*
Source: Nation (n.d.)

for example *accept*, does not imply that he or she knows all its related forms, in particular its complex derivatives such as *unacceptability*, which are not always related to the base form in a transparent way. Second, the frequencies of different word forms of the same headword are very different. That is why some researches decided to use in their research lists consisting of lemmas rather than word families, and, in addition, to use word frequency information from lists compiled with more consideration given to different modalities and to different text types (Vermeer, 2004) or derived from spoken and written corpora separately, according to the analysed text type (cf. Lindqvist et al., 2011; Lindqvist et al., 2013). Such a solution, however, can pose some technical challenges in the case of English as different parts of speech have identical forms. Thus, producing an index based on lemmas requires much more pre-processing of a text. Gardner (2013) proposed another solution. He compiled the Common Core List, which was created by trimming the list of 4000 most frequent BNC word families, as provided by *Range*, of items which do not occur in the list of 4000 most frequent lemmas retrieved from the Corpus of Contemporary American English (COCA, Davies, 2008).

A more complex approach to representing the lexical sophistication of a text was taken by Meara and Bell (2001). They argue that infrequent words occur rarely in texts produced even by advanced learners and native speakers, so an index based solely on proportions of infrequent items (however they are defined) will not be sensitive enough to distinguish between learners at different proficiency levels. Moreover, in order to produce reliable results, it requires longer texts of over 200 words which are difficult to obtain from learners, in particular those at lower proficiency levels. In addition, the researchers challenge the claim made by Laufer and Nation (1995) that the index is independent of text length, and assert that much longer texts produce lower proficiency profiles. Therefore, the researchers propose a new index of lexical sophistication. They divide a text into segments of ten words, irrespective of punctuation, and calculate the number of infrequent words in each segment. In their formula, infrequent words are defined as any items beyond the 1000-word frequency level, excluding proper names. Generally, ten-word segments contain very few, if any, infrequent items. Next, the segments containing 0, 1, 2, 3, 4, 5, etc. infrequent words are tallied and the tallies are turned into proportions (e.g. the proportion of segments containing three infrequent words). The proportions of segments containing from 0 to 10 infrequent words can be plotted in a graph. The plotted line will always be highly skewed to the left as the larger the number of infrequent items in a segment, the fewer such segments can be found in a text. The formula describing best such a skewed distribution is called Poisson distribution and it is expressed by the following algorithm:

$$P_N = \frac{\lambda^N e^{-\lambda}}{N!}$$

The critical parameter in the equation is λ (lambda) and its value decides on the shape of the curve. Similarly to the computation of the index of lexical diversity D, the value of λ, which produces the curve fitting best the distribution of the proportions of different segments in a text, is taken to represent the index of lexical sophistication of this text. Lambada's values typically range from 0 to about 4.5, with higher figures corresponding to a higher proportion of infrequent words. Its advantage is that it can produce reliable indices even for short texts. Again, such an index is not calculated easily without appropriate software, thus a programme called *P_Lex* was created by Meara and Bell (2001) to produce the λ for a text. The programme is freely available at the Lognostics website.[6]

Crossley, Cobb and McNamara (2013) propose yet another approach to capturing lexical sophistication of a text and solving the problem of arbitrary cut-off points. Instead of reporting proportions of words belonging to different proficiency bands, they provide a single figure which is a mean frequency of all the words in a text. This index is calculated as follows: the frequency of each token is checked in a relevant frequency list selected for the analysis. The frequencies of all the items are added and then divided by the number of tokens. The resulting value is an index of lexical sophistication of the text. The higher values indicate that a text does not contain many infrequent words. Since the mean frequencies of words in a text are usually large numbers in a range of a few thousands of words, Crossley et al. suggest that they can be made more manageable by applying logarithmic transformation. The mean frequency of this paragraph is 1009.25 and its logarithmic transformation equals 3.00. The researchers claim that their index of lexical sophistication is more accurate and, unlike the FLP, sensitive enough to capture even small differences and gains in L2 learners' vocabulary knowledge. Also, unlike the LFP, it is not influenced by text length. Its disadvantage is that its value is less meaningful and harder to interpret in comparison with the LFP. The mean frequencies of content words and log mean frequencies of all words for a text are provided by *Coh-Metrix*.

Daller et al. (2003) compare the measures of lexical sophistication and diversity. They observe that "both aspects of lexical richness are related since it can be assumed that a greater lexical variation will lead to a greater use of less frequent or rare words" (p. 203; cf. Jarvis, 2013). They also claim that each measure has certain limitations and by itself does not fully capture the lexical development. That is why they introduce two measures which are the combination of measures of lexical diversity and lexical sophistication:

$$\text{Advanced TTR} = \frac{\text{number of advanced types}}{\text{number of tokens}}$$

$$\text{Guiraud advanced} = \frac{\text{number of advanced types}}{\sqrt{\text{number of tokens}}}$$

The interpretation of lexical sophistication for making judgements about the learner's L2 lexical proficiency is based on the assumption that the acquisition of vocabulary follows roughly the order of frequency (Palmer, 1917; Meara, 1992; Milton, 2007). Barring special needs or interests of particular learners, learners generally start building their L2 lexical capacity from the simplest and most useful words which happen to be at the same time the most frequent words in a language. Thus, the more of sophisticated words the learner uses in his or her writing, the larger his or her L2 lexicon is assumed to be.

The applicability of different indices of lexical sophistication for measuring the vocabulary knowledge of L2 learners has been analysed in numerous studies (Laufer & Nation, 1995; Meara & Bell, 2001), and also for languages other than English, for example Dutch (Vermeer, 2004), French and Italian (Lindqvist et al., 2011; Lindqvist et al., 2013). The results indicate that this measure remains stable across two comparable texts written by the same learners and it can also discriminate between learners at different proficiency levels. Moreover, the effectiveness of various sophistication indices was compared with the metrics of lexical diversity (Linnarud, 1986; Laufer, 1991; Laufer, 1994; Daller & Xue, 2007; Tidball & Treffers-Daller, 2007). Here the results are ambiguous. Some researchers have demonstrated that lexical sophistication can discriminate better between learners at different proficiency levels (Laufer, 1994), others that lexical diversity is more effective for this purpose (Daller & Xue, 2007), still others that both are equally useful (Tidball & Treffers-Daller, 2007). Vermeer (2000) and Daller et al. (2003) claim that while lexical diversity may distinguish better between lower-level learners, it is the use of rare words, i.e. lexical sophistication, which characterises best the language produced by more advanced learners.

Lexical sophistication is also an important component of readability formulas, which indicate how difficult a text can be for its potential readers. Over the years, several mathematical formulas have been proposed to measure the difficulty level of a text. They are all algorithms based on two main variables: the average sentence length and a gauge of sophistication of vocabulary in a text. In early readability measures the sophistication of words was often defined not through their frequency but the number of syllables they consist of, based on the observation that long words are more difficult to process (e.g. Gunning fog formula, Gunning, 1952; or Flesch-Kincaid reading level, Kincaid, Fishburne, Rogers, & Chissom,

1975). However, some readability measures define sophisticated words through their frequency in language (e.g. Dale-Chall formula, Dale & Chall, 1948; and Lexile text measure[7] developed by an educational company MetaMetrics, Stenner, Burdick, Sanford, & Burdick, 2006). In the recently developed web-based programme measuring the difficulty level of Polish texts, *Jasnopis*[8] (Gruszczyński & Ogrodniczuk, 2015), word sophistication is calculated based on a list arranged according to perceptual frequency of words, which had been compiled in psycholinguistic tests (Imiołczyk, 1987).

The readability formulas highlight the fact that too many sophisticated words not only stop contributing to the quality of a text, but may even have a counterproductive effect by affecting its intelligibility. Writing manuals, in particular those addressed to authors of texts with wide readership, like newspaper or magazine articles, in fact recommend using simpler vocabulary in order to render a text more accessible to readers. The same trend is now visible in official documents. Thus, the contribution of lexical sophistication to the quality of a text is rather complex and the number of sophisticated lexemes used in written discourse can be influenced by its purpose and its intended reader. Advanced learners can be aware of these constrains and their texts may not always reveal the full potential of their lexicons. Yet it is safe to assume that for learners below near-native proficiency this measure is a good indication of the size and the complexity of their word stores.

4.4.3 Older Measures: Lexical Density and Lexical Originality

Another indicator of text quality proposed in the literature is lexical density (Linnarud, 1986; Read, 2000). It refers to how many content words a text contains in relation to function words. It is expressed by the following formula.

$$\text{Lexical density} = \frac{\text{number of lexical words}}{\text{number of all words}} \times 100\%$$

In contrast to function words, which represent grammatical relationships, content (lexical) words are primary carriers of information. The more lexical words in a text and in consequence the higher its lexical density, the more information this text conveys. It should be pointed out that this lexical measure can indirectly tap structural complexity. The larger proportion of lexical words in text, the more likely it is that the text contains complex phrases, as it is demonstrated in the example:

She was beautiful and nice.
She was divinely beautiful.

The lexical density ratio was first proposed by Ure (1971), who demonstrated that the proportion of lexical words in writing is generally higher than in speech. This observation was explained by the differences in information structure in both modalities. Written texts tend to pack information more efficiently, so the same information is usually expressed in smaller number of words in writing.

Lexical density has been primarily used for measuring readability of texts, along with lexical sophistication. It has also been applied to distinguish between texts written by native and non-native speakers (Linnarud, 1975, 1977) and for measuring progress and differences in proficiency between foreign language learners (Linnarud, 1986; Laufer, 1991; Lu, 2012b). Such application of this measure was motivated by the assumption that function words are learned relatively early in the process of second language acquisition. The more advanced a learner becomes, the more lexical words his or her lexicon contains, which should manifest itself in an increased number of lexical words in his or her text. But the precise relation between lexical density and the size and complexity of the learner's lexicon is not clear. The studies which analysed the performance of this measure when applied to L2 learner production failed to demonstrate its ability to distinguish texts written by native speakers and/or learners at different proficiency levels.

Lexical density is independent of text length, but it is greatly affected by text type. For example, non-interactive texts, both written and spoken, show higher density than interactive ones (Ure, 1971). Thus, similarly to the gauges of lexical diversity and lexical sophistication, this measure has to be applied with caution when comparing results of learners engaged in different tasks. Its value can be calculated either with the help of a concordancer or other text analysis software such as *Range*, *AntWordProfiler* or *Lexical Complexity Analyzer*, provided that a pre-prepared list of function words is available.

Lexical originality, also referred to as lexical individuality, focuses on words a particular learner used in his or her text, which were not used by other learners within the group. It is expressed by the following formula:

$$\text{Lexical originality} = \frac{\text{number of words used exclusively by a learner}}{\text{number of all the other words}} \times 100\%$$

The relevance of lexical originality for assessing the size and complexity of a learner's lexicon is based on the assumption that a more advanced learner knows more words than a less advanced student, and this is reflected in the uniqueness of vocabulary the more advanced learner uses in his or her writing in relation to the less advanced learner. It can be pointed out that lexical originality conveys a comparable information to the gauge of lexical sophistication. If the order of acquisition of vocabulary depends, even roughly, on frequency, then the unique words the more advanced learner knows are most likely to be less frequent words.

The advantage of this method over the lexical sophistication ratio is that it defines less frequent words in relative terms, that is as words other learners do not know, and not as items beyond some arbitrary cut-off point (like 1000, 2000 or 5000 most frequent words of language). Thus, it allows for more subtle comparisons between learners of different proficiency levels. Yet, this advantage is at the same time a shortcoming of this measure, as a learner's result depends in equal way on his or her performance and on the performance of other learners engaged in the same task. Needless to say, this measure is also task-dependent and the results of learners engaged in different tasks, in different settings and analysed in relation to different learner groups cannot be compared (Laufer & Nation, 1995). This measure can be automated although at the moment there is no software dedicated to its calculation.

4.4.4 More Recent Measures: Word Psychological Properties and Semantic Relations

The measures previously discussed have been well established in the literature and research devoted to assessing foreign language vocabulary. However, in the last several years researchers have been engaged in finding new and more refined indicators of the growth of lexical knowledge in learners, based on learners' spontaneous (uncontrolled) production. These new measures not only endeavour to describe the L2 learner's lexical proficiency, and by extension the size of his or her lexicon, but also attempt to tap this lexicon's conceptual levels, the semantic properties of its elements as well as the intricacy of its network, that is the number and the quality of links between the items in the L2 word store (Crossley et al., 2009, 2010a, 2010b; Salsbury, Crossley, & McNamara, 2011). Recent studies have explored the applicability of psycholinguistic measures of various word qualities as well as measures of textual cohesion and coherence for tracing the increase in L2 learners' lexical proficiency. So far, these explorations have not produced assessment methods of the kind described in the previous sections (Sections 4.4.1 and 4.4.2), yet the preliminary results are claimed to show some promise for their relevance.

Psycholinguistic experiments have revealed that native speakers' and L2 learners' performance in different lexical tasks (such as word recognition or pair associate learning) varies depending on words' linguistic attributes. A multitude of lexical qualities influencing these results have been singled out in such experiments. Some of these qualities pertain to the form of the word: its length in letters and syllables, its pronounceability or its similarity with other words. Some others are related to the words' use: frequency of occurrence, familiarity and age of acquisition. Several characteristics refer to the meaning aspects of a word such as its concreteness, imagability, ambiguity (polysemy), meaningfulness and specificity (hypernymy). Still other more specific attributes related to

meaning are, for example, emotionality, pleasantness, goodness or gender ladenness. Different behaviour in lexical tasks of words with different attributes can imply that such attributes influence underlying mechanisms in the representation, development and processing of words in the mental lexicon. Some of these qualities are hypothesised to indicate the size of the lexicon, some others to represent lexical access, still others to reflect a word's interrelatedness with other items in the mental word store (Crossley et al., 2011a). The qualities which have been foregrounded in recent research are defined next.

Word frequency, as an important attribute of a word has already been discussed in previous sections. This property reflects the word's rate of occurrence in language and by extension the rate with which a language user encounters this word. **Word familiarity** is also related to frequency, but in subjective terms. This attribute does not reflect corpus-based ratings of a lexical item but perceptions of adult native speakers related to the frequency with which they encounter a particular item. These perceptions do not always converge with objective norms, that are based on a very wide range of topics and text types. For example, the word *spoon* is not very frequent in objective frequency norms but scores high in familiarity ratings. Another word attribute related to the rate of occurrence is the **age of acquisition**, which reflects how early or late a native speaker believes to have learned a particular word. The word *duck* is learned much earlier than the word *business*, even though it is objectively much less frequent in language. The three properties: objective frequency, familiarity and age of acquisition are indicative of the ease of access to words, and of the strength of links in the mental lexicon. Words which are more frequent, more familiar or acquired earlier are generally observed to be processed quicker.

In the category of attributes related to meaning, the quality of **concreteness** designates the degree to which a word refers to a thing or a class of things that can be perceived by senses. For example, the word *table* is more concrete than the word *ability*. Concrete words are easier to remember and quicker to access than abstract words. The attribute strongly related to concreteness is **imagability**, which refers to the property of a word or concept reflecting how easy or difficult it is to visually or acoustically imagine. Words as *drink* or *sing* are easier to imagine than the word *become*. Although concreteness and imagability are strongly associated (r=0.883; Toglia & Battig, 1978), they are not identical qualities. Words such as *concert* or *gaiety* are considered fairly imaginable but not very concrete. Similarly to concrete words, lexical items characterised by high imagery are accessed and processed quicker. The next frequently researched attribute is word ambiguity, which refers to the number of meanings a word has. This lexical property is also known as **polysemy**. For example, the words *bank* and *lean* can refer to at least two different senses. Ambiguous words have been reported by a number of researchers

to be responded to faster than words with one meaning such as *food* and *yellow* (see for example, Hino, Lupker, & Pexman, 2002 for a review), which again suggests that they are processed quicker. The three qualities are also indicative of the semantic properties of the words in L2 lexicon and the depth of L2 vocabulary knowledge.

The next set of word attributes reflects the complexity of the language user's lexical network. The property of word **meaningfulness** refers to a degree to which a word is connected with other related lexical items. This property was first described by Noble (1952), who defined it as the number of associations a verbal stimulus can activate. An example of a meaningful word is the adjective *beautiful* and of words with low meaningfulness are adjectives *aural* or *copious*. Meaningfulness has been demonstrated to influence the rate of verbal learning (Noble, 1952). Words which have more connections with other items are learned and processed quicker. The next word characteristic, word specificity, also known as **hypernymy**, implies how specific the meaning of a word is and where the word is located in the hierarchy of more general items called hypernyms or superordinate words and more specific items called hyponyms or subordinate words (e.g. *animal—dog*). Hypernymic relations between words can go over several levels of specificity (e.g. *animal—dog—poodle*) and inherent in them is the idea of basic-level categories. A basic-level category is a word which is used most often to name and discuss an object (Brown, 1958) and the concept which encompasses the largest number of perceptually salient features distinguishing it from other categories at the same level. It is easier to tell a difference between *a dog* and *a cat* than between *a golden retriever* and *a labrador*. Basic-level categories form the reference point for defining superordinate and subordinate concepts. For instance, the word *car* is the basic-level category, and the words *sedan* or *Mercedes* are best defined as kinds of car, whereas the word *vehicle* is explained most adequately by enumerating its different examples, *a car* being one of them. Research in first and second language acquisition shows that the development of hypernymic relations among words is parallel to general cognitive development and to the growth of the mental lexicon. Basic-level categories are learned first, followed by superordinate categories and finally by subordinate categories (Berlin, Breedlove, & Raven, 1974).

The information on most of these qualities for individual words, such as familiarity, age of acquisition, concreteness and imagability, is gathered by directly probing language users. Large groups of respondents are asked to rate a particular quality of words presented to them. The judgements are usually made on a scale from 1 to 7. The final rating assigned to a particular word is the mean value of scores given to it by respondents. Lists of words together with their subjective ratings on different qualities are published as reference so as other researchers can find suitable verbal stimuli which meet specific criteria. The information on

some word properties is gathered not through direct rating tasks but in psycholinguistic experiments. For example, Noble (1952) proposed to measure word meaningfulness (m) as the mean number of associations that a verbal stimulus can evoke from language users in a fixed time interval. The information on word frequency, polysemy or hypernymy, on the other hand, is extracted from dictionaries or hierarchical wordlists. As already mentioned, the information on word frequency is also gathered from language corpora.

The ratings and scores on most of the qualities already discussed are available in the Medical Research Council (MRC) Psycholinguistic Database,[9] which provides information for up to 26 different properties for 150,837 words. The MRC database was assembled by merging a number of smaller databases of limited availability: Paivio (unpublished), Paivio, Yuille, and Madigan (1968), Toglia and Battig (1978) and Gilhooly and Logie (1980) among others. In addition to pulling together various lists in one place, the advantage of the MRC Database is that all the information is structured in a single machine readable dictionary (Wilson, 1988). The subjective ratings available in the MRC Database were converted to integer values between 100 and 700. Table 4.1 lists the values for different qualities of words used as examples in this section.

The measurements of different attributes discussed previously pertain to individual words in isolation and most of them are not gauges of the overall lexical content of a text in the same way as lexical diversity, sophistication, density or originality are. Yet, computing their mean values for all the words in a text may allow comparisons between vocabulary of different documents or essays. For example, a higher mean score for

Table 4.1 Information on Various Qualities for Selected Items Drawn From the MCR Psycholinguistic Database

WORD	*KFFRQ*	*FAM*	*AOA*	*CNC*	*IMG*	*CMEAN*
SPOON	6	612	186	614	584	425
DUCK	9	529	164	606	632	473
BUSINESS	392	563	436	389	441	461
TABLE	198	599	–	604	582	423
ABILITY	74	563	453	273	327	–
DRINK	82	628	211	549	553	488
SING	34	576	–	421	527	472
BECOME	361	603	–	–	259	–
CONCERT	39	520	386	252	578	–
GAIETY	8	464	–	275	546	–
BEAUTIFUL	127	609	–	393	532	617
AURAL	1	192	–	309	253	197
COPIOUS	1	213	–	273	304	244

concreteness indicates that a text uses more concrete words than another text with a lower mean score. That is why more recently these measures have been explored as gauges of L2 lexical proficiency. This idea was based on the assumption that for example abstract words, more specific words or words with fewer links with other items are learned later than more concrete, less specific or more meaningful words, as the following quotation proposes (see also Salsbury et al., 2011 for a discussion).

> The results from the current article provide strong evidence that psycholinguistic properties of words impact the learnability of those words. Specifically, words with higher scores for concreteness, imagability, and meaningfulness appear to carry lower learning burdens in L2 lexical acquisition and therefore tend to emerge earlier than words with lower scores on these same indices.
>
> (Salsbury et al., 2011, p. 15)

It has to be noted, however, that similarly to the measures described earlier, the values of these gauges may be influenced by a particular text type (Salsbury et al., 2011). For example, it seems intuitive that descriptions or narratives will contain more concrete and imaginable words than argumentative essays dealing with abstract topics.

A recent study by Crossley et al. (2010b) also explored the applicability of a measure which goes beyond the qualities of individual words and captures the way lexical items are used to create the meaning of the whole text. This measure expresses how words contribute to a text's deeper coherence by interacting with a reader's relevant world knowledge. This gauge is fairly complex as it needs to capture interferences between a text and the reader's mental model of the subject matter described in the text. It assumes that semantic similarity between text segments influences its coherence. However, the semantic similarity is not achieved by the use of lexis which is related semantically in an explicit way, for example by sharing the same stem. The relations between words are implicit and depend on the reader's general knowledge. This idea can be illustrated by the following sentence:

> He did not go to university. He hates studying.

The interpretation of semantic similarity between these two sentences depends on the reader's understanding that studying is the essential part of going to university. As this knowledge is implicit in the text, it is difficult to capture it statistically. One possible way of encapsulating such interferences is by applying the computational model derived from the fields of artificial intelligence and natural language processing called **Latent Semantic Analysis** (LSA, Landauer, Foltz, & Laham, 1998). It is a method for extracting and representing the meaning of words in

context by statistical computations performed on a large corpus of texts. It is based on the premise that the similarity of meanings of words or word sequences (texts or text segments) is the derivative of all contexts in which given words do (and do not) occur. Two words are closely related in meaning if they always appear in the same context and they do not appear separately. This idea is expressed by the following quotation:

> Another way to think of this is that LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains.
>
> (Landauer et al., 1998, p. 6)

Unlike collocation which is also a co-occurrence phenomenon, LSA is much more complex and based on a different context. Whereas collocation is a repeated incidence of two lexical items in the immediate vicinity defined in terms of the number words to the left and/or right, LSA measures co-occurrence of words within sentences, paragraphs or entire texts and the relationships discovered within this framework represent much broader semantic associations between words. LSA is believed to be not only a measure of semantic relatedness but also a model of the computational processes and representations underlying substantial portions of the acquisition and utilisation of knowledge in general and the mental lexicon in particular (Landauer et al., 1998).

LSA has been demonstrated to model human conceptual knowledge. Word and text meaning representations derived by LSA have been demonstrated as capable of simulating a variety of human cognitive phenomena such as human scores on standard vocabulary and subject matter tests, human word sorting and category judgements; lexical priming data and children's word learning rate. LSA can be used to evaluate text coherence, learnability of texts by individual students, and the quality and quantity of knowledge contained in an essay (Landauer et al., 1998). Crossley et al. (2010b) have also postulated that the increase in LSA values of semantic similarity of text segments produced by L2 learners can reflect their lexical growth in terms of both size and interrelatedness of items in the L2 lexicon and can also be used as a measure of lexical proficiency. However, its quantification requires very complex computing and large bodies of texts in registers and styles comparable to the analysed samples of the learner productions, thus its calculation is still impossible for many genres of student writing and speech.

All the measures discussed in this section are calculated for any text by the recently developed computer tool *Coh-Metrix* (Graesser et al., 2004; McNamara, Louwerse, Cai, & Graesser, 2005). Although the primary objective in the creation of *Coh-Metrix* was the development a better gauge of text readability which would produce more accurate

information on a difficulty of a text based on many different linguistic and discourse gauges, its creators had also another objective. They also aimed at creating a linguistic workbench capable of providing a wide array of measures on language and discourse within one tool (McNamara & Graesser, 2011). The public version of *Coh-Metrix* produces 108 different frequencies, ratios and indices which tap the following characteristics of a text: simple descriptive statistics, text easability, referential cohesion, latent semantic analysis (LSA), lexical diversity, use of connectives, situational model, syntactic complexity, syntactic pattern density and word information, which all contribute to the final score on text readability.

Among word measures provided by *Coh-Metrix*, there are gauges of different lexical qualities discussed earlier, which take the form of mean values calculated for all content words in a text. *Coh-Metrix* provides mean scores for familiarity, age of acquisition, concreteness, imagability and meaningfulness. All these measures are calculated using the information available for individual words in the MRC database. In addition, *Coh-Metrix* provides mean scores for frequency and polysemy, as well as mean scores for hypernymy (for nouns and verbs separately and jointly). Average scores on polysemy and hypernymy are based on information available in WordNet (Miller, 1995; Fellbaum, 1998), a large lexical database of English. *Coh-Metrix* reports a mean number of senses for all content words in a text. The hierarchical structure of sysnsets in WordNet, on the other hand, allows for measuring the number of superordinate words above and subordinate words below a target word. *Coh-Metrix* reports an average level in the super-subordinate hierarchy for all context words. A lower mean value for a text reflects an overall use of more general words, whereas a higher value indicates an overall use of more specific words (McNamara, Graesser, McCarthy, & Cai, 2014; Crossley et al., 2010b). *Coh-Metrix* also reports scores for semantic similarity of fragments of a text based on Latent Semantic Analysis. LSA similarity scores are computed for all pairs of adjacent sentences in a text, all possible pairs of sentences in each paragraph and for adjacent paragraphs.

One of the veins of research in validating *Coh-Metrix* and exploring its potential for various applications, is the application of the tool for assessing text produced by L2 English learners and thus inferring information on the size, depth, network complexity and growth of their lexicons (Crossley et al., 2011a). The details of these studies and their results will be presented in Chapter 5. Although the preliminary results demonstrate that the information on the linguistic qualities of words used by L2 learners in their written output can shed some light on their L2 proficiency, there has been so far no attempt to turn these ratings into formulas of the kind described in the previous sections which would describe the lexical quality of a text.

4.4.5 Phraseological Measures

All the measures of lexical complexity discussed so far focused on individual words. The quality of lexical texture in all these approaches was viewed mostly in terms of the repetitiveness/variability or inherent properties of individual words,[10] and disregarded one important feature of lexis which refers to its collocability. This is a reflection of the trend predominant for a long time in linguistics which regarded the mental lexicon as a repository of words with no internal structure and connections between its component items. If phrasal characteristics of words were recognised in lexical measures at all, this was only in the gauges of accuracy, where erroneous combinations were recognised as one possible type of lexical error (e.g. Linnarud, 1986; Engber, 1995).

In the last 20 years the use of phraseology by language learners has been studied extensively within the field of second language acquisition, mainly through the analysis of learner corpora. Two types of phraseological units were mostly examined: collocations (e.g. Lorenz, 1999; Nesselhauf, 2005; Laufer & Waldman, 2011) and lexical bundles (e.g. De Cock, 1998, 2000; Groom, 2009; Reppen, 2009); usually in the production of advanced learners, but some studies also targeted lower-level students (Vidakovic & Baker, 2010; Leńko-Szymańska, 2014). A lot of attention was also devoted to academic phraseology (Howarth, 1996; Ädel & Erman, 2012; Chen & Baker, 2010; Chen & Baker, 2016; Salazar, 2014). All these studies demonstrate "a complex picture of overuse, underuse, misuse and use of idiosyncratic sequences, which may well play a significant part in the foreign-soundingness of [learners'] speech and writing" (De Cock, 2000, p. 65).

However, these studies did not analyse the use of phraseology in individual acts of production but collapsed them in a corpus or subcorpora representing different proficiency levels and examined them holistically. Such approach has two weakness. First, the learners' levels were usually not determined carefully, thus a corpus or a subcorpus could contain the production of students which were not completely uniform in terms of proficiency (Carlsen, 2012). Second, it ignored individual variation in the use of multi-word units which may exist between learners at the same proficiency levels. In addition, these examinations were based on simple counts of phraseological units of different kinds, and used their frequency in a corpus as their only analysed property, without further scrutiny of the strength of co-occurrence of the retrieved items (Crossley & Salsbury, 2011).

Durrant and Schmitt (2009) were among the first researchers who tried to overcome these shortcomings. They analysed the use of phraseology in the written output of individual learners and native speakers. The researchers focused on directly adjacent pre-modifier-noun word pairs (including both adjective noun and noun-noun combinations).

However, their analysis went beyond mere comparisons of numbers of these units in texts written by different groups of L2 learners and native speakers. The researchers applied two statistical association measures introduced in the literature: Mutual Information and t-score to further examine the collocations found in the production of individual informants.

Mutual Information is an association measure derived from the field of information theory and applied in linguistic analysis by Church and Hanks (1990). This measure compares the observed frequency of two words occurring together in a corpus with the frequency which could be expected if this co-occurrence was only due to chance. The expected frequency is estimated by the following formula:

$$E_{w1,\,w2} = \frac{f_{w1} f_{w2}}{N}$$

The formula for computing Mutual Information is a log transformation of a ratio between the observed and expected frequencies (Evert, 2004, 2008):

$$MI_{w1,\,w2} = \log \frac{O_{w1,\,w2}}{E_{w1,\,w2}}$$

According to Church and Hanks (1990, p. 24) MI ratio becomes unstable when the frequencies are very low so they propose to discard the combinations that occur five or fewer times in a corpus. They also suggest that the MI value above 3 can be taken as indicative of a collocation.

Another association measure proposed in the literature is t-score (Church & Hanks, 1990). Unlike MI, which gauges the effect size, that is the strength of association between two words, this measure gauges the confidence of the existence of an association, that is the amount of available evidence for a positive association between two words (Evert, 2004, 2008). It gives priority to recurrent combinations made of frequent of words whose co-occurrence is higher than chance. T-score is a statistical hypothesis test and it is computed by the following equation (Evert, 2004, 2008):

$$t_{w1,\,w2} = \frac{O_{w1w2} - E_{w1,w2}}{\sqrt{O_{w1,\,w2}}}$$

Research using t-scores for retrieving collocations adopts very conservative confidence levels—p=0.005—for rejecting the null hypothesis that there is no association between two items. The critical value of t for this significance level equals 2.576, thus if the result of the aforementioned equation is higher than this value, it can be stated with 99.5% confidence level that the two words for which it was calculated are positively

associated. Many studies accept a slightly lower cut-off value of t>2 for two words to qualify as a collocation (Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Bestgen & Granger, 2014).

The application of these two different measures results in a retrieval of different types of word combinations. T-score gives prominence to collocations made of relatively frequent words, such as *post office* whereas MI favours word combinations where at least one of the components occurs rarely on its own, such as *post mortem*.

Durrant and Schmitt (2009) divided retrieved collocation into several bands based on their MI values and t-scores and compared the frequencies of phraseological units belonging to different bands in individual texts written by native speakers and L2 learners. Inspired by this analysis, Granger and Bestgen (2014) conducted a similar investigation, yet based on all bigrams retrieved from a text. They collapsed several collocational bands proposed by Durrant and Schmitt (2009) into four: items below the collocational threshold, and three bands grouping items with low, middle and high collocational strength. These bands were produced separately for MI values and t-scores. Bestgen and Granger (2014, p. 28) went a step further and proposed CollGram, "a technique that assigns to each pair of contiguous words (bigrams) in a learner text two association scores (mutual information and t-score) computed on the basis of a large reference corpus". CollGram produces three measures which together form a CollGram profile and which according to the authors, "quantify the collocation strength of each text" (p. 31):

- the mean MI value for all the bigrams in a text;
- the mean t-score for all the bigrams in a text;
- the proportion of idiosyncratic bigrams that are absent from the reference corpus and thus cannot be assigned any association score

The two association indices gauge the use of different types of collocations. A mean MI value reflects the use of rare and rather salient collocations made of infrequent words. The higher it is, the larger number of these collocations can be found in a text. The mean t-score, on the other hand, taps the use of more common word combinations, which are less noticeable individually, but which contribute to the overall perception of natural language flow. The two indices can be computed based on bigram tokens and types, the latter giving more weight to the diversity of word combinations. The status of idiosyncratic bigrams is unclear. They can be instances of errors, but they can also be a result of a creative use of language.

So far the CollGram has been applied by its authors in one study, whose aim was to analyse the L2 learners' production (Bestgen & Granger, 2014). This study is discussed in detail in Chapter 5. The results suggest that with a growing proficiency the mean t-score decreases and

the mean MI value increases. However, more research is still needed in order to validate these two indices as gauges of lexical quality of a text.

4.5 Conclusion

Although researchers generally agree on various components of (lexical) proficiency, they keep searching for appropriate measures which will tap this proficiency and its components adequately. The search for text-based statistical gauges of L2 lexis has intensified in the last ten years with the advances in corpus and computational linguistics. In order to validate the proposed indices, the researchers compare their results with other types of assessment of learners' (lexical) proficiency. A selection of such validation studies will be discussed in the following chapter.

In addition to treating the proposed gauges as measures of lexical proficiency, researchers often make far-fetching claims about how they tap various aspects of lexical competence and L2 lexicon. The question to what extent this is the right assumption will be explored in Chapter 6. It presents the results of a study analysing a whole host of various statistical measures of lexical proficiency applied to texts written by advanced learners of English at different levels and native speakers of English.

Notes

1. www.personal.psu.edu/xxl13/downloads/lca.html (accessed 22 December 2018)
2. http://dali.talkbank.org/clan/ (accessed 22 December 2018)
3. www.lognostics.co.uk/tools/ (accessed 22 December 2018)
4. https://umdrive.memphis.edu/pmmccrth/public/software/software_index.htm (accessed 22 December 2018)
5. www.lextutor.ca/vp/ (accessed 22 December 2018)
6. www.lognostics.co.uk/tools/ (accessed 22 December 2018)
7. www.lexile.com (accessed 22 December 2018)
8. www.jasnopis.pl (accessed 22 December 2018)
9. http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm (accessed 22 December 2018)
10. With the exception of the measures based on Latent Semantic Analysis (see Section 4.4.4).

5 Statistical Measures and Raters' Scores of L2 Production—Review of Literature

5.1 Introduction

The previous chapter introduced a range of measures proposed in the literature as tools for measuring L2 lexical proficiency. These tools were validated in numerous studies, which demonstrated that the values of these indices differed between native speakers and L2 learners, or between learners at different levels, or that they correlated with results produced by an independent instrument tapping more directly L2 learners' lexical competence (a vocabulary test). However, it is becoming increasingly obvious that each of these measures alone cannot represent fully the multifaceted nature of vocabulary ability and thus cannot be used as a single yardstick of the changes which happen during the development of L2 lexical proficiency. Thus, at present many researchers use a combination of various lexical measures and test how these measures taken together can tap into the growth in L2 learners' lexical proficiency. Researchers are also interested in the relationship between lexical proficiency and more general writing or speaking proficiency. Therefore, not individual measures, but several measures taken together are now used as instruments which make it possible to objectively assess and quantify gains in L2 linguistic ability and communicative competence. Their joint effect is validated in the same manner as for individual gauges. An even more ambitious research agenda involves probing to what extent these indices applied to texts written by L2 learners and native speakers can predict scores assigned to these texts by human raters. The first section reviews the studies whose aim was to validate the proposed indices by testing if they discriminate between different kinds of learners and native speakers. The subsequent parts of the chapter focuses on the relationship between lexical indices and human raters' scores.

5.2 Lexical Measures for Discriminating Between Different Proficiency Levels

One of the earliest studies employing several lexical measures to texts written by L2 learners was **Arnaud (1984)**. His aim was not so much to investigate the measures, but to validate a new vocabulary translation test

as a yardstick of L2 lexical proficiency. This test was administered to a group of 100 first-year students of English at a French university. A week later the same students were asked to write an essay. Three indices of lexical proficiency were produced for each essay: lexical diversity (referred to by Arnaud as lexical variation), lexical sophistication (referred to as rareness) and the number of lexical errors. Arnaud observed that several indices of lexical variation remained influenced by text length in the analysed range of 180 to 600 words. Thus, the calculations of lexical diversity and of error rate were based on the length of the shortest essay. A random sample of 180 words was drawn from each essay and two indices were produced for it: the number of lexical types and the number of lexical errors. The index of lexical sophistication applied by Arnaud was the number of rare types divided by the number of all lexical types. Rare types were defined as those not belonging to the list of 1522 items published by the French Ministry of Education as a target for lower-secondary (middle) school education in English.

Arnaud correlated the three indices of lexical proficiency with the test scores as well as with each other. His results demonstrated that the test scores correlated significantly but weakly with the indices of lexical diversity and lexical errors ($r=0.36$, $p<0.01$ and $r=-0.21$, $p<0.025$, respectively) but they did not correlate with the index of lexical sophistication ($r=0.09$, n.s.). Arnaud concluded that the low correlations can still be interpreted as confirming the validity of the discrete-item vocabulary test, as (1) solving the test and writing the essay represented different measurement situations, (2) the students' texts were considerably interfered with (shortened, edited and interpreted in the case of errors) and (3) the students formed a very homogenous group which by itself lowered the results of the statistical analysis. The author also noted weak but significant correlations between the three indices (lexical variation and lexical rareness $r=0.27$, $p<0.01$, as well as lexical variation and the number of errors $r=-0.24$, $p<0.01$); however, he emphasised that the concurrent validity of the three chosen measures of lexical proficiency was not examined in his study and more research is needed in this area.

Another early study which employed several indices to investigate the lexical quality of EFL learners' written production was conducted by **Linnarud (1986)**. She collected 42 essays written by Swedish secondary school students and 21 essays written by native speakers, equivalent to the Swedish students in terms of age and education. Both groups wrote a narrative on the same topic. For each essay Linnarud produced several measures: the number of words, the average number of words in a sentence, the number of errors per essay, the percentage of errors, lexical originality, lexical sophistication, lexical diversity (variation) and lexical density. Errors were defined as belonging to three categories: spelling, grammatical and lexical, but they were totalled as one category. Lexical variation was calculated based on lexical words and it was not adjusted

for text length. Lexical sophistication was defined as a percentage of lexical items belonging to a list of advanced words used in the Swedish national EFL education system.

The comparison of the essays written by the two groups demonstrated that the native speakers scored higher on most of the analysed indices: number of words, lexical originality, lexical sophistication and lexical variation (even though this index was not adjusted for essay length and NS wrote longer essays). All the differences were statistically significant ($p<0.05$). The measures which did not demonstrate a statistically significant difference were the average number of words per sentence and lexical density. For obvious reasons, errors were not taken into account in this comparison.

Linnarud also checked if the different measures she had produced for each essay correlated with each other. She was particularly interested to see if longer texts demonstrated smaller numbers of errors and higher values of the lexical indices. Even though she did not say this explicitly, such a correlation would indicate that the three components of lexical proficiency—fluency, accuracy and complexity—develop in a parallel way. The results of the study demonstrated that the number of words in learner essays indeed correlated with lexical originality ($r=0.49$, $p<0.01$) but at the same time the students who wrote longer texts made more errors, as expressed by raw numbers but not by percentages, ($r=0.44$, $p<0.01$) and used less varied vocabulary ($r=-0.37$, $p<0.05$). Interestingly, except for lexical originality and lexical sophistication, the other lexical measures did not demonstrate statistically significant correlations between each other. In the case of native speakers, the only variables which correlated moderately were lexical sophistication and lexical variation ($r=0.51$, $p=0.02$), which indicates that the writers who used more varied vocabulary achieved this effect by using more advanced words.

The four lexical measures discussed by Linnarud (1986) were also employed by **Laufer (1991)**, but the aim of her study was to trace the progress of advanced learners of English. Laufer's specific goal was to investigate whether learners whose vocabulary is sufficient to meet all their communicative needs still develop their lexical command as a result of extensive exposure to comprehensible input, but without explicit vocabulary instruction. She also wanted to check to what extent such a growth, if it indeed exists, depends on learners' general proficiency. In order to answer her research questions, Laufer collected 47 essays written by Israeli students as part of an entrance exam to an English department at an Israeli university. 22 of these students, referred to as Group One, were required to write another essay on the same topic after one semester. The remaining 25 students, forming Group Two, were given the same task but after two semesters of studies.

Laufer produced four indices for each essay: type/token ratio based on lexemes and calculated for the first 250 words of each text, as a measure

of lexical variation; the percentage of lexemes belonging to the University Word List containing over 1400 items frequent in academic discourse, as a measure of lexical sophistication; the percentage of lexemes unique to the essay, as a measure of lexical originality; and the percentage of lexical words, as a measure of lexical density. Group means for each of the four indices were compared between the first and the second writing sessions. The results demonstrated that there were no statistical differences for all the four indices in the case of one-semester interval. In the case of two-semester span, the only measure which displayed a statistically significant growth was lexical sophistication. Laufer concluded that throughout the year students had been required to write many academic papers, which had resulted in an increase of their active academic vocabulary. However, the researcher also studied the change in the values of the four indices for individual students. She observed that some of the essay pairs in fact demonstrated a considerable increase in the index values after two and even one semester. She also discovered that in all but one of these cases such a growth was noted for the students whose results at the entrance exams had been below the average. Due to this lexical development these students had managed to catch up with the group means at the second writing session. Laufer proposed that the students whose vocabulary had reached an "active vocabulary threshold" (p. 445) and who could meet all their academic needs in a satisfactory way had not made any progress in their active lexical command as a result of extensive comprehensible input. At the same time the active vocabulary of the students whose level had been insufficient for the kinds of tasks they were required to perform did develop throughout one or two semesters to reach this threshold.

Laufer (1994) replicated her 1991 study using a more refined measure of lexical sophistication—the Lexical Frequency Profile (LFP, cf. Section 4.4.2). She collected 48 pairs of L2 learners' essays following the same procedure as previously, i.e. with one- and two-semester intervals. For each pre-processed text, Laufer computed a profile consisting of the percentage of types belonging to the first and second 1000 most frequent English word families, the University Word List and the 'not in any list' category. She also collapsed the percentages for the first and second 1000-word frequency bands, as well as for the University Word List and the 'not-in-any-list' category thus producing the Condensed Lexical Frequency Profile (CLFP) representing the proportions of basic and advanced words in each essay. In addition, she calculated a type/token ratio for each composition, yet she did not specify if this time she had taken text length into account.

The comparisons of the essays' LFPs across time, performed with paired t-tests and MANOVA, demonstrated a decrease in the mean percentages for the 1000 and 2000 levels and an increase in the mean proportions of UWL and 'not in any list' words. These changes, however, were not statistically significant for all but the UWL level and the

individual fluctuations presented a rather complex pattern, which was hard to interpret. On the other hand, the comparisons of the CLFP values across time, performed with paired t-tests, revealed a statistically significant decrease in the mean use of basic words and a statistically significant increase in the mean use of advanced words, both after one and two semesters. At the same time the comparisons of the mean TTR values did not demonstrate a similar effect either after one or two semesters. In addition, no statistically significant correlations were observed between TTRs and any of the LFP and CLFP levels. Laufer concluded that while the comparisons of CLFPs demonstrated an increase in the quantity of the L2 learners' word knowledge throughout a year, a juxtaposition of these profiles with native speakers' results (analysed in a different study) indicated that the growth in the learners' L2 lexicons was not sufficient and they still had a long way to go before their word stores reached a native-like size. Comparisons of the TTR values for the learners and native speakers also did not produce satisfactory results. Commenting on the usefulness of the two measures of lexical richness for measuring learners' lexical command, Laufer observed:

> This provides empirical evidence for the statement made earlier that lexical variation shows how well a person can express himself with whatever vocabulary he has, not whether his vocabulary size is large or small. Since the vocabulary quality of a piece of writing depends on the type of words used and also on an effective way of varying these words, both measures of lexical richness (lexical profile and lexical variation) seem to be necessary in the assessment of writing.
>
> (Laufer, 1994, p. 30)

Thus, according to Laufer (1994), while lexical sophistication—as measured by Condensed Lexical Frequency Profile—is a good indication of an L2 learner's vocabulary knowledge, lexical variation is a measure of the lexical quality of his or her writing.

Comparisons of the two measures of lexical complexity—lexical diversity and lexical sophistication—and various less popular indices representing these dimensions of lexical proficiency—were carried out in two studies published in 2007. Their aim was to investigate how various indices computed for oral production can discriminate between groups of learners at different proficiency levels. The studies employed a more sophisticated statistical instrument (effect size) in order to compare the performance of the analysed indices.

Daller and Xue (2007) compared oral picture descriptions produced by two groups of Chinese advanced learners of English: 24 university TEFL students in China and 26 university students studying in the UK. The latter group was more proficient, as confirmed by the results of a C-test. The researchers computed the following indices: type/token ratio,

Guiraud Index and D as measures of lexical diversity and Condensed Lexical Frequency Profile (beyond 2000), Advanced Guiraud and P-Lex as gauges of lexical sophistication (see Chapter 4 for a discussion of these measures). The results revealed that four indices demonstrated statistically significant differences between the groups: Guiraud Index, D, LFP and Advanced Guiraud, with the UK group achieving on average higher scores. Such findings confirmed the researchers' hypothesis that oral production of more advanced learners would be characterised by richer vocabulary. The researchers also analysed and compared the performance of the six measures with the measure of an effect size—eta squared, which is an indication of how well the metric differentiated between the groups. Its values indicated that the indices which discriminated best between the two groups of different proficiency was the Guiraud Index followed by D (η^2=0.442 and η^2=0.250, respectively). Not surprisingly the values for TTR and P-Lex were the lowest, as these metrics did not produce significant results in the first place. In addition, the UK group generated on average longer descriptions; the difference, however, was not statistically significant. The juxtaposition of the number of tokens and the values of the six indices for each description indicated that four metrics produced positive and significant correlations: Guiraud Index, D, LFP and Advanced Guiraud. The researchers concluded that more proficient learners, who generated longer descriptions, used more varied and sophisticated vocabulary. As expected, the correlation for TTR was negative.

Finally, the researchers correlated the six values between each other. The findings indicated that only two measures of lexical variation: Guiraud Index, and D, as well as two measures of lexical diversity, LFP and Advanced Guiraud, showed strong, positive and significant correlations between each other ($r=0.89$ and $r=0.78$, $p<0.01$), indicating that these two measures tap into different aspects of lexical proficiency. Daller and Xue also remarked that the values of lexical sophistication indices depend strongly on the lists used for their calculations. Most studies employ lists based on written language but such lists may not be suitable for the analysis of learners' spoken production. In the introduction to their paper, Daller and Xue acknowledge that instead of looking for one perfect measure and index of lexical richness, each time researchers should choose metrics which are most appropriate for their purpose and their data. Thus, the authors interpreted the findings of their study as indicating that in their research context the list-free indices: Guiraud Index and D, were more appropriate. These indices are independent of any arbitrary list of basic and sophisticated words, usually compiled from written corpora, thus producing less meaningful results when applied to oral production.

An analogous study was conducted by **Tidball and Treffers-Daller (2007)**, who also analysed the application of the two measures of lexical

proficiency: sophistication and diversity—and various indices tapping into these measures—for describing differences among oral picture descriptions produced by participants at different proficiency levels. The originality of this study lay in proposing a new metric of lexical sophistication, as well as in applying various indices of lexical proficiency to a language other than English—French[1]. Tidball and Treffers-Daller collected oral samples from two groups of Year 1 and Year 3 students studying French at a British university (consisting of 24 and 16 students respectively) and a group of 23 French native speakers. All the participants took the same C-test, which confirmed statistically significant differences in proficiency between the three groups (Year 1 < Year 3 < Natives). The transcripts of the subjects' oral production were analysed with two metrics of lexical diversity and two indices of lexical sophistication. In order to avoid inflated results due to rich inflectional morphology in French, Guiraud Index and D were calculated on the basis of lexemes rather than types, as was the case in Daller and Xue's (2007) study. The metrics of lexical sophistication applied by the researches were Advanced Guiraud and a new index, which was a version of Limiting Relative Diversity (LRD), first proposed by Malvern et al. (2004). Their formula applied the following algorithm:

$$\mathrm{LRD}(\text{basic/all}) = 1 - \sqrt{\frac{D(basic)}{D(all)}}$$

The inventory of basic words for computing the two indices of lexical sophistication was the list Francaise Fondamental Premier Degré (Gougenheim, 1964), containing 1445 lexemes. Although rather dated, this list had the advantage of being compiled based on frequencies of words in an oral corpus, and thus it was judged most suitable for the study.

The results of the comparisons between the oral descriptions produced by the three groups showed that significant differences could be found for all the four indices. Yet, a post hoc analysis revealed that while all the metrics reported a significant difference between Year 1 students and native speakers as well as Year 3 students and native speakers, only two indices of lexical diversity demonstrated significant differences between Year 1 and Year 3 subjects. The difference for Advanced Guiraud was only marginally significant and in the case of LRD a statistically significant difference was not recorded. As in the case of Daller and Xue's (2007) study, the values of eta squared were computed for each variable. The values indicated that all the four indices produced comparable effect sizes ($0.61<\eta^2<0.67$). Analogous to Daller and Xue (2007), Tidball and Treffers-Daller also computed correlations between the four indices. They were all very high ($0.72<r<0.94$, $p<0.01$). The researchers also produced correlations between each of the fours indices and the C-test scores ($0.73<r<0.76$, $p<0.01$).

The researchers concluded that the four metrics analysed in the study—Guiraud Index, D, Advanced Guiraud and LRD differentiated between the participants with varying proficiencies in French, yet the proposed new metric of lexical sophistication performed less well in this respect. In addition, the really high correlations between the four measures and the C-test scores confirmed that the examined metrics are all valid gauges of lexical proficiency. These findings are not identical with the conclusions drawn by Daller and Xue (2007), who pointed to lexical variation measures as more suitable for analysing learners' oral production. The researchers attributed this discrepancy in the outcomes of the two studies to the characteristics of the wordlists used for the calculation of the list-based metrics. The inventory applied by Tidball and Treffers-Daller (2007) was more relevant, even if not ideal, for the analysis of oral data, and thus the results of the measures of lexical sophistication were more valid in their study.

Crossley, Salsbury and McNamara (2009); Crossley et al. (2010a, 2010b) and Salsbury et al. (2011) conducted a series of studies to explore the applicability of various indices computed by *Coh-Metrix*—the tool described in detail in Chapter 4—for accounting for the growth of lexical proficiency in L2 learners. Their research was based on spoken longitudinal data elicited from six foreign students taking a year-long English course at an American University. Eighteen sessions of unstructured interviews were carried out with these learners at two-week intervals. The resulting corpus consisted of 99 transcripts, with mean length for each learner ranging from 1121 to 2360 words. Various lexical, syntactic and discourse measures were computed for each text and variations in their values over the course of time were analysed statistically and linked to different aspects of the development of the L2 mental lexicon. A repeated-measures analysis of variance (ANOVA) for selected variables was computed only on trimester bases with six data collection points, that is for 36 texts, and all 99 texts were used for a separate growth modelling analysis (a linear curve estimation). The students were also administered the *TOEFL* test at six points of time, but only the results of four tests were analysed statistically.

Crossley et al. (2009) focused on the development of hypernymic relations in L2 learners' output. They hypothesised that mean hypernymy values should decrease over the time since lower values would indicate that students employed more superordinate categories in their speech. Changes in two more indices—the measure of lexical diversity (MTLD) and the mean value for concreteness—were also analysed as control variables. The results showed that MTLD increased over the course of time thus confirming a significant growth in the size of the learners' lexicons. At the same time mean hypernymy values demonstrated a significant decrease thus confirming the initial hypothesis. The hypernymy values were positively correlated with the concreteness index which proved

that an amplified use of superordinate words was paralleled by a higher abstractness of students' vocabulary. They were also negatively correlated with the L2 learners' time spent studying English and the MTLD values. Thus, this study has confirmed the assumption that as time studying an L2 increases, there is a corresponding increase in the range of hypernymy levels available to L2 learners. It has demonstrated that hypernymy growth is related to a growth in the size of the L2 lexicon.

A similar analysis was conducted for mean polysemy and lexical frequency values (**Crossley et al., 2010a**). The researchers were interested how polysemy relations develop in L2 learners over a course of time. The results demonstrated a significant growth of mean polysemy and word frequency values. However, further pairwise comparisons revealed that this growth occurred only in the first trimester and then levelled off in the next trimesters. The correlation between the L2 learners' time spent learning English and their WordNet polysemy values were non-significant and the same analysis for word frequency values showed a positive, but weak correlation. The correlations between polysemy and word frequency values were positive, which proves that more frequent words are more polysemous. The researchers also analysed the participants' scores on the *TOEFL* test, which showed a significant increase, thus confirming that the learners' general language proficiency developed over a year-long study. The quantitative study was followed by a qualitative analysis whose aim was to investigate if and when learners actually started using different senses of polysemous words. For the purpose of this study the corpus was divided into two parts: one containing transcripts from the first trimester, and the other—the interviews from the second and third trimesters. The researches focused on six highly polysemous words (with more than ten senses listed in WordNet) which appeared frequently in both sections of the corpus: *know*, *name*, *place*, *play*, *think* and *work*. The results of analyses of a random samples of concordance lines including these words provided evidence that there was indeed an increase in the production of different senses in the later trimesters.

This study has demonstrated that at the initial stage of language learning L2 students begin to produce words that are frequent and polysemous. However, learners' production is usually limited to the core meanings of the ambiguous words. While students' L2 proficiency grows, they extend their lexical knowledge to encompass less frequent and less polysemous items. At the same time they increase their command of different senses of the polysemous words whose core meanings they have already acquired. Thus, the study provides evidence not only for the development of L2 learners' lexical knowledge over a one-year study in a natural setting but also for the dynamics of the growth of their lexical networks.

Another study carried out by **Crossley et al. (2010b)** also concerned the growth of L2 semantic network. This time the researchers focused on the gauge of semantic co-reference of words as measured by LSA scores

(see Section 4.4.4 for details). The same analysis of the *TOEFL* scores and the MTLD values, as reported in the previous studies, were used to tap into learners' general linguistic and lexical proficiency and the size of their lexical command. Paragraph-based LSA scores, on the other hand, were taken to represent the complexity of learners' semantic networks, with higher values indicating more semantic overlap between adjacent utterances and thus more frequent use of semantically related items. The gauges of lexical overlap were used to ensure that the observed changes in LSA scores were not a result of mere increased repetitiveness of the same lexical items but a real growth of lexical networks in L2 learners' lexicons. The results demonstrated the growth in mean *TOEFL* scores, MTLD values and LSA gauges over the course of time. At the same time the overlap measures did not increase significantly. Crossley et al. concluded that while general proficiency develops and learners acquire more words, they also develop closer semantic similarities between speech segments which can be taken to reflect the development of learners' lexical networks. This means that the growth in the size of the mental lexicon is caused (at least partly) by the acquisition of more semantically related items and that meaning associations between words can facilitate lexical acquisition. The researchers also propose that LSA scores can be a good gauge of the development of lexical networks in L2 lexicon. However, they admit that more research is needed in order to validate this instrument, in particular studies based on more data which would also include L2 learners' written production. They also observe that using LSA as a model which mimics lexical learning may be a little premature as its learning mechanism is only an approximation of human learning, which has other sources of knowledge than context.

The fourth study conducted by the same research team (**Salsbury et al., 2011**) concerned word psycholinguistic information. The researchers analysed scores for word concreteness, imagability, meaningfulness and familiarity. They hypothesised that within the course of one year a decrease in the mean values for all the four properties would be observed. Such a decrease could indicate that as learners' lexicons grow in size, they encompass more words that are more abstract and less prone to arouse mental images, words that are less frequent in learners' input and words which have fewer connections with other words.

The statistical analysis indeed demonstrated a significant decrease in mean values, but only for three attributes: concreteness, imagability and meaningfulness. The familiarity index stayed at the same level throughout the data collections points. The researchers' interpretation of this last finding was that familiarity scores, which indicate the frequency with which words are encountered in input, are more relevant to the growth of receptive rather than productive knowledge. However, the remaining three results suggest that concreteness, imagability and meaningfulness play a role in the order in which learners acquire new words and in which

they are able to use them. According to the researchers, the study has also provided evidence that psycholinguistic information on word semantic properties can serve as a gauge of lexical development shedding light on the development of depth and access dimensions of the L2 word store and the growth of its lexical network.

The same three researchers carried out a more complex analysis of all the variables previously discussed. **Crossley, Salsbury and McNamara (2012)** endeavoured not only to barely trace to what extent various lexical indices change over time along the development of learners' general proficiency, but to explore if these measures can be employed to classify texts produced by L2 students at different proficiency levels. In other words, the aim of this new study was to demonstrate that a learner's level can be estimated based on a computational analysis of lexis in his or her output.

Unlike their previous studies discussed earlier, the researchers based this investigation on written texts produced by adult ESL students at two large American universities. One hundred students participating in the study were assigned to three broad proficiency levels: beginner, intermediate and advanced based on their scores on a standard test, with each proficiency level represented by a similar number of students. The students were assigned a free-writing task and a random selection of a 150-word segment was made from each text.

Twenty-four different measures were calculated for each text with *Coh-Metrix*. They represented gauges of:

- word frequency (six indices: mean frequency for all words and content words, based on spoken and written frequencies separately)
- lexical diversity (three indices: Maas, D, MTLD)
- polysemy (two indices: mean values and standard deviation for content words)
- hypernymy (three indices: mean values for nouns and verbs and for nouns and verbs combined)
- semantic co-referentiality LSA (two indices: LSA mean values between adjacent sentences and all sentences)
- word concreteness, imagability, familiarity and meaningfulness (eight indices: mean values for all words and content words)

These measures were classified by the researchers as representing three dimensions of lexical competence: breadth (lexical diversity and word frequency), depth (polysemy, hypernymy, semantic co-referentiality and meaningfulness) and access (word concreteness, imagability and familiarity).

An ANOVA test was carried out in order to trace differences between the three proficiency groups for all the variables. As a result of this analysis, the researchers singled out four unrelated variables which showed

the highest effect sizes when comparing the three proficiency levels: word imagability for every word, content word frequency based on written language, Maas lexical diversity and word familiarity for content words. Next, the four selected variables were entered into a discriminant function analysis performed on a training set of 77 texts. The results demonstrated that the model correctly classified 70.1% of texts into the three levels. All levels were distinguished with a similar accuracy rate, with advanced learners' texts noting the highest classification results. A further analysis of partial contribution of each variable to the discriminant function (the so-called discriminant function coefficient) revealed that the word imagability index contributed the most to distinguishing between the groups (DFC=0.773), followed by the Maas lexical diversity index (DFC=0.311), the content word frequency index (DFC=0.286) and the word familiarity index (DFC=0.224). The predictive power of the developed model was used on the remaining 33 texts. It classified correctly 69.7% texts into the three levels; however in this case the accuracy rates varied: the highest was noted for the beginner level, followed by the advanced and the intermediate levels (92%, 70% and 46% of texts classified correctly, respectively).

The results of this study demonstrated that lexical indices, in particular if applied jointly, can serve as a good indicator of general language proficiency determined independently by standardised tests. According to the researchers, the dimensions of lexical competence which are particularly indicative of the development of language proficiency are the breadth and accessibility features, with depth of knowledge characteristics not being informative in this respect.

Crossley et al.'s (2009, 2010a, 2010b, 2012) and Salsbury et al.'s (2011) findings are an important step forward in investigating the nature and dynamics of the growth of the L2 lexicon. They also offer an important argument for an automated method of assessing both lexical and general language proficiency. Yet, while the measures of lexical diversity and sophistication have been extensively researched and used in a variety of studies and for a variety of purposes, the indices related to the psycholinguistic properties of individual words in the lexicon and the structure of the L2 lexical network await further examination before they become well-established measures of lexical quality of L2 learners' production. The researchers' conclusion about the relative importance of the three dimensions of the L2 lexicon in a statistical model of lexical proficiency and in an automated assessment method also has to be treated with caution. The assignment of various variables into the three aspects of lexical competence is appealing, but to some extent arbitrary. The lexical variables proposed by the researchers can in fact be indicative of different aspects of the L2 lexicon. For example, the information on word frequency can be treated as an indicator of the size of the lexicon, based on the assumption that the order of acquisition of individual

words by L2 learners broadly follows the order of their frequency in language. Yet, as suggested by several studies in psycholinguistics discussed earlier, word frequency can also be suggestive of the accessibility of words in the lexicon. Frequent words are generally responded to quicker in psycholinguistic tasks, thus suggesting easier access. A similar ambiguity can be seen in the case of meaningfulness. This index represents the complexity of L2 learners' lexical networks and as such has been classified by the researchers as indicative of the depth of lexical competence. Yet, words which are within dense semantic networks are also accessed quicker, which again has been demonstrated in various psycholinguistic experiments. Thus, a neat classification of various lexical variables into three broad dimensions of L2 lexical competence is an oversimplification not fully supported by theoretical consideration and empirical evidence.

Since text-based measures of phraseology are the most recent development in the field, there are relatively fewer studies investigating their contribution to the measurement of lexical proficiency. **Granger and Bestgen (2014)** compared the use of collocations by learners at two different proficiency levels. They based their study on 223 texts drawn from the International Corpus of Learner English (ICLE). The essays were assessed and divided into two groups representing intermediate and advanced levels. The researchers produced a list of all bigrams for each essay. Next, they computed an MI value and a t-score for each bigram based on frequency information from the British National Corpus. They grouped the bigrams into four bands according to the strength of association; one grouping made according to MI and another according to t values, using the following thresholds for the groupings (Table 5.1).

The fifth category (BT, below the threshold) included the percentage of bigrams in a learner text that were either absent from the corpus or occurred only five times or fewer in it and thus the association measures could not be computed for them. The stratification of the five bands for each essay was performed twice: based on tokens and types. Thus, each essay was described by two sets of nine percentages. The percentages of

Table 5.1 Thresholds Applied to the Association Measures

Categories of bigrams	*MI*	*t*
Non-collocational	< 3	< 2
Collocational: Low	≥ 3 and < 5	≥ 2 and < 6
Collocational: Medium	≥ 5 and < 7	≥ 6 and < 10
Collocational: High	≥ 7	≥ 10

Source: Granger and Bestgen (2014, p. 236)

the four MI bands plus in the BT category added to 100%, and the same was true to the t-score.

Granger and Bestgen conducted a statistical analysis of the differences between the individual bands in essays produced by intermediate and advanced learners. The results confirmed the researchers' initial hypothesis that the intermediate learners' essays were characterised by a smaller proportion of strongly associated but infrequent collocations (identified by MI) and a larger proportion of frequent collocations (identified by t-score) in comparison with the advanced learners, and this trend was visible for both tokens and types. The differences were particularly large in the high association and non-collocational bands for both MI values and t-scores. In addition, the researchers observed that intermediate students used fewer below-threshold bigrams; however they did not comment on this finding. Granger and Bestgen concluded that:

> Unlike the analysis based on frequencies only, the analysis in terms of collocational strength reveals very different behaviour in intermediate and advanced learners. It shows that L2 phraseological competence is characterised by a mixture of high frequency and low frequency collocations. As proficiency in the language increases, the balance between the two types of units changes: the proportion of high scoring, high frequency sequences tends to decrease and that of high scoring but less frequent sequences to increase, but both types of units play an important role in language and continue to co-exist.
>
> (Granger & Bestgen, 2014, p. 240)

A different methodology was applied by **Bestgen and Granger (2014)**, who examined the development of phraseological competence in L2 learners over a period of time. They used 171 essays written by 57 international students at an American university. Each student contributed three essays written at the beginning, in the middle and at the end of one semester. The authors produced CollGram profiles for each text (cf. Section 4.4.5) which included mean MI values and mean t-scores computed for the bigrams retrieved from each text (based on tokens and types separately). In addition, the profiles featured the percentage of bigrams (both types and tokens) absent in the reference corpus. The reference corpus selected for this study was Contemporary Corpus of American English (COCA), as American English was the language variety the learners were exposed to.

In the longitudinal analysis, the means for paired samples were compared only between the essays written at the beginning and the end of the semester to maximise the effect of increasing proficiency. The results demonstrated a statistically significant difference only for t-score computed for both bigram tokens and types (tokens: $F(1, 56)=4.71$, $p<0.05$, $h2=0.08$; types: $F(1, 56)=6.53$, $p<0.05$, $h2=0.12$), but none of the other

measures showed a significant change over one semester. The observed change in the mean t-scores was a decrease. This replicated the authors' earlier results already discussed and was interpreted as a confirmation of the hypothesis that language acquisition starts with learning frequent chunks of language which at later stages are analysed into smaller units, i.e. words. The lack of change in the remaining indices was surprising as it went against earlier findings. The authors concluded that one semester is not sufficient to observe a growth in the phraseological competence, in particular related to the infrequent word combinations gauged by MI scores. Bestgen and Granger concluded that the CollGram profile is a promising instrument for measuring the development of phraseological competence of L2 learners.

The studies discussed previously are just a small sample of the extensive research on the metrics presented in the previous chapter. All of the earlier analyses prove that that various lexical indices proposed in the literature do indeed differentiate between L2 learners and native speakers or learners at different proficiency levels, as measured by the length of learning or independent tests. However, it needs to be acknowledged that the research in this area produces somewhat equivocal results. Certain projects support one set of indices as more valid and reliable instruments to tap the differences in proficiency levels, others advocate a different selection of measures. Meara (2005b, p. 35) points out that some of the indices (such as LFP) have quickly become common currency in SLA studies, which results in the fact that they have become taken for granted and not evaluated critically. He also remarks that research that does not produce significant results is harder to publish (which is a general problem of all experimental fields), thus studies which fail to demonstrate that widely accepted indices perform successfully are less likely to be disseminated. At the present moment it can be observed that no one index or a set of indices has as yet emerged as the most practical, valid and reliable gauge of lexical proficiency and lexical competence. As Crossley et al. (2012) note,

> the examination of lexical competence using these contemporary, computational indices is still in its infancy and the orchestration of these indices into a practical algorithm from which to investigate lexical competence has only just begun.
>
> (Crossley et al., 2012, p. 244)

The numerous studies reviewed in this section have also demonstrated that lexical proficiency develops in parallel to general linguistic proficiency and can serve as its predictor. The research discussed previously has also confirmed that various lexical measures, in particular if taken together, can tap well into changes in learners' global proficiency and reflect the growth of their vocabulary knowledge. This last finding has

important implications for language assessment. The following section will consider much more refined interdependencies between lexical scores and human rater's evaluations of the lexical texture and general quality of a text.

5.3 Lexical Measures vs. Raters' Scores

5.3.1 Correlational Studies

The first studies comparing human ratings and statistical measures of different aspects of language performance were conducted at the end of the 1970s and at the beginning of the 1980s. Inspired by the main trends in research on second language learning, teaching and assessment, these studies concentrated on measures of fluency, accuracy and syntactic complexity such as the number of words, the number of errors, mean sentence length or subordination (Larsen-Freeman & Strom, 1977; Flahive & Snow, 1980; Mullen, 1980; Perkins, 1980; Homburg, 1984). The only researcher who at this time was concerned with the relation between lexis and human perceptions of text quality was **Moira Linnarud (1975, 1986)**. In her 1986 study, already discussed in the previous section, she collected evaluations of 42 essays written by Swedish secondary school students and 21 texts produced by comparable native speakers of English. Each essay was assessed by 15 rates representing three different groups: Swedish secondary school English teachers, English-native-speaking university lectures working at Swedish universities and native speakers living in Great Britain with no TESOL experience. The raters were asked to evaluate the compositions on the scale from 1 to 5, but they were not provided with specific criteria or level descriptors, thus the assigned scores reflected the raters' intuitions rather than standardised assessment norms.

The results demonstrated that out of the lexical complexity measures, the index whose correlation with the raters' scores was highest was lexical originality ($r=0.47$, $p<0.01$). This feature seemed more relevant to both groups of native-speaking raters than to Swedish teachers. Interestingly, lexical sophistication, the measurement which—as also confirmed by Linnarud's results—overlaps with the index of lexical originality, did not produce statistically significant correlations. Statistically non-significant results could also be observed for lexical density and lexical variation; however the latter finding was not meaningful, since the type/token ratio was not adjusted for text length. Text length demonstrated moderate positive correlation with the scores ($r=0.51$, $p<0.01$), with little variation between the three groups of raters. The strongest correlation with the marks was demonstrated by the percentage of errors ($r=-0.77$, $p<0.01$) while raw numbers of errors produced a lower coefficient ($r=-0.47$, $p<0.01$). Native-speaking university teachers showed more

tolerance to errors than the other two rater groups. Interestingly, further analysis revealed only a low and non-significant correlation between lexical errors and the marks. Finally, mean sentence length did not produce statistically significant results.

Linnarud concluded that all the factors, both lexical and other, which discriminated between native and non-native writers (see the previous section) also correlated with the raters' evaluations, except for lexical sophistication, which did not produce statistically significant results in the latter analysis. The percentage of errors in the compositions also converged with the global ratings. Linnarud also hypothesised that lexical errors taken separately did not correlate significantly with the holistic evaluations, as they did not affect comprehension due to availability of wider context. A more detailed quantitative and qualitative analysis of the results led Linnarud to conclude that while accuracy and length combined distinguished between poor and average compositions, they were not sufficient to differentiate between average and very good essays. It was lexical originality and well as collocations and idiomatic language which separated "the merely competent writers from the truly successful" (p. 83).

A similar study was conducted by **Engberg (1995)**. Her aim was to establish to what extent raters take into account lexical richness and lexical errors when assigning global scorers to essays written by learners of English. The participants represented lower and more varied proficiency in comparison with Linnarud's investigation. Sixty-six timed argumentative essays written on the same topic by learners with diverse L1 backgrounds were evaluated globally by ten experienced raters on a scale from 1 to 6. However, unlike Linnarud, Enberg provided the evaluators with precise assessment criteria. In addition, four lexical indices were produced for each essay. Three of them were: the proportion of lexical items to all tokens in the text; the mean ratio of lemmas to the total number of lexical tokens based on 126 word-segments; the average percentage of lexical errors in a text segment. In addition, Enberg invented one more metric, which she called error-free lexical variation and which was calculated by subtracting the number of error types from the number of lemmas per segment and dividing it by the number of lexical tokens in a segment (cf. Section 4.3).

Engberg's results showed that three indices produced statistically significant (positive and negative) correlations with global scores at $p<0.01$ level. The gauges tapping separately into lexical variation and lexical accuracy produced similar moderate correlations ($r=0.43$, $r=-0.45$ respectively). The highest, but still moderate correlation was demonstrated by a metric which combined lexical variation and lexical accuracy—error-free lexical variation ($r=0.57$). Engberg concluded that diversity of lexical choice and lexical accuracy have a significant effect on raters' scores. She explained the discrepancy between her results and these of Linnarud's concerning the

correlation of lexical errors with the evaluations by the fact that her participants were at lower proficiency levels and were likely to produce more serious lexical errors which could have affected the comprehensibility of the results.

Yet another attempt at establishing the role of lexical factors in the human perceptions of learners' overall proficiency was made by **Daller and Phelan (2007)**. Their aim was to investigate empirically the usefulness of several measures of lexical richness in predicting human ratings of learner compositions. They used several indices which they divided into word-list-free approaches: TTR, Guiraud's Index and D, and word-list-based approaches: Condensed Lexical Frequency Profile (CLFP, the percentage of words beyond 2000), P-Lex and Advanced Guiraud's. The index values were computed for 31 essays written by learners of English from a variety of L1 backgrounds taking a preparatory course to study in the UK. The compositions were also rated by four experienced teachers. Unlike in the study by Engberg, the teachers used both a holistic and an analytic scale. The scale used for all evaluations spanned from 1 to 9, but specific band descriptors were only provided for the holistic scale. After a scrutiny of the correlations between the holistic and analytic scores, the researchers concluded that global ratings were the most reliable of all the evaluations. In addition, a high correlation between the global marks and the scores for vocabulary range (alpha=0.947) implied that the former can well represent human judgements of lexical richness. Thus the lexical indices were further correlated only with the holistic ratings.

The results demonstrated that all the word-list-based measures correlated highly with the holistic evaluations (CLFP rho=0.594 $p<0.01$, P-Lex rho=0.494 $p<0.05$, Advanced Guiraud rho=0.471, $p<0.01$). On the other hand, two of the word-list-free measures produced very low and statistically non-significant correlations. The TTR's result was not a surprise as this measure was not corrected for essay length; however low correlation for D (rho=0.235, n.s.) indicated that lexical variation was not taken into account by the judges. The researches did not know how to interpret a high correlation for another index of lexical diversity, Guiraud's Index (rho=0.577, $p<0.01$), especially because its values also correlated highly with all the measures of lexical sophistication (rho=0.591 to 0.832 $p<0.01$). Daller and Phelan's examination pointed to lexical sophistication as the measure which captured better the influence of vocabulary on human judgements of the quality of L2 learners' written production. Unfortunately, the researchers did not include indices tapping into lexical accuracy in their analysis.

Even more ambitious in its scope of the analysed indices was the study conducted by **Lu (2012b)**. In his paper, he examined 26 different metrics as measures of lexical density, lexical sophistication and lexical variation, and explored their relationship to human perceptions of the quality of L2 learners' productions. His analysis, however, was based on spoken rather

than written data. Lu's aim was to investigate how the individual components of lexical complexity—and different indices representing these measures—compare with and relate to each other in evaluation of L2 learners' oral performance. He used 408 transcripts of one task from a Chinese nationwide oral examination, retrieved from the Spoken English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005). The exam scoring procedure was fairly complex, and the only evaluations available to the researcher were task performance rankings based on initial scores given by two and sometimes more raters, in groups of 32–35 transcripts.

Among the 26 indices examined by Lu, there was one metric reflecting lexical density, five metrics corresponding to lexical sophistication and 20 metrics related to lexical variation. The lexical sophistication indices were calculated as the ratios of: sophisticated lexical tokens and types to all lexical tokens and types, sophisticated verb types to all verb tokens, and two corrected versions of this last metric which reduced the effects of sample size. The definition of sophisticated items was based on the 2000 frequency threshold. The lexical variation indices included families of metrics based on the number of different words, (four indices), on type/token ratio of entire vocabulary (seven indices) and on type/token ratio applied to one or more word classes (nine indices). Lu calculated the 25 lexical indices for each pre-processed transcript using a self-developed computer tool, *Lexical Complexity Analyser* (Lu, 2012a). For the computation of one metric, D, he used the *vocd* programme (cf. Chapter 4). Since the human evaluations were only available in the form of relative rankings within groups, a more complex statistical procedure had to be used to relate them to the values of individual metrics. Lu computed Spearman rho correlations between the task performance rankings and each of the 26 indices and then performed a fixed-effect meta-analysis to arrive at average correlations for the whole set of essays. At the same time each essay was classified into a proficiency band based on its ranking position thus enabling an analysis of variation in the mean values of the analysed indices across four proficiency levels. Thus, Lu analysed quality judgements both as an ordinal variable, by means of the correlation, and as a categorical variable, by means of the analysis of variance among group means.

The overall results demonstrated that the three aspects of lexical complexity (density, diversity and sophistication) did not correlate with each other. This observation was interpreted by Lu as an indication that the three measures did not tap into the same construct. Out of the three measures, lexical variation had the strongest relationship with the human judgements and the index which produced the highest correlation was the number of different words in an entire sample ($\rho=0.526$, $p<0.001$). Interestingly, this is the simplest metric of lexical variation,[2] which does not capture the repetition rate of individual words nor is corrected for the effect of text length. Moreover, a one-way ANOVA revealed

highly significant differences between the four levels ($F(3, 404)=41.675$, $p<0.001$). Other indices belonging to the three families of lexical variation metrics also demonstrated significant differences between their mean values across the four levels as well as significant combined correlations, but these correlations were either weak or very low. Lexical sophistication, in turn, produced very small effects in the analysis. Neither of the two general sophistication indices showed significant combined correlations with the raters' rankings and only one (the ratio of sophisticated lexical tokens to all lexical tokens) revealed a significant difference between its mean values across the four levels ($F(3, 404)=0.896$, $p<0.05$). Only the two transformed verb sophistication metrics demonstrated significant correlations with the ratings, but they were very low ($\rho=0.166$, and $\rho=0.165$, respectively $p<0.001$). These two sophistication metrics also demonstrated statistically significant differences between the four proficiency levels ($F(3, 404)=3.772$; $F(3, 404)=2.760$ $p<0.05$). Lexical density did not produce significant correlations with the human evaluations nor significant differences between the four levels.

Lu's study demonstrated a moderate relationship between human judgements of L2 texts and the values of lexical indices representing lexical variation. No effects were found for lexical density and lexical sophistication gauged by general indices. The researcher compares his outcomes to the results of the analyses performed by Linnarud (1986) and Laufer (1994), which pointed to lexical sophistication as an adequate measure discriminating between L1 and L2 writers on the one hand, and L2 writers at different proficiency levels at the other. Lu explains the lack of effect for lexical rarity in his investigation by a difference this feature may play in spoken and written proficiency (p. 198).

The influence of phraseology on human raters' perception of L2 learners' text quality was examined by **Bestgen and Granger (2014)**. They applied a set of indices, which they named a CollGram profile (cf. Section 4.4.5), to essays written by 57 learners at three different times in a semester. The essays were scored by two raters using two different scales, holistic and analytic, containing the same main five criteria. Next, the two grades were averaged for each essay to produce a single score for each of the main criteria and for the holistic grade. Three assessments were selected for further analysis: the overall scores, the vocabulary scores and the language use scores, as the other criteria, i.e. content, organisation and mechanics were deemed unrelated to phraseological competence. The researchers correlated each of the CollGram indices with the raters' scores. The results demonstrated positive and statistically significant correlations between the mean MI values and the three averaged scores. In addition, negative and statistically significant correlations were observed for the proportions of n-grams absent from COCA. There were no statistically significant correlations between mean t-score values and the raters' grades. All the significant correlations were rather low and

ranged for tokens and types from r=0.22 to r=0.43, p<.001 and p<.01 for MI values and from r=–0.15 to r=–0.37, p<.001 and p<.05 for the absent bigrams. However, the authors find them meaningful, as they note that other criteria not related to phraseology also have their influence on each of these grades. Interestingly, all the six indices showed the lowest correlations with vocabulary grades and highest with language use scores. The researchers suggest that the computed indices seem to capture the grammatical dimension in addition to the lexical dimension. Many retrieved combinations in fact include one or even two grammatical words, thus they seem more relevant for the perception of grammatical accuracy rather than formulaicity subsumed under the vocabulary criterion. Another observation was that there was little difference between correlation coefficients calculated for tokens and types, which could suggest that several repetitions of the same bigrams did not influence the grade. The qualitative analysis of 200 randomly selected absent bigrams indicated that in 70% of the cases they were erroneous combinations of words, only 30% represented possible sequences which just happened not to be present in the reference corpus. As the last step of the analysis, Bestgen and Granger (2014) conducted a multiple regression with MI values and percentages of absent bigrams as predictors and the language use grades as the dependant variable. The model, however, produced a weak improvement in relation to the correlations. The two indices could jointly explain 48% of variance in the scores.

5.3.2 Regression Studies

The studies discussed in the previous section focused on investigating and explaining the motivation behind human raters' evaluations of L2 texts in relation to the texts' vocabulary. More recent studies have taken a step forward by attempting to employ a set of linguistic indices, including lexical metrics, in models predicting human judgements. A series of such studies was conducted by Crossley and his colleagues. They applied a host of metrics computed by a suite of text-processing tools *Coh-Metrix* (cf. Chapter 4). These measures were classified as relating to a text's surface code level (e.g. the number of words, average word length or instances of passive voice), its textbase level (e.g. lexical diversity or the number of connectives), as well as its deeper-level, situation model (e.g. semantic co-referentiality or given/new information). The researchers' aim was to establish which gauges can best account for holistic and analytic scores assigned to written and oral texts produced by native speakers and learners of English in different conditions. They employed multiple regression analyses. Each collection of analysed data was divided into a training and a test set using a 67/33 ratio. In the first step, the variables in the training set were checked for their collinearity (r>0.7) and only the unrelated gauges with the highest correlation to human scores were retained for further

analyses. The selected indices were used as predictors in the subsequent regression analysis, whose aim was to establish their relative importance in predicting the human scores. Next, the derived regression model was applied to the test set to predict the scores. Finally, a correlation between the predicted scores and the actual scores in the test set was computed to assess the strength of the model.

In the **2012 study Crossley and McNamara** analysed a collection of 514 essays written by secondary school students in Hong Kong as part of a national exam. The essays were rated on a six-point scale by trained assessors. The researchers hypothesised that in addition to linguistic sophistication, as expressed by lexical and grammatical complexity metrics, the measures representing the cohesion aspects of a text will play a significant role in predicting raters' evaluations of L2 essays. To control for the effects of text length on raters' holistic marks (Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Reid, 1990), only essays including 485–555 words were analysed in the study. An initial scrutiny of many different metrics representing 12 banks of conceptually similar indices resulted in a selection of 12 variables, which were next entered into the stepwise linear regression performed on the training set of 344 essays. The regression yielded a significant model accounting for 30% of the variance in the ratings with five variables being significant predictors: D (lexical diversity), word familiarity, CELEX content word frequency, word meaningfulness and aspect repetition. The model was than tested on the test set of 170 essays, and its results demonstrated that the combination of the five variables accounted for 21% of the variance in the human grades. The application of the model to the entire collection of essays revealed that the combination of the five variables accounted for 26% of the variance in their evaluations.

The results indicated that that between 20–30% of variance in human evaluations of essays written by advanced L2 learners can be explained by the lexical quality of produced texts. Four of five measures, which were singled out as strong predictors of the marks, related to lexical complexity of essays (i.e. lexical diversity and lexical sophistication). Measures of grammatical complexity, which is also part of general linguistic sophistication, did not produce significant correlations with the evaluations. The index of aspect repetition was assumed by Crossley and McNamara to represent cohesion. However, this variable explained only 0.02% of the variance in the ratings. Moreover, higher proficiency essays were in fact characterised by a lower level of aspect repetition. This last result and the lack of other measures of cohesion in the final model suggested that higher proficiency learners do not appear to produce more cohesive texts. These findings corroborate the results of earlier studies by the same researchers analysing essays written by native speakers of English (McNamara, Crossley, & McCarthy, 2010; Crossley, Weston, McLain Sullivan, & McNamara, 2011).

The predictive power of various gauges of linguistic sophistication and cohesion for human ratings of L2 texts was explored further by **Guo, Crossley, and McNamara (2013)**. They analysed two writing tasks—independent and integrated—drawn from 240 *TOEFL iBT* examination scripts. The scripts were rated by two or three experienced assessors on the scale 1–5 following the standardised holistic rubrics (cf. Chapter 3). In addition, four essay characteristics were quantified in the study: text length, lexical sophistication, syntactic complexity and cohesion, and a host of different indices related to these four components were tested as predictors of the human ratings of the two essay types separately.

The results demonstrated both similarities and differences in the variables predicting human judgements between the two tasks. In both tasks, text length was the strongest predictor, accounting alone for 47.8% and 26.4% of the variance in the human scores for the independent and integrated essays respectively. Another aspect of text quality contributing significantly to the predictions of the human ratings in both tasks was lexical sophistication, as measured by average syllables per word and noun hypernymy values in the case of independent task, and by word familiarity information and CELEX frequency for content words in the case of the integrated task. Interestingly, like in the Crossley and McNamara (2012) study discussed earlier, the analysis did not demonstrate that syntactic complexity features were significant predictors of essay evaluations in either of the tasks. However, the frequency of some syntactic categories played an important role in accounting for the raters' scores. Finally, the cohesion features had their contributions in the regression models for the two tasks. In the case of the independent task the presence of conditional connectives had a negative effect on the scores, whereas in the integrated task semantic similarity (LSA sentence to sentence) played a significant and positive role in predicting the ratings. Altogether five variables accounted for 65% of variance in the human evaluations of the independent task and seven indices explained 58.7% of variance in the integrated task. It should be noted, however, that the researchers did not compare the values of the analysed indices for the individual learners. If the indices indeed reflect linguistic knowledge rather than contextual demands, these values should be stable across the two tasks for individual learners.

The studies discussed previously explored to what extent automatic indices describing a text's lexical texture and other aspects of language, correlated with or could predict human judgements of a text's overall quality. The results indicated that while various features related to vocabulary used by an L2 learner in his or her production—mainly lexical diversity and lexical sophistication—have an important influence on the perceived quality of this production, there are a number of other yet undefined variables which contribute to its final evaluation.

As Crossley and McNamara (2012) observe in the conclusion to their investigation:

> future studies might consider what features of the analysed texts outside of the linguistic features might play a role in writing proficiency. These features could include error production, contextual factors such as truthfulness and accuracy, world knowledge and rhetorical style. All of these features might help to explain the additional variance not predicted by the linguistic variables examined in this study.
>
> (Crossley & McNamara, 2012, p. 132)

5.3.3 Analytic Scores of Lexical Proficiency

Raters' holistic scores, used in most of the studies described in the previous two sections are reflective not only of the lexical texture of a text, but also of its other non-lexical characteristics. **Crossley et al. (2011a)** addressed this concern by asking raters to focus solely on the lexical aspects in their evaluations of L2 essays. They also developed a special lexical proficiency assessment rubric which is discussed in detail in Section 3.6. Samples of 140 words drawn from free-writes produced by 180 English learners and 60 native speakers were evaluated according to the lexical assessment criteria by three raters, who had first been trained in applying the rubric. At the same time ten indices were selected as measures of lexical proficiency: lexical diversity, polysemy, hypernymy, semantic co-referentiality (as reported by Latent Semantic Analysis, cf. Section 4.4.3), word frequency, word concreteness, word imagability, word familiarity, word meaningfulness and word length. These gauges were computed for the analysed texts with the help of *Coh-Metrix*. Only two out of the ten variables did not correlate significantly with the human scores: polysemy and word length. In addition, two indices—imagability and concreteness—highly correlated with each other, thus only imagability was retained for further analysis. Thus, altogether seven indices were used in a regression analysis of the training set of 160 essays. Its results demonstrated that only three of these variables were significant predictors of the human scores: lexical diversity, measured by D, word hypernymy and CELEX content word frequency. The three variables combined explained 46% of variance in the human scores. The results of the model applied to the test set demonstrated that the three indices accounted for 42% of the variance in the ratings of the 80 writing samples. The application of the model to the whole collection of samples yielded the result of 44%.

Crossley et al. (2011a) asserted that a simple and practical model can be proposed for predicting L2 learners' lexical proficiency based on short samples of their writing and this model can be a useful instrument in foreign language pedagogy for making decisions about learners and for

measuring their progress. Based on the regression analysis the researchers arrived at the following formula (p. 573):

$$\begin{aligned}\text{Predicted lexical proficiency} &= 4.701 + (0.022 \times \text{lexical diversity: D value}) \\ &+ (-1.130 \times \text{average of word hypernymy value}) \\ &+ (-0.736 \times \text{CELEX content word frequency value})\end{aligned}$$

The researchers stress that their model has not only been shown to have a predictive validity by being juxtaposed with human evaluations, but it also has a good grounding in the psychological theories of the mental lexicon and lexical processing. The findings of this study were corroborated by **Crossley et al. (2011b)** applying the same analytic procedures to a set of spoken data produced by L2 learners and native speakers. The results of this investigation pointed out to a similar combination of variables: lexical diversity D, word imagability, word familiarity (which measure is related to frequency) and hypernymy as strong predictors of human judgements of lexical proficiency, together accounting for 60% of variance in the raters' scores.

While Crossley et al.'s (2011a, 2011b) approach is undoubtedly an important step in the search for valid, reliable and practical methods of assessing L2 learners' lexical proficiency, it has to be borne in mind that their models explained a little below 50% and 60% of the variance in human judgements. This implies that other variables also had some influence on the raters' scores. These variables were either present in the lexical proficiency assessment rubric, but were not measured adequately by the statistical indices, or they could be absent from the rubric but implicit in the human perception of lexical quality. The first of these problems was addressed by **Crossley, Salsubury and McNamara (2013)** in a replication of the study discussed previously (Crossley et al., 2011a). The same data—240 free-writes—were used in the analysis and they were also scored according to the same lexical proficiency assessment rubric as in the previous study, described in detail in Section 3.6. However, in the replication study, prior to the lexical-holistic scoring, the raters were asked to assess the learners' texts analytically according to eight lexical criteria, which, according to the authors, represented four subcategories of lexical proficiency (Table 5.2).

Each of the criteria was briefly described in an analytic assessment rubric developed for this study and presented in Section 3.6. Out of the multitude of linguistic measures computed by *Coh-Metrix*, the gauges which were judged to match best the analytic evaluation criteria were selected. It is interesting to note that there was no one-to-one correspondence between the analytic criteria and the lexical measures assumed to represented them: some analytic features represented up to three different measures and some of the measures were represented by different analytic

Table 5.2 Lexical Analytic Components and Associated Lexical Measures

Analytic feature	*Lexical measure*
Basic category score	Word imageability Word concreteness Word hypernymy
Word concreteness	Word imageability Word concreteness Word hypenymy
Word specificity	Word imageability Word concreteness Word hypenymy
Semantic co-referentiality	Semantic similarity
Collocation accuracy	Word associations
Sense relations	Word frequency Word polysemy
Sense frequency	Word frequency Word polysemy
Word frequency	Word familiarity Word frequency
Lexical diversity	Lexical diversity

Source: Crossley, Salsubury and McNamara (2013, p. 111)

features. In addition, each lexical measure was gauged in *Coh-Metrix* by several indices. For example, three metrics of semantic similarity were used in the study: LSA overlap all sentences in paragraph, LSA overlap all sentences in text and LSA overlap adjacent sentences.

The human scores, both analytic and lexical-holistic, as well as all the selected indices computed by *Coh-Metrix* were subjected to a series of statistical analyses. The correlations between the average human scores related to each analytic feature and various corresponding indices ranged from r=–0.019—for the correlation of sense frequency scores and CELEX all-word frequency—to r=0.772—for the correlations of lexical diversity captured by TTR. The next step in the analysis involved entering the indices with the highest correlation in each analytic criterion (|r|=0.230 to 0.772) into a multiple regression against the human holistic ratings. The regression analysis was first performed on the training set of 160 texts. The results demonstrated that four variables were significant predictors in the regression: type/token ratio, word hypernymy, word frequency and word polysemy. The combination of these four variables explained 40% of variance in the training set for the holistic scores. The application of the model to the test set of 80 samples accounted for 37% of variance in the holistic judgements.

The researchers concluded that small to large effects of correlations between the statistical scores and the analytic criteria, the latter representing operationalised constructs of the components of lexical proficiency, provide evidence of a degree of convergent validity for most of the automated indices examined in the study. This meant that the indices were measuring the constructs which they were predicted to measure. It is worth noting that the two selection methods of variables to be used in the regression model resulted in the same measures as highest predictors of human lexical-holistic ratings. With the exception of polysemy, which was not a significant predictor in the previous study, both investigations pointed to lexical diversity, word hypernymy and word/sense frequency as strong predictors of human judgements. It needs to be noted, however, that different indices were selected to represent these measures in the two studies. It also needs to be pointed out that the four variables in this study, even though more carefully validated against human perceptions, explained only 40% and 37% of variance in the human scores in the training and test set respectively as opposed to 46% and 42% in the previous study.

The low correlations with the human analytic judgements of some groups of lexical indices confirm the questions already raised in the discussion of the previous study. In spite of the rather optimistic conclusion drawn by Crossley, Salsubury and McNamara (2013), these results clearly indicate that some of the analysed indices are not the best measures of the individual components of lexical proficiency. Their selection for the analysis was in fact motivated not so much by theoretical considerations but by their availability in *Coh-Metrix*. While this suite of programmes features a multitude of linguistic measures, it was designed with the primary aim to measure readability of texts and not to assess the lexical or overall linguistic proficiency of an L2 learner. For example, the choice of the index of word meaningfulness for content words in a text raises serious doubts as a measure of collocational competence. A potential collocability of a word, which is tapped by the word meaningfulness measure, does not indicate if the word occurs in the right company in a particular text. There are other metrics, discussed in the previous chapter, which could be used to measure this component more accurately, but which happen not to be computed by *Coh-Metrix*.

A critical evaluation of both studies (Crossley et al., 2011a; Crossley, Salsubury, & McNamara, 2013) indicates that the search for valid and reliable indices of various aspects of lexical proficiency is not concluded yet. More research is needed in this area to identify most suitable gauges of the individual components of lexical proficiency as revealed in L2 learners' written production. However, a better definition of the construct of lexical proficiency as reflected in writing, and of its relation to the constructs of vocabulary knowledge and the mental lexicon should be a starting point in the quest for automated measures for vocabulary assessment.

Such a need is addressed by **Jarvis (2013)**. In the introduction to his paper he argues against using existing statistical text-based indices of various aspects of linguistic performance for representing constructs without prior adequate definitions of these constructs. He expresses his concern in the following words:

> More worrisome, however, are measures that have been developed prior to or in the absence of an adequate theoretical construct definition, as well as measures that are used in ways that are incompatible with or reflect a poor understanding of the construct definition (assuming that there is a construct definition in the first place).
>
> (Jarvis, 2013, p. 14)

In response to his own criticism, Jarvis makes an attempt at dissecting the concept of lexical diversity as the measure most widely used in studies devoted to evaluating the quality of a text and to assessing L2 learners' lexical proficiency. First, he defines the construct theoretically and distinguishes its component properties. These theoretical considerations and the precise construct definition is described in detail in Section 3.6. Subsequently, he validates the choice of the construct components against human perceptions, which process was presented in Section 3.7.2. The next stage in Jarvis's analysis involved proposing suitable measures of each of the six properties and establishing their joint effect on human judgements of global linguistic proficiency and lexical diversity. The researcher proposed the following gauges of the individual properties:

- variability: MTLD
- volume: number of tokens in a text
- evenness: standard deviation of the number of tokens per type in a text
- rarity: mean rank of all words in a text as identified in a frequency-ordered lemmatised wordlist retrieved from a large reference corpus (in this case the British National Corpus)
- dispersion: the mean distance between different tokens of the same type, aggregated for all types in a text
- semantic disparity: the mean number of words in a text that share the same semantic sense, as measured by the WordNet semantic sense index. (Mihalcea & Moldovan, 2000)

As the next step, Jarvis compared and calibrated the values for the six indices with human ratings of the overall quality of essays produced by L2 learners and with evaluations of lexical diversity of texts written by L2 learners and native speakers of English. For the first comparison Jarvis used 210 narrative essays written by Swedish- and Finnish-speaking

learners of English. The essays were holistically scored by two trained raters who used a nine-point scale (0–8) with finer distinctions between the levels marked by pluses and minuses. Thus, the final evaluation represented 26 finely grained ranks, with the inter-rater reliability of 0.94, which was very high for such a level of distinction. At the same time, the data were lemmatised and the six indices were computed automatically for each narrative. Finally, the values of the six gauges for each essay were correlated with human holistic evaluations. The results demonstrated that five out of six variables produced statistically significant ($p<0.01$) Pearson correlation coefficients: dispersion ($r=0.77$), followed by volume ($r=0.72$), followed by evenness and disparity ($r=0.55$) and finally by variability ($r=0.41$). Interestingly, rarity did not produce a significant and strong correlation with the final scores ($r=0.12$) which is not only counter-intuitive but also stays in opposition to the results of most of the studies discussed in this chapter (except for Linnarud, 1986). It has to be noted, however, that Jarvis operationalised diversity through the mean rank rather than the mean frequency or frequency bands, which must have influenced the results. What Jarvis also found disappointing was that two measures (evenness and dispersion) correlated with volume above the 0.8 level, which clearly indicates that more research needs to address the problem of finding most adequate measures of individual lexical qualities.

The main aim of Jarvis's study was to compare the measures of the six properties with human perceptions of lexical diversity of texts. For this analysis he used 37 narratives and supplemented them with 13 essays written by equivalent American native speakers, thus arriving at the total of 50 texts. Eleven raters were asked to evaluate lexical diversity of the narratives, "defined simply as a variety of different words used" (p. 34). The scale used in evaluation had a range of ten points. No assessment rubric was provided and the raters were encouraged to rely on their intuitions. At the same time the six indices were computed for the 50 essays and their values were correlated with the human judgements. Out of the six indices, five correlated with the human scores. Interestingly, the order of the strength of correlation was almost identical as in the case of holistic judgements: volume, dispersion, evenness, disparity and variability ($r=0.67$; 0.64; 0.53; 0.43; 0.31 respectively) with rarity also not producing a significant result ($r=0.26$). The same high correlations between volume on the one hand, and dispersion and evenness on the other were an additional proof that the two latter measures of these two properties are sensitive to text length. However, the results of multiple regression analysis were rather disappointing. Jarvis performed it using three different methods, but each time he arrived at the score below 50% ($R^2=0.47–0.49$, $p<0.001$). The results indicate that all the six variables, including rarity, contributed to predicting human raters' judgements, yet the fact that the model could account for only less than 50%

of the variance in the scores suggests that both the individual properties of the diversity and the way they are measured need to be reconsidered. An additional problem was a fairly low inter-rater agreement between the human scores (r=0.45). One reason might have been that the raters had never performed such type of evaluation. Although it can be argued that lexical diversity is a perceptual phenomenon, in real assessment situations it is never evaluated separately from other components of linguistic proficiency, thus the judges might have had trouble separating it from other characteristics of the texts. Yet, Jarvis explicitly did not want to present the raters with an assessment rubric in order to avoid influencing the raters' intuitive perceptions. He elaborates on his decision in the following way:

> Although it would have been possible to create a lexical diversity rubric to assist the raters in their judgments, this would have resulted in a severe circularity of purpose because the rubric would have reflected the six proposed properties, yet the purpose of the study in the first place was to determine whether these six factors affect human judges' perceptions of lexical diversity without their being told what to look for. Therefore, in order to determine whether human judges already have an intuitive sense of what lexical diversity is, and in order to determine whether their intuition is grounded in the six proposed properties, they were given . . . minimal instructions.
>
> (Jarvis, 2013, p. 34)

5.4 Conclusion

All the studies discussed in this chapter illustrate an interesting attempt to apply lexical indices for measuring the growth in lexical proficiency. One line of research attempted to establish whether the values of selected lexical measures differ for native speakers and learners with different proficiencies. Another line of development was devoted to juxtaposing automated indices and human perceptions of L2 learners' general language proficiency and lexical proficiency. The initial attempts in this area, made in the 1970s and the 1980s, concentrated on grammatical metrics and did not include vocabulary in the analysis. With the growing interest in lexis and is role in second language acquisition, which started in the 1980s, we could observe an increased interest in including lexical indices in the examination or focusing solely on their performance in relation to human evaluations. At the same time a host of new measures have been proposed and their relation to raters' judgements has been researched.

The aims of these studies have also shifted. The first studies attempted just to compare the values of different indices at different proficiency levels. More recent studies in this paradigm explored if the different measures employed jointly can discriminate between learners at different

levels. In the same vain, earlier studies attempted to explain human judgements by examining which indices demonstrated strongest correlations with the scores. This was done with the aim of discovering the role of lexis in the final evaluation of an L2 learner's text and of identifying the best index which could capture this dependence. Such analyses employed simpler statistical procedures involving mainly correlational analyses. However, the studies carried out in the 2000s and 2010s took up a more ambitious goal. Instead of explaining human scores by correlating them with various lexical and non-lexical gauges, they tried to predict human scores using a variety of lexical indices and examining their relative contributions to the final evaluation of global language proficiency or its lexical component. These studies employed much more complex statistical procedure—multiple regression modelling, attempting to arrive at a mathematical formula calibrating the importance of individual lexical metrics in the final score. Yet, the drawback of these studies was the selection of individual lexical gauges for the model, which was motivated primarily by their availability rather than a sound theoretical justification. Little attention was given to defining what specifically the selected indices actually measured in the context of lexical proficiency and if they were the most appropriate measures of lexical proficiency in the first place. Little attention was also paid to the fact that lexical proficiency is shaped by other variables than lexical competence. The next chapter presents an attempt in this direction.

Notes

1. Research using various measures of lexical richness to analyse languages others than English remains scarce (see Vermeer, 2000, 2004; van Hout & Vermeer, 2007 for analyses of Dutch).
2. It is interpreted as a measure of lexical productivity by Bulté et al. (2008) (cf. Chapter 4).

6 The Study—Measuring and Assessing Lexical Proficiency of Advanced Learners

6.1 Introduction

The development of a range of automatically computed gauges of lexis and—more recently—of phraseology in learners' written and/or spoken production constitute an important line of research on second language vocabulary acquisition and assessment. Such gauges, described in detail in Chapter 4, take the form of various—simpler or more complex—mathematical formulas describing vocabulary and formulaicity of a text. They have been applied in automatic essay scoring systems used in second language testing on the one hand, and as a yardstick of learners' linguistic competence and development in SLA research, on the other. One way of validating these measures has been to compare them with other methods of assigning levels to students, for example according to the length of their language study or their results on an independent (vocabulary) test. Another way to demonstrate the validity of these gauges has been to juxtapose them with scores attributed to (the lexical aspects of) learners' speech or writing by human raters. All such validation studies have attempted to demonstrate a statistical relationship of one or several measures of lexis and phraseology in learners' texts with other measures either of their linguistic proficiency or of the quality of their spoken/written performance. A sample of such investigations was discussed in Chapter 5.

There are two important assumptions implicit in this line of research. It has been generally taken for granted that vocabulary knowledge, vocabulary use and a text's (lexical) quality are related in a straightforward linear fashion. The quality of a learner's writing or speech is assumed to relate to his or her use of varied, sophisticated—and even abstract and non-polysemous—vocabulary as well as to his or her selection of expressions which are formulaic and native-like. These premises were expressed in a quotation by Laufer (1994, pp. 21–22), who remarked that "[i]t is true that factors other than lexis affect the quality of a written composition, but a well used rich vocabulary is likely to have a positive effect on the reader". In the same vein, Li and Schmitt (2009, p. 86) observed that "learning to write well also entails learning to use formulaic sequences appropriately". Another implicit assumption is that

vocabulary use depends on and reflects a learner's lexical competence, in particular the size of his or her mental lexicon, as observed by Meara and Bell (2001, p. 9): "people with big vocabularies are more likely to use infrequent words than people with smaller vocabularies". Similarly, Groom (2009, p. 28) suggested that "the [more advanced] learners are relying less on an overused set of known lexical bundles". What follows from these assumptions is that more advanced learners, whose mental lexicon is larger and more complex, would use rarer, more abstract and more diverse vocabulary, and at the same time, their vocabulary use would be more formulaic in a subtle and non-clichéd way. All these characteristics are assumed to result in a higher quality of texts produced by more proficient learners.

It is inevitable for any research to be based on certain assumptions and the ones described earlier certainly have a strong intuitive appeal. They lie at the core of many practices in second language teaching, learning and assessment. Yet, few studies have ever looked closely into these beliefs and tested them empirically. Although it seems against common sense to question the existence of such relationships, it may be worthwhile to test their limits. Thus, the aim of this study is to critically examine these premises. This investigation will focus on written performance of upper-intermediate and advanced learners, and juxtapose it with texts produced by equivalent native speakers. More specifically, the study will (1) compare the information on the participants' proficiency levels and vocabulary knowledge established with independent methods with their vocabulary use described by various lexical and phraseological measures. The study will also (2) juxtapose the information on the quality of (the lexical aspects of) the participants' texts provided by human raters with the assessment of their texts' lexis and phraseology yielded by automatic indices. Finally, the study will (3) analyse human raters' perceptions of good and effective vocabulary use as well as (4) the extent to which automatically computed lexical indices describing vocabulary use converge with the perceptions of human raters concerning the quality of (the lexical aspects of) the participants' written production. Unlike the previous studies discussed in Chapter 5, this investigation will look more closely into other possible variables affecting vocabulary use and human perceptions of this use. In order to address these issues, the data collected for this study will be scrutinised both quantitatively and qualitatively.

6.2 Research Questions

The following specific research questions were posed at the outset of the study:

1. Do advanced learners have a larger receptive and productive lexicon than strong upper-intermediate learners? Or does vocabulary growth reach a plateau once learners become independent users of a

foreign language who can easily compensate for gaps in their lexical command?
2. Do advanced learners use more sophisticated, varied, abstract and less polysemous vocabulary, as well as more standard phraseology in their written production than strong upper-intermediate learners? Do equivalent native speakers surpass advanced learners in these respects? Or does vocabulary use by advanced learners demonstrate a ceiling effect?
3. To what extent are vocabulary knowledge of advanced learners (as measured by vocabulary tests) and their vocabulary use (as measured by various automatic indices) interrelated?
4. Do more proficient advanced learners produce texts of better (lexical) quality as assessed by human raters than less proficient advanced learners? Do equivalent native speakers surpass advanced learners in these respects?
5. To what extent do various automatic measures of lexical richness correspond to the raters' assessment of the quality of a text and of good and effective vocabulary use?
6. What is the perception of experienced teachers and raters of writing on the contribution of vocabulary to the quality of writing?
7. How do experienced teachers and raters define good and effective vocabulary use in writing?

6.3 Subjects and Instruments

In the study, two types of performance were elicited from Polish upper-intermediate and advanced learners of English. They are described in detail in the following sections.

6.3.1 Essays

The data used in this study were drawn from the PELCRA Learner English Corpus (PLEC), which contains samples of written and spoken language produced by Polish upper-intermediate and advanced students of English, mostly at the university level (Pęzik, 2012). Among its written data, the corpus includes a batch of 288 argumentative essays written by students in an English department in Years 1 through 4, on the same topic (*The curse or the blessing of mobile phones*), and in identical conditions (timed in-class writing in the first week of a new academic year, no access to reference materials). This collection was supplemented with 81 essays also written on the same topic and in similar conditions by American native-English-speaking students at the university level. The only difference between the Polish and American settings was timing: 90 and 50 minutes respectively. All essays were handwritten and then word-processed retaining the original spelling and layout as much as possible.

In addition to the homogeneity of the task, the Polish and American students themselves were also highly comparable. The Polish learners all

studied at the same institution, with an equal distribution of students in Years 1, 2, 3 and 4. All the students had to pass a very demanding entrance exam and, in the case of Year 2–4 students, end-of-the-year exams in English including Use of English, Writing and Speaking sections. Thus, the Polish data can be treated quasi-longitudinally. The native speakers matched the Polish learners in age (late teens, early twenties) and educational background. To ensure even more reliable and meaningful comparisons, three sets of 50 argumentative essays produced by the Polish learners in Year 1 and 4 and by the American students were drawn from this section of the corpus. Subsequently, each batch of essays was divided quasi-randomly[1] into two equal sets to be used for two different rounds of assessment by human raters. The composition of the non-native and native student data samples used in the study is presented in Table 6.1.

6.3.2 Vocabulary Tests

A week after the essays were collected, two vocabulary tests were administered to the Polish students: the receptive and productive versions of the Vocabulary Levels Test (Laufer & Nation, 1999; Schmitt et al., 2001; respectively) described in detail in Chapter 2. The students had 45 minutes to complete both versions. Fifty pairs of the receptive-productive tests which tallied with the essays in the Year 1 sample were drawn from the collection. Due to fluctuation in attendance, only 25 tests completed by Year 4 participants matched the texts in the sample. An additional 17 pairs were selected among the remaining tests written by Year 4 students. The summary of the available test pairs and their match with the essays in the written sample are presented in Table 6.2.

Table 6.1 The Composition of the Data Samples Used in the Study

Set 1	*Set 2*
75 essays:	75 essays:
25—Year 1	25—Year 1
25—Year 4	25—Year 4
25—Native	25—Native

Table 6.2 Vocabulary Levels Tests Available for Analysis

Year 1	*Year 4*
50 test pairs	42 test pairs
25—matching essays in Set 1	17—not matching essays
25—matching essays in Set 2	25—matching essays in Set 2

6.4 Data

The collected samples of students' performance were processed in order to obtain data for statistical analyses.

6.4.1 Lexical Indices

A range of lexical and phraseological measures were computed for the essays in the sample. Chapter 4 presented a whole host of different automated gauges. As the aim of this study is to compare the information on the vocabulary and phraseology of a text captured by statistical measures and human perceptions, the choice of the indices was based on the meaningfulness and interpretability of the information they encapsulate as well as their theoretical motivation, discussed in Chapter 4, rather than their sheer availability.

The texts were pre-processed for the computation of the indices. Minor spelling mistakes were corrected and proper nouns, acronyms, numbers rendered in digits and non-existent words were deleted.[2] However, other lexical errors, such as incorrect word use were not handled. Contractions were replaced with full forms. The spelling conventions in learner texts were standardised to American English or British English, depending on the norm applied for the computation of individual indices. Table 6.3 presents all automated lexical measures selected for the study.

Table 6.3 Lexical Indices Computed for Each Essay

Aspect of Lexical Proficiency	*Index Type*	*Index*	*Details of Computation*
Lexical productivity	quantity	Lgth	tokens—number of running words
		Lex_Ty	types—number of unique word forms
Lexical diversity	ratio	GUIR	types—unique word forms (GUIR_Ty) lexemes—unique units of meaning (a lemma with its inflectional forms) (GUIR_Lex)
	permutation	D	types—unique word forms
		MTLD	types—unique word forms
Lexical sophistication	profile	2K	tokens—percentage of running words with frequencies beyond 2K (2K_To) lexemes—percentage of unique units of meaning (a lemma with its inflectional forms) with frequencies beyond 2K (2K_Lex)
		AWL	tokens—percentage of running words from AWL (AWL_To) types—percentage of unique word forms from AWL (AWL_Ty)
	mean	FQLog	mean log word form frequency for all words (FQLog_AW) mean log word form frequency for content words (FQLog_CW)

Aspect of Lexical Proficiency	*Index Type*	*Index*	*Details of Computation*
Lexical density	ratio	LD	ratio of all lexical lexemes to all lexemes
Lexical elaborateness—word psycholinguistic properties	mean	FAM	familiarity for content word forms
		CNC	concreteness for content word forms
		IMG	imagability for content word forms
		MEA	meaningfulness for content word forms
Lexical elaborateness—word meaning relations	mean	POL	polysemy for content word forms
		HYP	hypernymy for nouns and verbs (word forms)

The index values were computed with the help of several text-processing programmes mentioned in Chapter 4: *Lexical Complexity Analyser* (Lu, 2012a), *Coh-Metrix* (McNamara et al., 2014), *AntWordProfiler* (Anthony, n.d.) and *TAALES* (Kyle & Crossley, 2014). The 2000 level as the cut-off point for sophisticated vocabulary (see Chapter 4 for justification) was based on the rank-ordered lemmatised frequency list retrieved from the American National Corpus (Reppen, Ide, & Suderman, 2005). The percentages of tokens and types constituting academic vocabulary were based on Academic Word List (Coxhead, 2000). Means of logarithmically transformed word frequencies for all words in a text as well as for content words only were calculated using frequency information retrieved from the written section of the British National Corpus (The BNC Consortium, 2007) and normalised for 1000 words.[3] They were selected over raw frequencies as they had been shown to reduce effects of few very rare words on mean values, and thus to produce more stable results (Baayen, 2001).

Several recently proposed gauges of phraseology were also computed for each text. They are all founded on bigrams and include frequency counts as well as two most frequently used association measures: pointwise mutual information (MI) and t-score. Table 6.4 lists the phraseological indices used in the study.

A custom-built programme was used to generate the gauges of phraseological complexity for this study.[4] It followed the procedure described in Bestgen and Granger (2014). The computation of formulaic indices was based on frequency information for unigrams and bigrams retrieved from Corpus of Contemporary American English (COCA, Davies, 2008–) and available at its related site Word Frequency Data (Davies, n.d.). The boundaries for individual bigram categories followed the benchmark set by Granger and Bestgen (2014) and presented in Chapter 4.

Table 6.4 Phraseological Indices Computed for Each Essay

Aspect of Lexical Proficiency	*Index Type*	*Index*	*Details of Computation*
formulaicity	ratio	BiAbs	tokens—proportion of bigram forms in text not found in COCA (BiAbs_To) types—proportion of unique bigram forms in text not found in COCA (BiAbs_Ty)
		BeTh	tokens—proportion of bigram forms in text whose frequency in COCA <5 (BiAbs_To) types—proportion of unique bigram forms in text whose frequency in COCA <5 (BiAbs_Ty)
		MI	tokens and types—proportion of bigram forms and unique bigram forms with freq ≥5 for which: • 3≤MI<5 (LowMI_To, LowMI_Ty); • 3≤MI<5 (MidMI_To, MidMI_Ty); • MI≥7 (HiMI_To, HiMI_Ty)
		T	tokens and types—proportion of bigram forms and unique bigram forms with freq ≥5 for which: • 2≤t<6 (LowT_To, LowT_Ty); • 6≤t<10 (MidT_To, MidT_Ty); • t≥10 (HiT_To, HiT_Ty)
	mean	LogBiFQ	tokens—mean log frequency of bigram forms (LogBiFQ_To) types—mean log frequency of unique bigram forms (LogBiFQ_Ty)
		BiMI	tokens—mean MI score of bigram forms (BiMI_To) types—mean MI score of unique bigram forms (BiMI_Ty)
		BiT	tokens—mean t-score of bigram forms (BiT_To) types—mean t-score of unique bigram forms (BiT_Ty)

Table 6.5 summarises descriptive statistics for all the indices analysed in this study.

The number of lexical—and even to a larger extent—phraseological indices used in this study may seem daunting at first glance. It should be borne in mind, however, that in many cases two gauges in fact represent the same index, but computed twice for each essay using a different unit as a basis of calculation—tokens, types or lexemes. This was done not so much with an aim of establishing which method of calculation is more informative; such concerns have already been tackled in literature (cf. Vermeer, 2000, pp. 78–79; Treffers-Daller, 2013,

Table 6.5 Descriptive Statistics for the Indices Used in the Study

Index	*Year 1*				*Year 4*				*Native*			
	mean	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*
Lgth	375.50	64.92	260.00	578.00	435.84	90.71	225.00	658.00	426.24	88.26	242.00	669.00
Lex_Ty	134.02	23.67	95.00	199.00	158.18	29.12	79.00	240.00	125.30	21.55	64.00	167.00
GUIR_Ty	9.97	0.79	7.28	11.23	10.64	0.78	8.69	12.47	9.17	0.87	6.45	10.67
GUIR_Lex	9.46	0.77	6.91	11.01	10.10	0.74	8.24	11.79	8.51	0.81	6.13	9.79
D	98.69	15.53	54.70	141.39	100.77	17.64	68.15	163.15	86.51	15.83	49.00	122.30
MTLD	97.47	16.58	47.30	135.05	101.92	19.87	69.01	187.04	78.15	16.32	45.43	115.89
2K_To	14.29	2.59	7.74	21.51	15.47	2.93	8.38	23.38	8.80	2.66	3.94	15.09
2K_Lex	21.91	3.93	12.35	31.00	24.72	4.64	13.30	36.87	16.53	3.78	9.79	28.08
AWL_To	3.10	1.34	0.31	7.10	3.91	1.64	1.21	8.54	2.21	1.07	0.00	5.00
AWL_Ty	5.33	2.09	0.57	11.68	6.78	2.71	2.29	15.52	4.07	1.77	0.00	8.24
FQLog_AW	4.92	0.08	4.70	5.09	4.90	0.08	4.74	5.07	4.93	0.09	4.68	5.13
FQLog_CW	4.26	0.10	3.99	4.44	4.20	0.10	4.00	4.38	4.30	0.10	4.04	4.52
LD	0.50	0.03	0.41	0.57	0.50	0.03	0.45	0.57	0.50	0.03	0.43	0.60
FAM	576.35	5.02	563.35	585.42	573.95	5.15	562.96	583.32	577.06	4.39	566.53	586.96
CNC	367.95	11.93	343.97	391.77	365.03	13.62	334.60	390.19	391.32	19.46	331.32	427.39
IMG	394.68	10.35	376.79	415.61	394.56	12.22	363.22	418.14	419.98	18.27	373.47	453.08
MEA	422.80	8.65	404.54	445.26	419.93	9.69	390.68	440.26	429.31	12.32	401.43	461.26
POL	4.09	0.33	3.39	5.05	4.04	0.28	3.43	4.85	4.61	0.36	3.73	5.40
HYP	1.58	0.15	1.13	1.96	1.60	0.16	1.27	1.99	1.87	0.21	1.26	2.38
BiAbs_To	0.03	0.01	0.01	0.06	0.03	0.02	0.00	0.09	0.02	0.01	0.00	0.06
BiAbs_Ty	0.03	0.01	0.01	0.07	0.04	0.02	0.00	0.09	0.02	0.01	0.00	0.07
BeTh_To	0.06	0.02	0.02	0.11	0.07	0.02	0.02	0.14	0.04	0.02	0.00	0.12
BeTh_Ty	0.07	0.02	0.02	0.12	0.07	0.03	0.03	0.15	0.05	0.02	0.01	0.13

(*Continued*)

Table 6.5 (Continued)

Index	*Year 1*				*Year 4*				*Native*			
	mean	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*
NonMi_To	0.60	0.03	0.51	0.68	0.59	0.03	0.50	0.67	0.63	0.03	0.56	0.71
LowMI_To	0.23	0.03	0.14	0.30	0.23	0.03	0.18	0.29	0.22	0.03	0.15	0.29
MidMI_To	0.07	0.01	0.05	0.09	0.07	0.01	0.04	0.09	0.06	0.01	0.03	0.11
HiMI_To	0.04	0.01	0.02	0.07	0.04	0.01	0.01	0.07	0.05	0.02	0.02	0.10
NonMi_Ty	0.61	0.03	0.54	0.67	0.60	0.04	0.52	0.69	0.65	0.03	0.59	0.71
LowMI_TY	0.23	0.03	0.14	0.29	0.22	0.03	0.17	0.30	0.21	0.03	0.16	0.28
MidMI_Ty	0.07	0.01	0.05	0.10	0.07	0.01	0.04	0.10	0.06	0.01	0.03	0.09
HiMI_Ty	0.02	0.01	0.01	0.04	0.03	0.01	0.01	0.05	0.03	0.01	0.01	0.06
NonT_To	0.17	0.02	0.12	0.22	0.16	0.02	0.12	0.21	0.19	0.03	0.13	0.25
LowT_To	0.06	0.02	0.03	0.13	0.07	0.02	0.03	0.11	0.06	0.02	0.02	0.11
MidT_To	0.05	0.01	0.02	0.07	0.05	0.01	0.02	0.09	0.04	0.01	0.01	0.08
HiT_To	0.66	0.04	0.58	0.75	0.66	0.04	0.56	0.78	0.67	0.04	0.57	0.76
NonT_Ty	0.18	0.02	0.12	0.24	0.17	0.02	0.13	0.23	0.21	0.03	0.14	0.27
LowT_Ty	0.07	0.02	0.03	0.13	0.07	0.02	0.04	0.11	0.06	0.02	0.02	0.11
MidT_Ty	0.05	0.01	0.02	0.08	0.05	0.01	0.03	0.09	0.05	0.01	0.01	0.08
HiT_Ty	0.64	0.04	0.56	0.73	0.63	0.04	0.55	0.76	0.64	0.05	0.52	0.72
LogBiFQ_To	0.79	0.13	0.51	1.10	0.75	0.11	0.54	0.96	0.86	0.16	0.43	1.30
LogBiFQ_Ty	0.69	0.13	0.38	0.94	0.64	0.11	0.38	0.92	0.76	0.14	0.41	1.07
BiMI_To	2.36	0.19	2.07	2.76	2.41	0.19	1.91	2.81	2.35	0.22	1.75	2.92
BiMI_Ty	2.18	0.16	1.85	2.52	2.24	0.17	1.77	2.68	2.01	0.22	1.42	2.55
BiT_To	103.56	14.20	74.59	133.40	100.81	16.23	38.62	132.30	93.74	20.33	16.27	144.43
BiT_Ty	80.11	11.53	59.37	100.39	77.68	12.24	30.11	107.22	74.88	21.04	–45.28	105.26

p. 84). A more important aim of double calculation for almost all gauges was to confirm the observed effects. Another reason explaining the multitude of variables in this study is the fact that several phraseological indices are in fact complementing proportions representing one construct. The split into several collocational bands is meant to capture the fluctuations in the formulaicity of language used by different groups of subjects. For example, the indices BeTh_Ty, NonMi_Ty, LowMI_Ty, MidMI_Ty and HiMI_Ty represent the distribution of bigrams in a text among different levels of association strength and they all add up to 1 (100%).

6.4.2 Raters' Grades

Four experienced raters—three males and one female—were asked to evaluate the essays in two rounds. They were all university lecturers in the same institution where the Polish data were collected, with a considerable experience in teaching advanced writing courses and evaluating writing exams. Two of the raters (N1 and N2) were native speakers of English: a Canadian (N1) and an American (N2) with their BA and MA degrees related to journalism (N1), creative writing (N2) and English (N1 and N2) awarded by American universities. They had worked as teachers of English as a foreign language in Poland (N1) and Russia and Poland (N2) for six and four years respectively. One rater (N1) also worked occasionally as a journalist, the other (N2) was a practising poet and sporadically published scholarly essays on poetry. The other two raters (P1 and P2) were Polish, with their MA degrees in English Philology and their PhD degrees in translation and legal English (P1) and pre-modern and early modern literature (P2). P1 applied corpus methodology for his research and thus was familiar with the notions of frequency patterns, lexicogrammar and automatic indices describing texts. Both Polish raters published regularly in English in academic journals and edited volumes. Thus, in addition to their experience as teachers of writing, the four raters were active professional writers.

The raters were asked to perform two rounds of assessment with separate instructions for each round. In the first round, they performed holistic evaluation of 75 essays in the first set. In the next round, they marked the other 75 essays in set 2 focusing only on lexis. The information on the students' profiles was communicated to the raters but the essays were coded and their order in the sets was mixed, so the authorship of individual texts was not disclosed to the raters. Furthermore, no specific marking guidelines were provided for two reasons. First, the four raters were experienced teachers of writing working at the same institution, thus they shared a general understanding of the institutional standards applied to student writing. The native raters also had had previous experience as writing instructors at American universities. Second, the study

targeted raters' personal criteria of good writing, in particular their perceptions of role of vocabulary as a factor contributing to the quality of text. Providing the raters with specified rubrics for assessment would have influenced their marking behaviour by forcing them to give more importance to certain criteria (such as, for example, vocabulary use) that they normally may not perceive as important (cf. Lumley, 2005; Jarvis, 2013). This decision was expected to have an unfavourable effect on inter-rater reliability, however, since the grades were not part of students' formal evaluation; this consequence was accepted as an inevitable outcome of the study.

The raters were given the first batch of printed essays with an instruction to evaluate their effectiveness as written discourse. They had two weeks to complete the task. After finishing this round, the judges were provided with the next set of printed essays and only then given the instruction to focus solely on vocabulary use in their assessment. They had another two weeks for this task. For both rounds of assessment the raters used the Polish university grading scale covering six bands:

2 (fail),
3 (satisfactory),
3+(3.5),
4 (good),
4+(4.5),
5 (very good).

The scores were then converted into a continuous scale from 0 (fail) to 5 (very good). The summary of the scores attributed to the essays by each rater in both rounds of assessment (global and lexical) are provided in Table 6.6.

Table 6.6 Descriptive Statistics for the Raters' Grades

	N	*mean*	*s.d.*	*min*	*max*
Global Assessment					
N1	75	2.05	0.96	0	4
N2	75	2.57	1.31	0	5
P1	75	2.75	1.46	0	5
P2	75	1.40	1.14	0	5
Assessment of Vocabulary					
N1	75	3.03	0.87	1	5
N2	75	2.53	1.36	0	5
P1	75	2.77	1.02	1	5
P2	75	1.77	1.07	0	5

6.4.3 Vocabulary Test Scores

Both versions of the tests were marked. A clear procedure was adopted for the treatment of grammatical and spelling mistakes in the productive version. The following guidelines were adopted:

- all inflected forms of a target word (e.g. singular or plural nouns, verbs with or without the third person singular *–s*) were accepted, even if they were not used correctly in context
- spelling mistakes were accepted, but only if they involved omission, addition or transposition of no more than one letter

Several observations, briefly touched upon in the discussion of word knowledge in Chapter 1, motivated the selection of these guidelines. Incorrect inflected forms of a word can be interpreted as reflecting gaps in learners' grammatical rather than lexical knowledge. They may also be caused by students' attention focused on word choice rather than its correct use in context. Spelling mistakes involving one letter, on the other hand, may be taken as evidence of incomplete but satisfactory knowledge of a word.

Scores were recorded for each section of the receptive and productive tests as well as for the whole test. Mean scores for Year 1 and Year 4 are presented in Table 6.7.

Table 6.7 Descriptive Statistics for the Vocabulary Levels Tests

Group	*Year 1*				*Year 4*			
	mean	*s.d.*	*min*	*max*	*mean*	*s.d.*	*min*	*max*
Receptive Vocabulary Levels Tests								
R2K	29.52	0.84	27	30	29.67	0.69	27	30
R3K	28.08	1.76	24	30	29.81	0.45	28	30
R5K	26.50	3.20	16	30	29.40	1.04	26	30
RAWL	34.74	1.17	32	36	35.55	0.74	33	36
R10K	17.62	4.92	8	26	23.07	4.36	13	30
RTot	136.46	7.83	120	152	147.50	5.16	137	156
Productive Vocabulary Levels Tests								
P2K	17.14	0.81	15	18	17.74	0.73	14	18
P3K	14.74	1.85	10	18	16.33	1.34	13	18
P5K	7.66	2.52	3	13	12.19	2.92	7	18
PAWL	12.78	2.61	0	16	15.48	1.80	10	18
P10K	6.52	2.58	0	11	9.33	2.87	2	16
PTot	58.84	7.58	36	71	71.07	7.20	54	88

6.4.4 Interviews

Interviews with the raters were conducted individually after they had completed the two rounds of assessment. The interviews with the native raters were carried out in English, the Polish judges were interviewed in Polish since this was also the interviewer's native tongue. However, the Polish raters frequently code-switched and expressed some of their opinions in English.

Two sets of questions concerning each round of assessment were used to guide the interviews. These questions are presented in Figure 6.1.

During the interviews the raters were presented with the two sets of questions separately and were encouraged to answer them with only minimal prompts from the interviewer. The interviews lasted from 20 minutes to half an hour and were recorded. Afterwards, the raters' responses were transcribed and cleaned of false starts, hesitations and repetitions and subsequently, detailed written reports on the judges' responses to the questions were produced in English. Care was given to retain the order of expressed comments and the original wording as much as possible. Finally, the reports were sent to the raters for corrections, clarifications and authorising. The authorised reports ranged from 673 to 1086 words.

Interview questions

Global Assessment

1. What criteria did you take into consideration when assessing the essays?
2. What was the weight of each criterion in the overall mark?
3. Did you assign points for each criterion to arrive at the final mark?
4. What role did vocabulary play in your global assessment? Did you see any interesting patterns?
5. How, on the whole, did you evaluate the essays?
6. Could you tell the three groups of essays apart? Did you see any differences between the levels?
7. Do you have any additional comments?

Assessment of Vocabulary

1. What criteria did you take into consideration when assessing the vocabulary?
2. What was the weight of each criterion in the overall mark for vocabulary?
3. Do you assign points for each criterion to arrive at the final mark?
4. How on the whole did you evaluate the essays?
5. Could you tell the three groups of essays apart? Did you see any difference between the levels?
6. Do you have any additional comments?

Figure 6.1 Questions Used in the Interviews With the Raters

6.5 Data Analysis

The data were analysed quantitatively and qualitatively. Three types of quantitative analyses were performed: correlation, analysis of variance and regression. The examination was conducted in ten steps, which are listed in Table 6.8.

The results of each analyses will be described in the following sections.

6.5.1 Analysis—Associations Between the Indices

Preliminaries

Lexical indices may correlate for two reasons: (1) they can measure the same construct, or (2) they can measure different but related constructs

Table 6.8 Analyses Performed in the Study

Step	*Type of Data*	*Type of Analysis*	*Statistical Test*
Analysis 1	indices	associations between indices	Spearman rank-order correlation, Pearson product-moment correlation
Analysis 2	indices	comparison of indices between groups	one-way ANOVA
Analysis 3	8 indices	prediction of group membership based on selected indices	multinomial logistic regression
Analysis 4	raters' grades	associations between raters' grades (inter-rater reliability)	Spearman rank-order correlation
Analysis 5	raters' grades	comparison of mean values between groups	one-way ANOVA
Analysis 6	raters' grades, indices	associations between raters' grades and indices	Spearman rank-order correlation
Analysis 7	raters' grades, 8 indices	prediction of raters' grades based on selected indices	linear regression
Analysis 8	vocabulary test scores	comparison of vocabulary scores	t-test
Analysis 9	vocabulary test scores, raters' grades, indices	associations of vocabulary scores with indices and raters' grades	Spearman rank-order correlation, Pearson product-moment correlation
Analysis 10	interviews	raters' perceptions of assessment	qualitative analysis

which vary in parallel. That is why correlations were analysed first for the indices linked to the same traits of lexical proficiency and then among the indices capturing different aspects of lexical competence. The former analysis should help to establish concurrent validity of the selected indices to measure the proposed constructs and the latter the relationships between different aspects of lexical proficiency proposed by researchers.

Data Analysis

The correlations were compared between the indices produced for all the texts in the analysed sample jointly, irrespective of the subject groups they referred to. Almost half of the indices did not demonstrate normal distribution (Shapiro-Wilk, $p<0.05$). In such cases Spearman's rank correlation coefficients were calculated instead of Pearson product-moment correlation coefficients. Table 6.9 presents the correlations between the lexical indices, Table 6.10 contains the same results for the phraseological measures and Table 6.11 tabulates correlations between the lexical and phraseological gauges.[5]

Out of the total 946 pairwise correlations in the three tables, over two-thirds (647) coefficients are statistically significant, over half (563) are at least weak (rho>0.200), and over one quarter (266) are moderate or high (rho>0.400; rho>0.700 respectively). Only 14% (19) of all the correlations between the lexical indices (total 136) presented in Table 6.9 are statistically insignificant or significant but negligible. A far larger number of insignificant or negligible relationships can be found between the phraseological measures (128 [29%] of the total 325 of correlations) tabulated in Table 6.10.

Discussion

Several observations can be made after the detailed analysis of the correlation matrixes. First, the choice of the computation unit (token/type/lexeme) has only a marginal effect on the information captured by individual measures. This claim is supported by the fact that the indices calculated with the same formula, but applying a different unit of counting, are strongly correlated and in most cases they show the same patterns of relationship with other gauges.

While the computation unit usually does not influence the information provided by the particular measure, the choice of an algorithm to capture a particular aspect of lexical proficiency has a larger effect on the obtained results for individual texts. The most robust correlations can be found between the ratio- and mean-based indices representing lexical sophistication, i.e. between the proportions of beyond-2K items and the mean log frequency of content word tokens (cf. Crossley, Cobb, & McNamara, 2013). However, only moderate relationships can

Table 6.9 Correlations Between Lexical Indices

	LexTy	*GUIR_ Ty*	*GUIR_ Lex*	*D*	MTLD	*2K_To*	*2K_Lex*	AWL_To	AWL_Ty	*FQLog_ AW*	*FQLog_ CW*	LD	*FAM*	CNC	IMG	*MEA*	POL	HYP
Lgth	.700**	.299**	.238**															
LexTy	*1*	***.844****	***.823****	*.340***	*.366***	*.383***	*.524***	.393**	.484**		*–.332***	.204*	*–.336***	–.209*	–.181*		–.317**	
GUIR_Ty		*1*	***.984****	*.642***	.673**	*.628***	*.669***	.510**	.519**	*–.296***	*–.519***	.221**	*–.400***	–287**	–.257**	*–.186**	–.453**	–.218**
GUIR_Lex			*1*	*.626***	.686**	*.677***	*.697***	.560**	.568**	*–.304***	*–.559***	.207*	*–.420***	–.316**	–.294**	*–.208**	–.518**	–.236**
D				*1*	**.851****	*.395***	*.338***	.257**	.208*	*–.499***	*–.314***	.288**		–.220**	–.200*		–.183*	–.165*
MTLD					1.000	.576**	.530**	.379**	.350**	–.448**	–.415**	.258**	–.226**	–.307**	–.322**	–.190*	–.343**	–.229**
2K_To						*1*	***.902****	.501**	.459**	*–.445***	***–740****	.176*	*–.517***	–.391**	–.377**	*–.241***	–.650**	–.341**
2K_Lex							*1*	.480**	.473**	*–.441***	***–.743****	.191*	*–.527***	–.314**	–.305**	*–.226***	–.594**	–.290**
AWL_To								1.000	**.946****	–.317**	–.452**	.300**	–.505**	–.244**	–.257**	–.170*	–.481**	
AWL_Ty									1.000	–.229**	–.408**	.237**	–.478**	–.270**	–.280**	–.205*	–.468**	
FQLog_AW										*1*	***.710****	–.611**	*.434***	–.203*	–.189*		.184*	–.228**
FQLog_CW											*1*	–.431**	*.675***				.480**	
LD												1.000	–.244**	.199*	.176*	.243**		.349**
FAM													*1*		–.074	*.137*	.439**	–.229**
CNC														1.000	**.945****	.674**	.296**	**.760****
IMG															1.000	**.742****	.342**	**.768****
MEA																*1*	.302**	.518**
POL																	1.000	.363**

Note: Italics—Pearson moment correlation; normal font—Spearmen rank-order correlation; bold—strong correlations; ** p<0.01; * p<0.05

Table 6.10 Correlations Between Phraseological Indices

	BiAbs_Ty	BeTh_ To	BeTh_ Ty	*Non Mi_To*	*Low MI_To*	*MidMI_ To*	Hi MI_To	*Non Mi_Ty*	*Low MI_Ty*	*MidMI_ Ty*	Hi MI_Ty	*NonT_ To*
BiAbs_To	**.990****	**.864****	**.849****					–.192*		.192*	.296**	
BiAbs_Ty	1.000	**.865****	**.863****					–.167*		.188*	.272**	
BeTh_To		1.000	**.993****			.184*		–.256**		.221**	.282**	
BeTh_Ty			1.000			.178*		–.237**		.212**	.258**	
NonMi_To				*1*	**–.814****	*–.489***	–.201*	***.907****	***–.832****	*–.473***	–.187*	*.466***
LowMI_To					*1*		–.251**	***–.731****	***.918****			*–.380***
MidMI_To						*1*		*–.530***	*.167**	***.912****		*–.239***
HiMI_To							1.000				.581**	
NonMi_Ty								*1*	***–.856****	*–.566***	–.304**	*.558***
LowMI_Ty									*1*	*.170**		*–.458***
MidMI_Ty										*1*		*–.312***
HiMI_Ty											1.000	–.230**
NonT_To												*1*
LowT_To												
MidT_To												
HiT_To												
NonT_Ty												
LowT_Ty												
MidT_Ty												
HiT_Ty												
LogBiFQ_To												
LogBiFQ_Ty												
BiMI_To												
BiMI_Ty												
BiT_To												

Note: Italics—Pearson moment correlation; normal font—Spearmen rank-order correlation; bold—strong correlations; ** p<0.01; * p<0.05

	Low T_To	*MidT_ To*	*Hi T_To*	NonT_ Ty	Low T_Ty	*MidT_ Ty*	*Hi T_Ty*	*LogBi FQ_To*	*LogBi FQ_Ty*	*Bi MI_To*	Bi MI_Ty	BiT_To	BiT_Ty
BiAbs_To	.622**	.275**	−.438**		.597**	.262**	−.348**	−.619**	−.611**		.302**		
BiAbs_Ty	.593**	.256**	−.439**		.578**	.247**	−.367**	−.596**	−.613**		.267**		
BeTh_To	.599**	.302**	−.411**		.579**	.281**	−.317**	**−.726****	**−.751****		.335**		
BeTh_Ty	.579**	.277**	−.414**		.568**	.260**	−.336**	**−.703****	**−.750****		.301**		
NonMi_To			*−.322***	.429**			*−.326***			***−.764****	−.661**	−.315**	−.411**
LowMI_To			*.377***	−.338**			*.354***	*.210***	*.164**	*.400***	.410**	.417**	.382**
MidMI_To	.225**			*−.253***	.232**					*.379***	.453**		.190*
HiMI_To								−.189*		.615**	.248**	−.174*	
NonMi_Ty	-.239**		*−.275***	.561**	−.223**		*−.363***			***−.713****	**−.820****	−.256**	−.354**
LowMI_Ty			*.364***	−.472**			*.391***			*.520***	.579**	.319**	.388**
MidMI_Ty	.208*	*.219***		−.312**	.217**			*−.200**		*.400***	.480**		
HiMI_Ty	.356**	.293**		−.286**	.326**	.258**		−.367**	−.318**	.410**	.519**		
NonT_To	−.178*	*−.182**	*−.610***	**.971****	−.161*	*−.181**	*−.661***			*−.649***	**−.723****	−.350**	−.404**
LowT_To	1.000	.285**	−.528**	−.243**	**.984****	.282**	−.459**	−.658**	−.657**	.192*	.358**		
MidT_To		*1*	*−.444***	*−.231***	*.259***	***.936****	*−.338***	*−.500***	*−.469***		.271**	−.272**	−.225**
HiT_To			*1*	*−.506***	*−.523***	*−.420***	***.948****	*.632***	*.648***	*.339***	.248**	.460**	.480**
NonT_Ty				1.000	−.214**	−.227**	−.601**			−.607**	**−.785****	−.276**	−.394**
LowT_Ty					1.000	.253**	−.479**	−.632**	−.654**	.194*	.330**		
MidT_Ty						*1*	*−.373***	*−.469***	*−.464***	*.166**	.251**	−.295**	−.278**
HiT_Ty							*1*	*.514***	*.599***	*.361***	.365**	.410**	.525**
LogBiFQ_To								*1*	***.941****		**−.276****	.460**	.369**
LogBiFQ_Ty									*1*		−.196*	.346**	.447**
BiMI_To										*1*	**.778****		.360**
BiMI_Ty											1.000	.201*	.372**
BiT_To												1.000	**.700****

Table 6.11 Correlations Between Lexical and Phraseological Indices

	Lgth	*LexTy*	*GUIR_Ty*	*GUIR_Lex*	*D*	MTLD	*2K_To*	*2K_Lex*	AWL_To	AWL_Ty	*FQLog_AW*	*FQLog_CW*	LD	*FAM*	CNC	IMG	*MEA*	POL	HYP
BiAbs_To		.349**	.547**	.555**	.497**	.566**	.683**	.647**	.381**	.342**	−.550**	−.597**	.364**	−.381**	−.265**	−.217**		−.346**	
BiAbs_Ty		.363**	.536**	.545**	.474**	.540**	.670**	.640**	.385**	.354**	−.521**	−.587**	.351**	−.398**	−.270**	−.223**		−.351**	
BeTh_To		.367**	.565**	.584**	.457**	.557**	**.713****	**.709****	.445**	.407**	−.580**	−.681**	.426**	−.440**	−.272**	−.269**	−.171*	−.409**	
BeTh_Ty		.372**	.538**	.556**	.420**	.515**	.687**	.692**	.438**	.406**	−.561**	−.663**	.423**	−.449**	−.265**	−.265**	−.174*	−.400**	
NonMi_To	.187*				*−.165**	−.212**	*−.162**				*.274***								
LowMI_To													−.218**		−.280**	−.285**			−.316**
MidMI_To	−.232**		*.180**	*.202**	*.423***	*.445***	*.254***				−.272**		.168*		−.167*	−.209*		−.190*	−.193*
HiMI_To											−.366**	−.236**	.407**	−.300**	.461**	.405**	.182*		.501**
NonMi_Ty	.244**	*−.161**	*−.215***	*−.266***	*−.302***	−.409**	*−.373***	*−.286***	−.251**	−.168*	*.289***	*.210**			.221**	.250**	*.165**	.280**	.245**
LowMI_TY	−.168*					.226**	*.175**								−.240**	−.262**	*−.176**	−.167*	−.293**
MidMI_Ty	−.258**		*.185**	*.223***	*.350***	.387**	*.323***	*.177**	.190*		*−.271***	*−.225***	.180*			−.190*		−.248**	−.167*
HiMI_Ty		.178*	.300**	.321**	.240**	.312**	.376**	.336**	.320**	.245**	−.383**	−.417**	.319**	−.398**				−.170*	.191*
NonT_To				*−.200**			*−.331***	*−.273***	*−.225***	−.171*					.174*	.216**		.288**	.179*
LowT_To		.285**	.488**	.507**	.500**	.508**	.602**	.594**	.408**	.331**	−.625**	−.632**	.361**	−.462**				−.372**	
MidT_To			*.279**	*.310**	*.218**	.191*	*.416**	*.413***	.258**	.214**	*−.440***	*−.487***	*.255***	*−.327***					
HiT_To		*−.289**	*−.275***	*−.255***	*−.289***	−.252**	*−.263***	*−.298***	−.162*		*.453***	*.467***	−.314**	*.340***					−.172*
NonT_Ty	.181*		−.217**	−.252**	−.176*	−.267**	−.397**	−.336**	−.242**	−.175*	.191*	.212**			.185*	.222**		.321**	.197*
LowT_Ty		.276**	.448**	.469**	.448**	.460**	.581**	.577**	.394**	.332**	−.600**	−.620**	.338**	−.472**				−.375**	
MidT_Ty			*.274**	*.295**	*.230**	.195*	*.391**	*.419***	.237**	.198*	*−.462***	*−.477***	*.253***	*−.307***					
HiT_Ty		*−.301**	*−.182*					*−.206**			*.382***	*.371***	−.265**	*.322***					−.214**
LogBiFQ_To		*−.357**	*−.555***	*−.561***	*−.507***	−.524**	*−.579***	*−.573***	−.511**	−.443**	***.768****	***.732****	−.625**	*.544***				.321**	
LogBiFQ_Ty		*−.465**	*−.592**	*−.592**	*−.428**	−.465**	*−.577***	*−.626***	−.541**	−.499**	*.680***	***.726****	−.544**	*.604***				.332**	
BiMI_To						.176*	*.221***	*.198**	.173*		*−.485***	*−.280***	.268**	*−.197**	.167*				
BiMI_Ty	−.244**		.362**	.402**	.395**	.486**	.549**	.479**	.373**	.276**	−.469**	−422**	.214**	−.258**		−.178*		−.396**	
BiT_To		−.162*			−.179*						.304**	.171*	−.391**		−.298**	−.333**	−.315**		−.311**
BiT_Ty	−.323**	−.366**	−.243**	−.220**						−.185*		.175*	−.232**	.211**		−.161*	−.180*		−.209*

Note: Italics—Pearson moment correlation; normal font—Spearmen rank-order correlation; bold—strong correlations; ** p<0.01; * p<0.05

be observed between simple and complex measures of lexical diversity (cf. Jarvis, 2002; McCarthy & Jarvis, 2010). These observations may cast doubt on the concurrent validity of the indices claimed to tap into lexical variation of a text.

As far as formulaicity is concerned, much fewer correlations can be observed between the indices. This is to some extent explicable, because these gauges are based on three different types of information: frequency and two association measures. Each of these types highlights very different kinds of word combination: novel word strings, frequent and conventional expressions as well as infrequent but strongly associated word pairs. However, some associations can be observed between individual indices within each of the categories. There is a strong positive association between the proportion of all word combinations whose MI score is above the cut-off point of 3[6] and the mean MI score calculated for all bigrams in a text, which confirms the concurrent validity of these two indices. Weaker, but statistically significant correlations can be observed between the proportions of items belonging to the individual strength-based bands of MI collocations and the mean MI score. However, in the case of indices derived from the t-score a less clear picture can be observed between individual ratio-based indices and the mean-based gauge. In addition, the analysis casts doubt on the validity of some of the cut-off points used to classify bigrams into several strength-related bands. In particular, the boundary between two kinds of creative word combinations—those absent from the reference corpus and those with very small frequency is problematic as the two indices are strongly and positively correlated. Similarly, the boundaries defining mid-strength bands for both MI and t collocations may be questioned. This observation was also made by Durrant and Schmitt (2009) and Granger and Bestgen (2014).

Several interrelations have been observed between distinct aspects of lexical proficiency. First, lexical productivity, as captured by the number of content types, correlates with all other dimensions of lexical proficiency except for phraseological complexity tapped by bigram MI scores. Lexical diversity correlates moderately with lexical sophistication. Such effect was already demonstrated in earlier studies (Daller & Xue, 2007). These results confirm Jarvis's (2013) definition of word diversity which encompasses lexical sophistication (rarity; see Chapter 3 for a discussion) and confirms the claim that lexical diversity and lexical sophistication are different but interrelated constructs describing the vocabulary of a text. Lexical sophistication and diversity show a fairly complex relationship with the formulaicity of a text. The essays with more advanced and varied vocabulary tend to contain a larger proportion of creative (novel or very infrequent) word combinations. In addition, such essays contain a larger proportion of less common collocations (LowT) and a wider range of strongly associated collocations (HiMI_Ty). In other words, it can be said that less advanced and varied vocabulary goes in hand with larger

conventionality in language use, the overuse of highly clichéd collocations and a poor range of strongly associated word pairs.

The analysis has also demonstrated that although each index representing psychological properties or meaning relations of words in a text captures a different aspect of lexical elaborateness, these aspects are interrelated and link to two common underlying features: specificity/concreteness/imagability/meaningfulness on the one hand, and perceived frequency/polysemy on the other. The latter trait is also moderately associated with indices of lexical sophistication and diversity, thus suggesting that information derived from objective corpus-based counts does not differ much from human intuition on the pervasiveness of individual lexical items. The former group of features also correlates (negatively) with the measures of lexical sophistication and diversity. These associations are weak, but they indicate that the essays with more sophisticated and varied vocabulary contain fewer concrete, imaginable, meaningful and specific words.

Lexical density produces mostly weak correlations with the indices tapping lexical productivity, lexical sophistication and diversity. This suggests that the texts whose lexis is more advanced and varied are also more lexically dense.

On the whole, the matrix of correlations between the analysed indices validates the claims about the existence and interdependence of various aspects of lexical proficiency which were discussed in Chapter 1. With a few exceptions, it also confirms the validity of the formulas proposed to measure these aspects.

6.5.2 Analysis 2—Comparison of the Indices Between the Groups

Preliminaries

The three groups of subjects whose essays were analysed in this study represent three different proficiency levels. It is expected that the lexical and phraseological measures computed for the essays will reflect this difference. Due to the nature of individual indices, their mean values are predicted to either increase or decrease between the groups. Thus, the hypothesis underlying the subsequent analyses can be expressed by the following formulas:

$$M_{Year1} < M_{Year4} < M_{Native}$$
$$M_{Native} < M_{Year4} < M_{Year1}$$

However, the measures of lexical productivity can also depend on the time available for producing a text. Since it was shorter for the native group, the following hypothesis can be proposed:

$$M_{Native} < M_{Year1} < M_{Year4}$$

Data Analysis

The mean values of the 19 lexical indices and the 26 phraseological indices were compared across the three groups of subjects. Almost two-thirds of the indices were normally distributed for the three groups, as assessed by Shapiro-Wilk's test ($p>0.05$)[7]. The assumption of homogeneity of variances was met by all but six variables, as assessed by Levene's test ($p>0.05$). One-way ANOVA was applied in order to trace the differences between the groups. Since this test is fairly robust to deviations from normality, particularly for equal-sized samples (Lix, Keselman, & Keselman, 1996; Maxwell & Delaney, 2004), the statistic was also applied to the variables which failed the condition of normality. However, Welsh's ANOVA was computed for the six variables which did not meet the assumption of homogeneity of variances. Finally, post hoc comparisons between the groups were performed using the Tukey procedure (for the regular one-way ANOVA) or the Games-Howell procedure (for the Welch's ANOVA). The effect size was computed for each statistically significant difference. Its negative values indicate the effect opposite to the underlying hypothesis for a particular index. The results of the tests, together with the information on the effect size in the population, estimated by omega squared,[8] and effect sizes for individual statistically significant differences estimated by Cohen's d (Cohen, 1998), are presented in Table 6.12 and Table 6.13.

The results indicate that out of the 19 lexical indices analysed in the study, one type of lexical diversity measures—the Giraud index—discriminated well between the three subject groups, in each case with a large effect ($d>0.8$). Its robustness was additionally confirmed by the fact that the same effects were noted in the case of two different units of calculation—types (GUIR_Ty) and lexemes (GUIR_Lex). Three measures of lexical sophistication also proved their good discriminatory power, but with less pronounced effects: LogFQ_CW, AWL and 2K_Lex (only lexemes). The last of these, however, turned out to be less successful when tokens were used for calculations (2K_To), and discriminated only between the learner and native texts. Both gauges of lexical productivity (Lgth, LexTy) and word familiarity (FAM)—which also relates to words frequency, but not in objective but perceptual terms—revealed the differences between the two learner groups and between native students and one group of learners (Year 4 or Year 1), with varying effect sizes. The remaining lexical indices with a statistically significant ANOVA result registered large-effect differences between the native and learner essays, but failed to discriminate between the two groups of learners. It should be noted that while the observed differences between the two learner groups confirmed the hypothesised direction of change, all the differences between native and learner essays displayed the opposite pattern.

None of the 26 indices depicting the use of phraseology in the analysed texts demonstrated a difference between Year 1 and Year 4 students. All

Table 6.12 Results of One-Way ANOVA for All the Lexical Indices

Lexical Indices									
Index	*Hypothesis*	*ANOVA*		*Differences of Means*			*Effect Size (d)*		
		F	ω^2	*Y1 vs. Y4*	*Y4 vs. N*	*Y1 vs. N*	*Y4–Y1*	*N-Y4*	*N-Y1*
Lgth	Y4>Y1>N	7.79**	**.08**	60.34**	−9.6	50.740*	.76		**−.65**
LexTy	Y4>Y1>N	624.25**	**.23**	24.16**	−32.88**	−8.72	**.91**	**1.28**	
GUIR_Ty	N>Y4>Y1	41.00**	**.35**	.67**	−1.47**	−.8**	**.86**	**−1.78**	**−.96**
GUIR_Lex	N>Y4>Y1	52.87**	**.41**	.63**	−1.59**	−.95**	**.84**	**−2.04**	**−1.20**
D	N>Y4>Y1	11.08**	**.12**	2.08	−14.26**	−12.18**		**−.85**	**−.78**
MTLD	N>Y4>Y1	*25.59***	**.25**	4.44	−23.77**	−19.33**		**−1.31**	**−1.17**
2K_To	N>Y4>Y1	84.79**	**.53**	1.18	−6.67**	−5.49**		**−2.38**	**−2.09**
2K_Lex	N>Y4>Y1	*50.68***	**.40**	2.82**	−8.19**	−5.38**	**.65**	**−1.94**	**−1.39**
AWL_To	N>Y4>Y1	*20.13***	**.20**	.81*	−1.7**	−.89**	**.54**	**−1.23**	**−.74**
AWL_Ty	N>Y4>Y1	*18.19***	**.19**	1.45**	−2.71**	−1.26**	**.60**	**−1.18**	**−.65**
FQLog_AW	Y1>Y4>N	.98							
FQLog_CW	Y1>Y4>N	12.78**	**.14**	.05*	−.1**	−.05*	**.54**	**−1.02**	−.46
LD	N>Y4>Y1	.42							
FAM	Y1>Y4>N	*5.62***	.06	2.40*	−3.11**	−.7	**.47**	**−.65**	
CNC	Y1>Y4>N	*33.44***	**.30**	2.92	−26.29**	−23.37**		**−1.56**	**−1.45**
IMG	Y1>Y4>N	*40.27***	**.34**	.11	−25.41**	−25.3**		**−1.64**	**−1.70**
MEA	Y1>Y4>N	10.81**	**.12**	2.87	−9.38**	−6.51*		**−.85**	**−.61**
POL	Y1>Y4>N	46.22**	**.38**	.05	−.57**	−.52**		**−1.76**	**−1.49**
HYP	Y1>Y4>N	41.56**	**.35**	−.03	−.27**	−.29**		**−1.41**	**−1.58**

Note: Normal font—ANOVA; italics—Welch ANOVA; * p<0.05; ** p<0.01; bold—large or medium effect size; grey shading—the difference in accordance with the hypothesis

Table 6.13 Results of One-Way ANOVA for the Phraseological Indices

Phraseological Indices									
Index	*Hypothesis*	*ANOVA*		*Differences of Means*			*Effect Size (d)*		
		F	ω^2	*Y1 vs. Y4*	*Y4 vs. N*	*Y1 vs. N*	*Y4-Y1*	*N-Y4*	*N-Y1*
BiAbs_To	Y1>Y4>N	*14.86***	**.16**	.00	.01**	.01**		**1.00**	**.88**
BiAbs_Ty	Y1>Y4>N	*12.51***	**.13**	.00	.01**	.01**		**.97**	**.85**
BeTh_To	N>Y4>Y1	17.39**	**.18**	.01	−.01**	−.01**		**−1.12**	**−.89**
BeTh_Ty	N>Y4>Y1	16.05**	**.17**	.01	−.01**	−.01**		**−1.06**	**−.85**
NonMi_To	Y1>Y4>N	14.64**	**.15**	.00	−.03**	−.03**		**−.98**	**−.91**
LowMI_To	Y1>Y4>N	3.95*	.04	.00	.01	.02*			.52
MidMI_To	Y1>Y4>N	4.86*	.05	.00	.006	.01*			**.60**
HiMI_To	N>Y4>Y1	8.85**	**.09**	.00	.01**	.01**		**.64**	**.74**
NonMi_Ty	Y1>Y4>N	32.73**	**.30**	.01	−.05**	−.04**		**−1.48**	**−1.35**
LowMI_Ty	Y1>Y4>N	5.68**	.06	.00	.01*	.02**		.51	**.65**
MidMI_Ty	Y1>Y4>N	9.41**	**.10**	.00	.01**	.01**		**.66**	**.83**
HiMI_Ty	N>Y4>Y1	1.91							
NonT_To	Y1>Y4>N	18.80**	**.19**	.01	−.03**	−.02**		**−1.16**	**−.81**
LowT_To	N>Y4>Y1	6.34*	.07	.01	−.01**	−.006		**−.74**	
MidT_To	N>Y4>Y1	5.08*	.05	.01	−.01*	−.002		**−.59**	
HiT_To	Y1>Y4>N	2.01							
NonT_Ty	Y1>Y4>N	24.75**	**.24**	.01	−.03**	−.03**		**−1.29**	**−.98**
LowT_Ty	N>Y4>Y1	5.46*	.06	.01	−.01**	−.006		**−.68**	
MidT_Ty	N>Y4>Y1	3.64*	.03	.01	−.01*	−.0008		−.48	
HiT_Ty	Y1>Y4>N	.74							
LogBiFQ_To	Y1>Y4>N	8.74**	**.09**	.04	−.11**	−.07*		**−.82**	−.46
LogBiFQ_Ty	Y1>Y4>N	10.68**	**.11**	.05	−.12**	−.07*		**−.94**	−.48
BiMI_To	N>Y4>Y1	1.34							
BiMI_Ty	N>Y4>Y1	21.63**	**.22**	.07	−.24**	−.17**		**−1.19**	**−.90**
BiT_To	Y1>Y4>N	4.38*	.04	2.75	7.06	9.82*			**−.56**
BiT_Ty	Y1>Y4>N	1.41							

Note: Normal font—ANOVA; italics—Welch ANOVA; * p<0.05; ** p<0.01; bold—large or medium effect size; grey shading—the difference in accordance with the hypothesis

but five gauges (HiMI_Ty, HiT_To, HiT_Ty, BIMI_To, BiT_Ty) revealed differences between native subjects on the one hand, and both or at least one learner group on the other. These differences were large in 20 cases, medium (d>0.5) in ten cases and small (d>0.3)—but on the border line with medium—in five cases. As in the case of the lexical measures, most of the differences between learners and native subjects ran against the hypothesised pattern. However, in the case of the following measures—BiAbs, LowMi, MidMI, HiMI_To (tokens only)—the observed differences between native and learner texts were consistent with the hypothesis set prior to the analysis.

The indices capturing the proportions of words and bigrams absent from the corpus require a more detailed scrutiny, as this category can contain qualitatively different types of bigrams: unusual and creative word associations as well as errors. For the interpretation of the results it would be helpful to know which types of bigrams are more frequent in this band. A total of 1330 bigram types and 1444 bigram tokens were absent from the reference corpus. A random sample of 150 concordance lines generated for these items was analysed manually. In the sample, only 29 (19%) bigrams were erroneous (e.g. *scientist find** and *got addtict**), a few others were indeed creative word pairings (e.g. *ringing friend* or *infernal development*), but the great majority were bigrams overlapping two adjacent syntactic units (e.g. *negatively since* or *curfew they*). In the pool of the erroneous items, 13 (45%) were found in Year 1 essays, 10 (35%) in Year 4 essays and 3 (10%) in native compositions.

Discussion

As discussed in Section 6.3.1, the three groups of participants formed very homogeneous groups. These groups were assumed to be distributed along a proficiency scale: Year 1—Year 4—Native. Positioning EFL learners on the scale was not based solely on their institutional status but also on the outcomes of their university exams (see Section 6.3.1). The language proficiency of native students was not assessed, yet assigning them the highest rank perpetuated the belief underlying all research in SLA that the ultimate—and in most cases unattainable—aim of language learning is native-like proficiency, and that second language acquisition stops short of native norms of language knowledge and use (cf. Coppieters, 1987; Medgyes, 1992; Gass & Selinker, 2001, p. 12; Cook, 2002). Admittedly, several more recent studies have demonstrated that language proficiency of native speakers also varies "mainly because of differences in age, intellectual skills, education, occupation, and leisure-time activities" (Hulstijn, 2011, p. 244). Such differences are particularly visible at the lexical and collocational level. Rich vocabulary and phraseology is intrinsically linked with the educational level and the knowledge of the world (Mainz, Shao, Brysbaert, & Meyer, 2017). Likewise, writing skills, especially an ability to produce academic genres, are also to a great extent a result of training

(Weigle, 2002, p. 4). Nevertheless, the native participants in this study could be assumed to have a good proficiency in their mother tongue. Similar to the Polish subjects, they were admitted to the university based on the results of one of the American standardised tests for college admission (SAT or ACT) and their high school transcripts. They were thus comparable in terms of age and educational level to the non-native participants, in particular the Year1 students. That is why the hypothesised distribution of participants into the proficiency bands seems justified.

The 45 indices computed for the analysed texts in this study have already been validated in the literature on language assessment and second language acquisition (cf. Chapter 5). All these studies have demonstrated that the values of most of the examined gauges change—increase or decrease—parallel to the proficiency of foreign language learners, or that they differ between native and non-native language users. Meanwhile, the analyses carried out in this study have produced several interesting and unexpected outcomes. Among the lexical indices, two measures—mean log frequency for all words (LogFQ_AW) and lexical density (LD)—failed to demonstrate statistically significant differences between any of the subject groups. Mean log frequency computed for all words, which was shown to discriminate between learners of different proficiency in earlier research (Crossley, Cobb, & McNamara, 2013), may be argued to be less useful in the case of fairly advanced students. Yet, when the function words are taken out of the computation, the differences in mean log frequency of content words in a text become large, as demonstrated by the LogFQ_CW index analysed in the study.

The index of lexical density also did not register the differences between the texts written by the three subject groups. Even though some early research demonstrated that the index had a discriminatory power among texts written by students and native speakers or students at different proficiency levels (cf. Wolfe-Quintero et al., 1998), this effect has never been repeated in more recent research (Linnarud, 1986; Laufer, 1991; Lu, 2012b).

The remaining 17 lexical indices did register differences between the subject groups, but only nine of them revealed differences between the texts written by the learners with different proficiency levels. They were gauges of lexical productivity, lexical diversity and lexical sophistication. They confirmed that texts written by more advanced students are longer and feature more varied and sophisticated vocabulary (cf. Laufer, 1991, 1994; Laufer & Nation, 1995; Daller & Xue, 2007; Tidball & Treffers-Daller, 2007). It is interesting to note that the index which demonstrated the most pronounced effect was a mathematically simple measure proposed early in the literature—Guiraud's index. This effect is particularly noteworthy in view of the fact that Guiraud is claimed to be negatively affected by text length (Malvern et al., 2004) and Year 4 students produced on average longer essays, as demonstrated by this

analysis. Interestingly, more recent and more complex formulas—vocd-D (D) and mean textual lexical diversity (MTLD)—failed to register the differences between the two learner groups.

Most of the measures of lexical sophistication demonstrated a difference between Year 1 and Year 4 students; however the learner texts did not differ in the proportion of advanced tokens, (2K_To). This fact sheds some doubt on the index's robustness in differentiating between two groups of learners at fairly advanced stages of language development. Interestingly, one more index capturing information on word frequency, but in perceptual rather than objective terms—familiarity (FAM)—also registered a difference between the two groups of learners even though its effect was smaller.

The two measures of lexical productivity—the number of tokens and content word types used in the text also discriminated well between the two learner groups. In particular the number of content word types produced the largest effect size among all the indices (cf. Lu, 2012b).

None of the indices summarising the psychological properties and meaning relations of words in a text (except for familiarity) demonstrated differences between Year 1 and Year 4 students. These indices, recently proposed in the literature as gauges of the internal structure of the lexicon (Crossley et al., 2009; Crossley et al., 2010a; Crossley et al., 2011a, 2011b; Salsbury et al., 2011; Crossley et al., 2012), have not been researched as extensively as the more established measures discussed previously. Particularly with more advanced learners they lack a sound theoretical justification, as already pointed out in Chapter 4. However, it should be noted that in all but one case the small changes in their mean values reflected the direction assumed in the hypotheses.

Even though several lexical indices did not register differences between the two groups of learners, almost all of them differentiated between the learner and native texts. The striking result, however, is that all of them noted changes in the direction opposite to intuition as well the theoretically and empirically grounded hypotheses set out at the beginning of the study (cf. Linnarud, 1986; Laufer & Nation, 1995; Jarvis, 2002; Tidball & Treffers-Daller, 2007). They demonstrated that native speakers produced more repetitive texts which used less sophisticated vocabulary, fewer academic words and a larger number of concrete, imaginable, unspecific lexical items. It is particularly surprising in view of the fact that all the subjects wrote essays on exactly the same topic.

The interpretation of the results for the phraseological indices is also quite challenging. Contrary to the results obtained by Granger and Bestgen (2014), the indices capturing the proportions of very rare bigrams (freq<5) as well as the ratios of MI and t-bigrams with different value ranges failed to demonstrate a statistically significant difference between the texts written by Year 1 and Year 4 students. The observed lack of a discriminatory effect can only be explained by the fact that the learners in

Granger and Bestgen's study displayed a much wider range of proficiency levels extending from B1 to C2, which were then collapsed into two broad bands—B and C—for further comparisons. In this study, the subjects formed more homogenous proficiency groups. These groups clearly differed from each other in the use of varied and sophisticated vocabulary, as demonstrated by several gauges of lexical diversity and sophistication. Yet, the differences in their use of phraseology, as captured by the phraseological indices, were too small to be registered by a statistical test. Nevertheless, it is worth pointing out that these slight differences in index means reflected almost perfectly the pattern of differences noted by Granger and Bestgen. The fact that this pattern of differences for a total of 18 indices is consistent with the previous research, even if not statistically significant, points to an existing trend which has to be further confirmed in future studies.

Another set of phraseological indices, i.e. the proportions of bigrams absent from the reference corpus, as well as the means for logarithmically transformed frequencies, MI and t-scores, also did not pick up a statistically significant change between Year 1 and Year 4 essays. Such a discriminatory effect was observed earlier by Granger and Bestgen (2014), but only for mean t-score. A lack of such a result in this study is particularly surprising, as the difference between the collection points was three years rather than three months (in this case, however, it was a pseudo-longitudinal design). Again, as with the ratio-based phraseological indices, all the means exhibited the expected direction of change, yet too small to be registered by a statistical test. Only one pair of measures showed differences opposite to the expected trend. Year 4 students used more bigram types and tokens absent from the reference corpus than Year 1 students. The hypothesis, which predicted the opposite effect assumed that the bigrams absent from the corpus were mainly errors. Yet the analysis of a random bigrams from both groups proved that error constituted not more than 20% of all the bigrams. The others were correct but unattested word choices.

On a more general level, the results for Year 1 and Year 4 indicate that the phraseological proficiency in advanced learners of English develops in the direction proposed in the hypotheses. Their writing becomes less stereotypical and more novel in word choice, yet at the same time they demonstrate a better mastery of strongly associated word pairings which contribute to the impression of formulaicity of a text. However, this development takes place at a slower rate than the purely lexical proficiency and it needs to be confirmed in future studies as the statistical significance has not been reached by any of the indices.

As in the case of the lexical indices, many phraseological gauges demonstrated a difference between the essays written by the learners and the native students. However, the expected direction of the change—based on previous results and theoretical considerations—was only confirmed

for selected categories of indices. The native students indeed used fewer bigrams which were absent from the corpus, which category includes both errors as well as highly creative word combinations, but they also employed fewer very rare bigrams. Contrary to the hypotheses they used larger proportions of non-collocational bigrams defined by both MI and t-scores, as well as more frequent word combinations.

To sum up, two main observations emerge from the analysis. First, most of the measures of lexical productivity, diversity and sophistication register a growth in lexical proficiency between two groups of advanced learners at uneven levels. The measures of lexical elaborateness do not demonstrate a statistically significant increase, but the small changes in their values occur in the expected direction. Second, almost all the measures used in the study show a difference between the learners and the native students, and with much larger size effects than in the case of differences between the learner groups. However, these changes are contrary to intuitions and the expectations based on the theoretical assumptions and the results of the earlier research.

6.5.3 Analysis 3—Prediction of Group Membership Based on Selected Indices

Preliminaries

The aim of the multinomial logistic regression performed in this study was to explore if the indices describing the lexical features of texts can jointly predict group membership of their authors. Since group membership was assumed to be a proxy for the subjects' proficiency, this analysis attempted to establish if automatic indices taken together can gauge lexical proficiency.

In the sample of only 150 essays, entering all 45 measures into the model would lead to its overfit. This effect—coupled with the collinearity of the variables demonstrated in Analysis 1—would hamper a meaningful interpretation of the results, especially the statistical significance of individual independent variables. That is why only eight indices were chosen for the model. They were selected on the basis of earlier theoretical considerations (Chapter 4) and empirical analyses (Chapter 5 and the present chapter). Each index was considered to represent best one dimension of lexical proficiency. The values of all the eight independent variables were standardised. The indices selected for the model are presented together with their short description in Table 6.14.

Data Analysis

The likelihood ratio test confirmed that the multinomial logistic regression model fitted the data significantly better than the null model

Table 6.14 The Indices Selected for the Multinomial Logistic Regression

Index	*Dimension of Lexical Proficiency*	*Information Expressed by the Index*
LexTy	productivity	the number of different content words used in the essay
GUIR_Lex	diversity	lack of repetition—use of varied vocabulary
2K_Lex	sophistication	general sophistication—use of advanced words
AWL_Ty	sophistication	register-specific sophistication—use of academic words
HYP	elaborateness	paradigmatic relations among words—use of specific words
BiAbs_Ty	elaborateness	syntagmatic relations among words—use of creative collocations
NonMI_Ty	elaborateness	syntagmatic relations among words—use of well-associated collocations
HiT_Ty	elaborateness	syntagmatic relations among words—use of very frequent collocations

Table 6.15 Predicted Classification of the Essays Made by the Regression Model

	Pred.: Year 1	*Pred.: Year 4*	*Pred.: Native*	*Percent Correct*
Obs.: Year 1	36	11	3	72.0%
Obs.: Year 4	18	31	1	62.0%
Obs.: Native	3	0	47	94.0%
Percent total	63.2%	73.8%	92.2%	76.0%

($\chi^2(16)=182.05$, $p<0.0001$). Nagelkerke's pseudo R^2 was 0.791, which indicates a good fit. The model correctly classified 76% of the essays. Table 6.15 presents the classification made by the equation.

As can be seen in Table 6.15, the model predicted group memberships for the native essays with a precision reaching 94%. Only three essays written by the American students were not classified accurately. The prediction of group membership for Year 1 and Year 4 essays turned out to be more problematic. Only 72% of the former and 62% of the latter were classified correctly. A great majority of misclassifications occurred between the two learner groups.

Table 6.16 presents the contribution of each independent variable to the model and its statistical significance.

Of the eight indices entered into the model, none was statistically significant in discriminating between Year 1 and Year 4 essays. Only three indices were statistically significant in the classification of Year 4 and

Table 6.16 The Contribution of Each Index to the Model

Group (Year 4 was set as reference)		*B*	*S.E.*	*p (in Wald's test)*
Year 1	Intercept	.528	.339	.119
	LexTy	–.876	.522	.093
	GUIR_Lex	.049	.687	.943
	2K_Lex	–.502	.384	.191
	AWL_Ty	–.290	.273	.290
	HYP	–.023	.328	.943
	BiAbs_Ty	–.140	.383	.714
	NonMi_Ty	.148	.306	.628
	HiT_Ty	–.426	.406	.294
Native	Intercept	–1.695	.798	.034*
	LexTy	–.110	.985	.911
	GUIR_Lex	–2.184	1.230	.076
	2K_Lex	–2.270	.926	.014*
	AWL_Ty	–1.199	.656	.068
	HYP	3.371	.973	.001**
	BiAbs_Ty	–1.793	1.116	.108
	NonMi_Ty	1.240	.926	.180
	HiT_Ty	–1.812	.831	.029*

Note: * p<0.05; ** p<0.01

The upper section contains model parameters in the equation predicting whether an essay was written by a Year 1 student rather than by a Year 4 student. The lower section contains model parameters in the equation predicting whether an essay was written by a native student rather than by a Year 4 student.

native texts: 2K_Lex, HYP and HiT_Ty. An increased probability of a text being written by a native student was associated with increasing hypernymy of its vocabulary as well as a decreasing proportion of advanced words and highly frequent collocations. It should be noted that the magnitude of the effect of some other indices (expressed by the value of standardised regression coefficients) is large, however not statistically significant. In particular, the effect of LexTy approaches the statistical significance in discriminating between Year 1 and Year 4 essays. This indicates that an increased probability of a text being written by a Polish student with a lower proficiency level was linked with a decreasing lexical productivity. In the same way the coefficient for GUIR_Lex approached the statistical significance in classifying native essays. Its value is large, but the direction of its effect is opposite of what is expected.

Discussion

The multinomial regression analysis of the data has demonstrated that the selected lexical indices are more successful in discriminating between the

native and learner essays than between the two groups of learners. Only a few indices have a statistically significant effect on predicting the authorship of the individual essays. This can cast doubt on whether this model could be applied equally successfully to classify some other comparable essays not included in the analysed dataset. On the other hand, the power of the tests is low because of the collinearity of the variables (as evidenced in the earlier analysis) and limited sample size. In consequence, weak or even moderately strong effects are hard to confirm as statistically significant. Taking this into account, scarcity of statistically significant effects should not be considered as clear evidence that the model results cannot be generalised. It should also be noted that out of the three indices predicting the classification of the essays as written by a native rather than a Year 4 student, two metrics (2K_Lex and HYP) had an effect contrary to the expected, which confirms the results of Analysis 2.

6.5.4 Analyses 4 and 5—Relationships Between the Raters' Grades and Their Comparison Between the Groups

Preliminaries

Since the judges had not been provided with rating scales for the two rounds of assessment and since they had evaluated three qualitatively different groups of students on the same scale, the standard of inter-rater reliability applied in regular testing situations was not expected. Yet, in the assessment of vocabulary, where the raters had been asked to focus solely on one aspect of the essays, a higher degree of agreement between the judges was anticipated. It was also assumed that in both rounds of assessment the raters would tend to award the lowest marks to Year 1 students and highest marks to the native subjects, with Year 4 writers' grades falling in the middle of the scale.

Data Analysis

First, inter-rater reliability between the grades awarded in two rounds of assessment was explored. In the holistic evaluation, only in ten cases (13%) the four marks differed by no more than one band (1 point), and in 36 cases (48%) by no more than two bands (2 points). Krippendorff's alpha=0.267 was low and confirmed a low level of inter-rater consensus. In the second round of assessment, the four raters achieved an agreement of no more than one band (1 point) in the case of 23 (31%) essays and a consensus of no more than two bands (2 points) in the case of 49 (65%) essays. Krippendorff's alpha=0.292 was also low and again pointed to unsatisfactory level of inter-rater consensus. However, it should be noted that the agreement between the raters, as expressed by percentages of overlapping grades, was much higher in the assessment of vocabulary than in the first round.

Table 6.17 Correlations Between the Grades in the Two Rounds of Assessment

	N1	*N2*	*P1*	*P2*
Holistic grades				
N1	1.000	.404**	.417**	.360**
N2		1.000	.266*	.380**
P1			1.000	.544**
P2				1.000
Vocabulary grades				
N1	1.000	.488**	.484**	.226
N2		1.000	.504**	.398**
P1			1.000	.372**
P2				1.000

** $p<0.01$

In addition to inter-rater consensus, correlations between the scores were examined. Since the grades awarded by all the four raters in both rounds of assessment did not follow a normal distribution (Shapiro-Wilk $p<0.05$), Spearman's rank-order correlations were computed. Their results are presented in Table 6.17.

Except for one pair of raters (N1 and P1) in the second round of assessment, all the correlations are statistically significant. The relationships between the raters are moderate and weak (in three cases). In spite of a relatively low strength of correlations between the rater pairs, Cronbach's alpha computed for the four sets of scores obtained in the first and second round of assessment was 0.706, and 0.731 respectively, which indicates a satisfactory level of inter-rater consistency. This statistics demonstrates that although the consensus between the raters is low, their grades vary consistently (co-vary).

One-way ANOVA tests[9] were applied in order to establish whether there are statistically significant differences between the marks assigned by the raters to the three groups in the two rounds of assessment. Not all the scores had homogeneous variances in the three groups, as diagnosed with Levene's tests ($p<0.05$). For these cases Welch ANOVA was used. Table 6.18 presents the results of the comparisons together with the effect sizes (ω^2). They indicate that the differences between the holistic and vocabulary marks in the three groups are statistically significant for all but one rater and the effect sizes are medium and large. Only the vocabulary grades awarded by N1 fail to discriminate in a statistically significant way between the essays written by the three groups of students. Post hoc analyses were performed in order to establish which pairwise differences between the groups were statistically significant. The results of the Tukey test, and Games-Howell's test in the case of Welch ANOVA, are reported in Table 6.18.

Table 6.18 Results of One-Way ANOVA for the Grades in the Two Rounds of Assessment

Rater	*ANOVA*		*Differences of Means*			*Effect Size (d)*		
	F	ω^2	*Y1 vs. Y4*	*Y4 vs. N*	*Y1 vs. N*	*Y4-Y1*	*N-Y4*	*N-Y1*
Holistic Grades								
N1	7.15**	**.141**	.80**	–.84**	–.04	**.92**	**–1.02**	
N2	6.27**	**.123**	1.08**	–1.04*	.04	**.90**	**–.92**	
P1	*11.29***	***.121***	1.48**	–.24	1.24**	**1.19**		**.99**
P2	5.98**	**.117**	.96**	–.12	.84*	**.90**		**.90**
Vocabulary Grades								
N1	2.55		.52	–.12	.4			
N2	3.83*	**.071**	1.00*	–.28	.72	**.76**		
P1	8.39**	**.165**	1.08**	–.56	.52	**1.13**		
P2	*8.40***	***.090***	1.20**	–1.04**	.16	**1.16**	**–.99**	

Note: Normal font—ANOVA; italics—Welch ANOVA; * p<0.05; ** p<0.01; bold—large or medium effect size

In the first round of assessment, Year 4 students scored on average higher than Year 1 students. In addition, the native raters gave higher marks to Year 4 texts than to the native texts, but they did not differentiate between the native and Year 1 compositions. The Polish raters, on the other hand, assessed the native essays as comparable to the Year 4 compositions, but they assessed the native essays as better than Year 1 production. All the differences observed in this round of assessment were large. In the second round, the grades awarded by three raters (N2, P1 and P2) assessed Year 4 learners higher than Year 1 students. The effect sizes were large and medium. Only one judge (P2) also gave on average higher marks to Year 4 than to native texts with a large effect size.

Discussion

The analysis of the scores has revealed that in spite of the lack of rating scales the judges were fairly consistent in the marks they awarded to the essays. Although their scores lacked agreement in the exact points assigned to individual texts, they were fairly consistent in depicting the variation in the global quality of the students' essays, and even more so in the differences in vocabulary use. It should be noted that the strength of the relationships between the marks awarded by individual raters is below the level commonly required in language testing. This fact, however, is not problematic in this study, as the aim of the assessment was not to arrive at a single reliable score for each essay, but to juxtapose the raters' marks with the information provided by the lexical indices and to analyse the two types of evidence against the raters' opinions on the factors influencing the quality of a text.

In spite of only moderate and weak correlations between the individual raters' marks, all four of them awarded on average higher holistic marks to Year 4 than to Year 1 students. The same tendency was observed in the case of the assessment of vocabulary, with an exception of one rater (N1), whose marks showed the smallest variation on the whole. However, the judges were not so unanimous about the quality of the essays written by the American students. The native raters perceived their holistic standard as comparable to the texts produced by less advanced Polish learners and lower than the compositions by more advanced learners. The Polish raters' scores demonstrated the opposite pattern. With one exception, the three judges did not note the differences in the lexical quality of the essays written by native and non-native students.

The results of the analysis point to the existence of robust and easily perceived differences between the general quality and the lexical excellence of essays written by learners of English at two different proficiency levels. Yet the essays written by the native students posed more challenges to the raters in the two rounds of assessment. The judges were far from unanimous in evaluating holistic and lexical quality of the American compositions. They certainly did not perceive them as best achievers in both general writing skills as well as in their vocabulary use.

6.5.5 Analysis 6—Relationships Between the Raters' Grades and the Indices

Preliminaries

The aim of this analysis was to examine if there are any associations between the raters' grades and the lexical aspects of student writing. It was expected that various lexical features would be associated with the holistic scores and that their relationship with the vocabulary grades would be higher and pertain to more indices.

Data Analysis

Since all the grades in the first and second round of assessment were not normally distributed (see Analysis 4), Spearman rank-order correlation coefficients were computed to explore the relationships between the raters' holistic and vocabulary scores and each of the 45 lexical indices generated for the essays. The results of the analysis are presented in Table 6.19.

The holistic marks demonstrate a range of statistically significant relationships with the raters' scores. With two exceptions (LexTy and POL for N1), they are weak. The native scores are associated with larger numbers of indices—13 in the case of N1 and 18 in the case of N2. Polish raters produce much fewer correlations with the metrics—4 in the case of P1 and 3 in the case of P1. The index which is most widely correlated with the holistic scores is essay length (Lgth). It demonstrates positive and statistically significant correlations with the grades awarded by three raters. Nine variables are weakly associated with the scores awarded by two

Table 6.19 Correlations Between the Lexical Indices and the Raters' Grades in the Two Rounds of Assessment

	Global (N=75)				*Vocabulary (N=75)*			
	N1	*N2*	*P1*	*P2*	*N1*	*N2*	*P1*	*P2*
Lgth	.390**	.282*		.265*		.575**	.270*	.381**
LexTy	.450**	.392**				.581**	.290*	.599**
GUIR_Ty	.382**	.389**				.346**	.247*	.550**
GUIR_Lex	.372**	.390**				.280*		.524**
MTLD	.281*							
2K_To	.242*	.254*						.473**
2K_Lex	.354**	.372**						.535**
AWL_To		.227*	.264*					.275*
AWL_Ty	.253*					.266*	.276*	.305**
FQLog_AW								–.248*
FQLog_CW	–.230*	–.284*						–.506**
FAM		–.263*	–.362**			–.246*	–.328**	–.552**
CNC			.236*					
IMG			.268*					
POL	–.414**							
BiAbs_To								.280*
BiAbs_Ty								.285*
BeTh_To								.293*
BeTh_Ty								.303**
HiMI_To						.239*		
NonMi_Ty		–.259*						
NonT_To				–.249*				
LowT_To	.282*	.297**						.339**
MidT_To		.300**						.237*
NonT_Ty				–.238*				
LowT_Ty	.304**	.281*						.330**
MidT_Ty		.280*						.243*
LogBiFQ_To		–.386**						–.322**
LogBiFQ_Ty	–.254*	–.364**						–.411**
BiMI_Ty		.301**						
BiT_Ty		–.234*				–.359**		
Statistically insignificant correlations	D, LD, MEA, HYP, NonMi_To, LowMI_To, MidMI_To, LowMI_Ty, MidMI_Ty, HiMI_Ty, HiT_To, HiT_Ty, BiMI_To, BiT_To							

Note: * p<0.05; ** p<0.01

native raters and the direction of these correlations confirms the expectations. AWL_To and FAM show weak relationships with the scores awarded by N2 and P1. The remaining indices are linked to the scores of individual raters, both native and Polish. Only two variables produce associations which are contrary to the expectations. P1's scores are positively correlated with the use of concrete and imaginable words, which seems rather surprising considering the topic and the genre of the essays.

A very different picture can be observed when analysing the relationships of lexical indices with the vocabulary scores. Although a similar number of the indices are statistically significantly linked to the vocabulary marks—36 coefficients as opposed to 38 in the previous matrix—an overwhelming majority of the correlations (21) are produced by one rater—P2. N2's and P1's scores are linked to nine and six indices respectively and N1 has produced no statically significant correlations. The strength of associations are generally higher, with 10 moderate associations (including 8 produced by P2). As many as five indices are linked with the scores awarded by three raters (N2, P1 and P2). GUIR_Lex demonstrates correlations only with the marks produced by two raters (N2 and P2). The remaining indices are correlated with the scores awarded by individual raters. Again, it is notable that the direction of the associations matches the expectations, with the exception of BiAbs, which produces positive correlations with P2's scores. This indicates that the use of novel expressions is associated with higher grades.

Discussion

On the whole, it can be noted that the measures of lexical productivity, diversity and sophistication are much more notably linked to the holistic scores (particularly those awarded by the native speakers) than measures of elaborateness, especially the gauges tapping the proportions of novel as well as infrequent and strongly associated collocations. Similarly to the global evaluation, in the assessment of vocabulary the measures of lexical productivity, diversity and sophistication (the last measured in perceptual rather than objective terms) as well as genre-specific sophistication are more strongly related to the raters' scores than measures of lexical elaborateness, especially the gauges tapping the proportions of infrequent and strongly associated collocations. Thus, in spite of the differences occurring between the patterns of correlations in the two analyses, the results point to the same gauges which are more strongly linked to the scores awarded in the two rounds of assessment. The strength of association of these measures is generally higher with the vocabulary marks than the holistic marks.

6.5.6 Analysis 7—Prediction of the Raters' Grades Based on Selected Indices

Preliminaries

A linear regression was performed on the raters' grades in order to explore to what extent the essays' lexical features, as captured by lexical

indices, can jointly predict human scores (global and vocabulary).[10] The grades awarded by the raters in this study were assumed to reflect the (lexical) quality of the essays, thus this analysis attempted to establish if automatic indices—and text lexical characteristics which they tap—together contribute to the human perception of writing quality. Since the marks awarded by the raters demonstrated at best only moderate correlations, a regression model was constructed for each rater separately.

Data Analysis

The standardised values of the same eight measures which were used in Analysis 3, were fed into the models. They represent lexical productivity (LexTy), lexical diversity (GUIR_Ty), lexical sophistication (2K_Ty, AWL_Ty) and lexical elaborateness (HYP, BiAbs_Ty, NonMI_Ty and HiT_Ty). The results of the regression—i.e. the coefficients of determination (R^2 and adjusted R^2) for each rater in both rounds of assessment—are presented in Table 6.20.

Only in the case of two raters—N1 and N2—R^2 of the model predicting raters' global scores turned out to be statistically significantly higher than 0. The adjusted R^2 was equal to 0.153 and 0.226 respectively. In the analysis performed for the vocabulary scores, the model predicting raters' global scores turned out to fit the data statistically significantly better than the null model ($p<0.001$) for two raters—N2 and P2 with the adjusted R^2 equal to 0.334 and 0.367.

The regression model has demonstrated that vocabulary of a text has some influence on its global grade only in the assessment performed by native raters. Lexical features, at least those tapped by the selected indices, had no effect on the holistic marks awarded by the Polish raters. Quite surprisingly, lexical qualities, as captured by the indices selected for the regression models, had an influence on the vocabulary marks only in the case of two raters—N2 and P2. The regression model constructed for P1 approached a statistical signification, but did not quite reach it. No statistically significant model could be fitted for the vocabulary marks awarded by N1. This last result may be due to the fact that N1 demonstrated little variance in his vocabulary scores. Interestingly, only

Table 6.20 Predictive Power of the Regression Models (With Raters' Grades as a Dependent Variable and With Lexical Indices as Predictors)

	Global				*Vocabulary*			
	N1	*N2*	*P1*	*P2*	*N1*	*N2*	*P1*	*P2*
R^2	.245	.310	.185	.168	.141	.406	.200	.435
adj. R^2	.153	.226	.086	.067	.037	.334	.103	.367
p	.013*	.001***	.080	.124	.234	<.001***	.052	<.001***

Note: * p<0.05; ** p<0.01; *** p<0.001

N2 demonstrated an influence of lexical qualities on his scores in both rounds of assessment.

Despite the fact that the authorship of individual essays was not known to the raters during the assessment process, another linear regression analysis was performed to explore if the raters applied—explicitly or implicitly—the same criteria in evaluation of learner and native texts. Two model specifications were considered. In the first specification, the author's status (learner vs. native) was added in addition to the indices used in the previous analysis. In the second one, the author's status as well as the interaction of the status with all the analysed indices were added to the model. Next, the new models were compared with the original one using the F-test in order to establish whether they fit the data statistically significantly better. The results of the analysis are presented in Table 6.21. The raw titled *difference* provides the change in the coefficient of determination (adjusted R^2) between the new models and original one.

The new models demonstrated no statistically significant improvements of prediction in the case of vocabulary assessment (p against the original model >0.05). In the case of global assessment, only the models for the Polish raters appeared to fit the data better than the original ones: the model including only the status for rater P1 (ΔR^2=0.119, p<0.01) and the model including both the status and its interactions for rater P2 (ΔR^2=0.264, p<0.01). That is why only the new models fitted for the Polish raters' holistic scores were examined further. This involved scrutinising general effects of the individual indices (P1 and P2), their effects in the learner and the native categories separately as well as the difference between the effects in the two writer categories (P2). The results of this analysis are presented in Table 6.22. For convenience in interpretation, the table pulls together the results of the best-fitting model of each rater.

In the case of holistic assessment, the simpler model chosen to describe the evaluation made by N2, and the more complex ones selected for raters P1 and P2 have similar values of the adjusted R^2 statistic ranging from 0.207 to 0.263. The simpler model fitted for N1 has a slightly lower predictive power with adjusted R^2 of 0.153. The raters' holistic grades seem to have been most commonly affected by lexical productivity. LexTy produced a general positive effect on the marks awarded to all the texts by N1 and N2, but it played an interesting role in the evaluation performed by P2. It boosted learners' grades, but it acted to the detriment of native writers' grades and its effect was almost twice as strong as in the case of the Polish students. Only P1's marks have not been influenced by the number of different content words used by the writer. Lexical sophistication represented by 2K_Lex affected the holistic marks awarded by the two Polish raters. In the case of P1 it produced a positive effect on all the marks. However, this quality produced a strong

Table 6.21 Predictive Power of the Regression Analysis (With Raters' Grades as a Dependent Variable and With Lexical Indices, Author Status and the Interactions as Predictors)

	Global				*Vocabulary*			
	N1	*N2*	*P1*	*P2*	*N1*	*N2*	*P1*	*P2*
models with status but without interactions								
R^2	.254	.310	.304	.257	.142	.432	.207	.436
adj. R^2	.150	.214	.207	.154	.023	.353	.098	.358
difference	[–.003]	[–.012]	[.121]	[.087]	[–.014]	[.019]	[–.005]	[–.009]
p—against the null model	.018*	.003**	.003**	.016*	.314	<.001***	.069	<.001***
p—against the original model	.374	.914	.001**	.007**	.778	.090	.448	.833
models with status and interactions								
R^2	.388	.450	.382	.432	.308	.486	.286	.483
adj. R_2	.205	.286	.198	.263	.094	.332	.073	.329
difference	[.052]	[.060]	[.112]	[.196]	[.057]	[–.002]	[–.030]	[–.038]
p—against the null model	.018*	.002**	.021**	.004**	.147	.001**	.200	.001**
p—against the original model	.177	.133	.052	.06**	.183	.464	.652	.806

Note: * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Table 6.22 Summary of Results for the Best-Fitting Models for Each Rater in the Two Rounds of Assessment

independent variables		*Global*						*Vocabulary*			
		N1	*N2*	*P1*	*P2*			*N1*	*N2*	*P1*	*P2*
		indices	*indices*	*indices+status*	*indices+status+interactions*			*indices*	*indices*	*indices*	*indices*
R^2		.245	.310	.304	.432			.141	.406	.200	.435
adj. R^2		.153	.226	.207	.263			.037	.334	.103	.367
p		.013*	.001***	.003**	.004**			.234	<.001***	.052	<001***
effects		*general*	*general*	*general*	*learner*	*native*	*diff.*	*general*	*general*	*general*	*general*
LexTy	B	.463	.424		.536	−.979	−1.516		.972		.542
	p	.025*	.032*		.047*	.008**	.002**		<.001***		.008**
GUIR_Lex	B					1.585	1.707		−.642		
	p					.007**	.006**		.014*		
2K_Lex	B			.575		−.978	−1.319				.543
	p			.001**		.023*	.007**				.003**
AWL_Ty	B										
	p										
HYP	B										
	p										
BiAbs_Ty	B		−.367								
	p		.026*								
NonMI_Ty	B		−.295								
	p		.027*								
HiT-Ty	B			.311							
	p			.040.*							
Status	B			1.403							
	p			.001**							

Note: * p<0.05; ** p<0.01; *** p<0.001

negative effect on the marks awarded by P2 to the American essays. In other words, the essays written by native writers tended to receive lower grades if they contained more sophisticated words. On the other hand, P2 gave credit to the American students (but not the Polish learners) for using varied vocabulary and it was the strongest effect observed in the analysis ($B=1.585$, $p<0.01$). No other rater seems to have been affected by lexical diversity. Phraseology had their effects on the marks of two raters. A negative general effect of BiAbs_Ty and NonMI_Ty could be observed for N2's scores. Academic vocabulary and word hypernymy had no effect on the marks awarded by any of the raters. Finally P1's grades were positively affected by frequent and stereotypical collocations (Hi_Ty). P1 tended to evaluate higher the overall quality of texts written by the native speakers, irrespective of their lexical characteristics.

As already stated, in the assessment focusing on evaluation of vocabulary, lexical qualities had an influence on the marks awarded by two raters—N2 and P2. This influence was higher than in the holistic assessment with the adjusted R^2 statistic equal to 0.334 and 0.367 for N2 and P2 respectively. Lexical productivity captured by LexTy had a positive effect also on the vocabulary marks and its effect was stronger than in the case of holistic scores, particularly for P2. In addition, the vocabulary scores awarded by P2 were affected by the occurrence of sophisticated lexemes (2K_Lex), and their effect was opposite than in the case of American essays in the holistic assessment. A rather counter-intuitive negative effect on the marks was produced by lexical diversity (GUIR_Lex) in the case of N2. His scores were lower for lexically more varied texts. Academic vocabulary, word hypernymy and phraseology had no effects on the vocabulary marks.

Discussion

The results clearly point to the fact that lexical qualities of a text were taken into account when texts were evaluated holistically. The values for the adjusted R^2 seem quite high considering that lexis is typically one of several criteria taken into consideration in global assessment. Interestingly, different raters seem to have paid attention to different lexical features in texts. The most consistent effect can be observed for lexical productivity. Lexical diversity, sophistication and formulaicity are relevant only for individual raters. The analysis also revealed that one rater (P2) applied—explicitly or implicitly—double standards in the holistic assessment. She tended to give credit to the American writers for lexical sophistication and to the Polish subjects for lexical productivity. In addition, some lexical qualities had an opposite effect on her scores awarded to the Polish and American students.

The most surprising result of this analysis is the fact that while in the holistic evaluation all the raters tended to be influenced by essays' lexical qualities in one way or another, in the assessment of vocabulary not all

of them seem to have paid a lot of attention to the qualities captured by the automatic indices. However, it should be noted that in the case of the raters who seem to be sensitive to these qualities, the predictive power of the models was quite strong. These two raters (N2 and P2) looked out in particular for lexical productivity as well as lexical sophistication and diversity in the essays. The effect of this last variable on the scores is rather ambiguous as it works to the detriment of the marks awarded by N2. In the assessment of vocabulary none of the raters seems to have used double standards in interpreting the lexical qualities of a text.

6.5.7 Analysis 8 and 9—Comparison of Vocabulary Scores and Their Relationships With the Indices and the Raters' Grades

Preliminaries

The administration of the Vocabulary Levels Test made it possible to assess the L2 learners' lexical competence with an independent instrument. An analysis of the students' test scores will help to establish if lexical competence grows in advanced learners with years of exposure, as demonstrated for lexical proficiency in Analysis 2 and 3. An examination of the relationships of the test scores with the indices (tapping lexical qualities of students' texts) and raters' vocabulary grades (expressing judges' perceptions of students' effective vocabulary use in their essays) will indicate the extent to which these three instruments and procedures measure the same underlying trait. These analyses will confirm the existence of links between lexical proficiency and lexical competence and validate the automatic indices and human grades as methods measuring both these constructs.

Data Analysis

Scores in each section of the two vocabulary tests and test totals were examined. Despite the fact that the test results were normally distributed in both groups only for the 10K section and the Total in the productive version (Shapiro-Wilk test $p>0.05$), the t-test for independent samples was used for all the comparisons. This test is better suited for comparing group means than its nonparametric equivalent in a situation when the assumption of equal (however non-normal) distributions in groups being compared has not been met. Levene's test was applied to estimate the equality of the variances between the groups ($p>0.05$). The outcomes of the statistical procedures are reported in Table 6.23. The table also reports the effect size expressed by Cohen's d.

The statistical procedures applied for the analyses of the data demonstrate that Year 4 students scored on average significantly better than Year 1 students on all but one section of the receptive test (R2K) and on all the sections of the productive test. The observed differences are large, except for the lowest band of the productive vocabulary (P2K).

Table 6.23 Comparisons Between the Scores of the Receptive and Productive VLT for Year 1 and Year 4

Test part	*R2K*	*R3K*	*R5K*	*RAWL*	*R10K*	*RTot*	*P2K*	*P3K*	*P5K*	*PAWL*	*P10K*	*PTot*
Equal variances	y	n	n	n	y	n	y	y	y	y	y	y
t	−.90	−6.69**	−6.06**	−4.01**	−5.57**	−8.10**	−3.68**	−4.65**	−7.98**	−5.67**	−4.95**	−7.87**
d		1.35	1.22	.82	1.17	1.67	.77	.99	1.66	1.20	1.03	1.65

Note: $N_{Year\ 1}$=50; $N_{Year\ 4}$=42; ** p>0.01

Table 6.24 Statistically Significant Correlations Between the Lexical Indices and the Productive VLT Scores

	Lgth	LexTy	*GUIR_Ty*	AWL_To	AWL_Ty	*MidT_To*	*MidT_Ty*
PTot	.398**	.330**	*.256**	*.278**	.302**	*.266**	*.271**

Note: Italics—Pearson moment correlation; normal font—Spearmen rank-order correlation; N=75; * p<0.05; ** p<0.01

Table 6.25 Correlations Between the Vocabulary Grades and the Productive VLT Scores

	N1	*N2*	*P1*	*P2*
PTot	.243	.337*	.590**	.397**

Note: N=50; * p<0.05; ** p<0.01

The relationships between the results of the productive vocabulary tests and the lexical indices were examined by means of Pearson product-moment correlation and Spearman's rank-order correlation. Table 6.24 lists only statistically significant correlation coefficients.

Seven (out of 45) lexical indices demonstrate statistically significant weak and positive associations with the total results of the vocabulary test. The highest coefficient was noted for text length (Lgth).

The vocabulary tests were available only for 25 Year 1 essays which were assessed in the first round of assessment, so the sample was too small for a meaningful analysis. That is why the correlations between the test results and raters' grades were performed only for the vocabulary marks. Since they did not show normal distribution for any of the raters (Shapiro-Wilk p<0.05), Spearman's rank-order correlations were computed and their coefficients are listed in Table 6.25.

The vocabulary marks awarded by three out of four raters demonstrate statistically significant positive associations with the cumulative results of the vocabulary productive test. The associations were weak for N2 and P2 and moderate for P1.

Discussion

The results of the vocabulary tests demonstrate that Year 4 students scored consistently better on all but one section of the receptive and productive vocabulary tests, with large effect sizes, particularly for higher frequency bands. The lack of the difference in the lowest band of the receptive test can easily be explained by the ceiling effect. Thus, the results testify to the fact that both the receptive and productive lexicons of even fairly advanced students continue to grow as a result of extensive exposure.

Another interesting observation following this analysis is that the results of the productive Vocabulary Levels Tests, which tap the size component of the learners' lexical competence, demonstrate associations with the indices describing the vocabulary used by the learners in their essays. Links exist between the test scores and the lexical indices representing productivity, diversity, genre-specific sophistication and elaborateness. This confirms the concurrent validity of the two types of assessment in evaluation of lexical competence as well as interdependence of the dimensions of vocabulary knowledge. As expected, the highest effect was observed for the correlations between the VLT results and the two measures of lexical productivity, as they both relate to the size component of the lexicon. The effects of the other indices were much less pronounced, and for most of the measures did not reach a statistical significance. It is particularly notable that the results of the productive vocabulary test did not correlate with the 2K indices in spite of the fact that the two instruments follow the same approach to measuring sophisticated vocabulary and apply the same frequency criteria. This outcome stays in tune with the results of Laufer (1998) who also noted a lack of correlations between the information produced by these two tools.

Comparable and even stronger effects could be noticed when examining the associations between the results of the productive vocabulary test and the grades awarded by the raters. This observation confirms the concurrent validity of the two assessment types, and together with Analysis 6 confirms the associations between raters' grades assigned to the lexical aspects of a text with its author's vocabulary size. As in the case of the previous analyses, it is clear that the vocabulary scores awarded by N1 do not behave similarly to the other raters' grades.

6.5.8 Analysis 10—Interviews

Preliminaries

No rating criteria were presented to the judges at the outset of the assessment process and they had not been trained specifically for this project. To the contrary, they were encouraged to apply their own standards in the evaluation of the essays. That is why the interviews carried out with the raters after the completion of the two rounds assessment are a very important part of the study. Their aim was to tease out the judges' expectations concerning the use of vocabulary in student compositions and their opinions on the factors contributing to lexical quality of a text.

Data Analysis

The four reports were reread carefully and the raters' comments and observations were coded based on the information that they contained.

Next, the coded fragments were matched with relevant questions. A summary of the raters' responses to the two sets of questions, backed up with quotations taken from the reports, are presented next.

GLOBAL ASSESSMENT

1. What criteria did you take into consideration when assessing the essays?

All the raters reported paying attention to several evaluation criteria when marking the essays. These criteria were almost identical and included: content, organisation of ideas, coherence, register, paragraphing and the quality of language (both accuracy and range). Three raters did not mention mechanics such as spelling or punctuation at all and one rater (P2) specifically stated they were not relevant.

2. What was the weight of each criterion in the overall mark?

Two raters (N1 and P2) reported that they treated the essays mainly as a task in academic writing rather than an exercise in language use and thus they gave more importance to the communicative (discoursal) aspects of the essays rather than to their purely linguistic features such as vocabulary, phraseology or grammar. Two other raters (N2 and P1) paid equal attention to the discoursal and linguistic features of an essay. In fact, N2 reported that the linguistic layer of an essay (sentence structure, collocation and vocabulary) could overbalance the lack of structure in a composition.

The raters had different expectations concerning the content of the essays and interesting and unexpected patterns emerged in this respect. The same raters who reported giving more focus to the communicative value of the students' compositions (N1 and P2) did not expect originality of the content, although they admitted that original ideas worked to the benefit of an essay and could positively influence the grade. The other two raters (N2 and P1), who treated linguistic features of writing on equal footing with its discursive side, looked for a fresh approach in the treatment of the topic. They also pointed out that the essays should tackle several aspects of the topic and use information coming from different sources rather than present one-sided opinion.

3. Did you assign points for each criterion to arrive at the final mark?

Only one rater (P1) evaluated the essays on each criterion separately (using the marks from 2 to 5) and his final marks were computed as averages of all the component marks. However, he also admitted that he had allowed an element of general impression to contribute to the mark.

The other three raters relied on their subjective perceptions of the overall quality of an essay and its effect and their marks were not weighted means of component scores. As N1 observed, "Weighing is artificial, because there is no simple recipe for what makes a good essay".

4. What role did vocabulary play in your global assessment? Did you see any interesting patterns?

The raters had divergent opinions on the role of vocabulary in the global assessment which were, however, consistent with their views on the importance of the purely linguistic criteria in the final mark. Raters N1 and P2 reported that they had paid little attention to the lexical aspects of an essay, whereas judges N2 and P1 claimed that good use of vocabulary was an important, albeit one of several, aspect of good writing. All raters voiced interesting comments about the characteristics of good vocabulary use in writing; however, since they were later repeated in the responses concerning the second round of assessment, their opinions on this issue will be summarised further.

5. How, on the whole, did you evaluate the essays?

The raters had mixed opinions about the general quality of the essays. N2 and P2 found it difficult to offer an assessment of the essays as a whole. P2 also admitted that since she had known that the essays were marked solely for research purposes, she had been less generous with good grades because she had not needed to pay attention to negative consequences of bad marks for the students. In other words, the marks she had awarded to individual essays reflected more closely her true and objective opinion of the quality of writing and were not affected by external factors, such as encouragement or appreciation of an effort. The opinions expressed by N1 and P1 differed considerably. N1 observed that generally the essays were not as good as he had expected of both Polish and American students. His comment referred in particular to the way the essays were structured, as in many cases the texts were not clearly divided into paragraphs. On the other hand P2 had a rather positive impression of the general quality of the essays.

6. Could you tell the three groups of essays apart? Did you see any differences between the levels?

The Polish raters reported that in spite of the fact that the essays had been coded and mixed they could easily distinguish between the essays written by American and Polish students. In spite of the same topic the content of their texts was different. The American students "took a more concrete approach" (P2), and tackled "more down-to-earth" (P1) issues.

In consequence, their style was "informal and chatty" (P1). Interestingly, the native judges reported having problems with distinguishing the essays written by American and Polish students, yet at the same time they offered the same insights about the American essays. N2 observed that

> native students more often gave their opinion and backed it up with personal stories, which suggested that their only sources of information were anecdotes about their own or their family and friends' lives, instead of information reflecting reading of newspapers or magazines, let alone books.

Both native judges remarked that American students "tended to use more colloquial language" (N1) which "sometimes was too idiomatic and too conversational" (N2). Both Polish and native raters observed that Polish essays "tended to tackle more general issues" (P1) and "present more complex treatment of the topic" (N2). N2 also observed that Polish students were more careful with expressing their own opinion on the topic.

The difference between the first- and fourth-year essays was "not so eye-catching", as observed by P1. Both Polish raters remarked that some Polish essays used simpler arguments and used simpler language, and these could be guessed to have been written by Year 1 students. The native raters declared that that could easily discriminate between the Year 1 and Year 4 students, but did not elaborate on the factors differentiating the two groups of texts. N1 remarked, however, that "the writing skills of some non-native students are phenomenal".

7. Do you have any additional comments?

The raters did not offer additional comments concerning the first round of assessment.

ASSESSMENT OF VOCABULARY

General remarks on this round of assessment

Although unprompted, the judges offered their general impressions on this round of assessment. All the four judges admitted that the assessment focusing solely on lexis was a difficult task. They found it challenging to separate vocabulary use from other aspects of writing: content (N1), discursive and communicative aspects and rhetoric (P2) and grammar (P1 and P2). P2 remarked that it would have been easier to mark the essays focusing on grammar. On the other hand, P1 stated that he considered the division into syntax and lexis artificial and instead he was in favour of a more unified lexical approach which spotlighted the grammar of a word.

Three of the four raters also underlined a relationship between vocabulary use and other aspects of writing, which is expressed by the following quotations:

- If the vocabulary of an essay is terrible, everything else (i.e. structure or support) tends to be terrible, although this is not always the case. (N1)
- Usually when vocabulary was simple, there were problems with both global organisation of an essay as well as grammatical structures. If a sentence is incorrect in terms of grammar, it is usually because students don't know what words mean. (N2)
- Essays which demonstrated good use of sophisticated vocabulary were good essays in the general sense as well. However, I gave a few very good marks which would not have been very good marks in the global assessment. (P2)

N2 in particular stressed a connection between vocabulary and content:

- The students who used simpler vocabulary usually just expressed their opinion without reflection and looking at the problem from various angles. Good vocabulary went with the ability to express subtle thought. (N2)

1. What criteria did you take into consideration when assessing the vocabulary?

The raters listed several criteria which played parts in the evaluation of lexis, both in the global and vocabulary assessment. The factors mentioned by all the raters included: collocations/idiomaticity and lexical sophistication; however, these two criteria were not related to the grades in a straightforward way, but were hedged with several conditions, discussed in the next paragraphs. Three raters (N2, P1 and P2) also mentioned appropriate register.

Wide and varied vocabulary was mentioned by two raters (N2 and P1) but was not elaborated on in the interviews:

- Non-native students were more aware of the existence of different words. On the whole there is a wider vocabulary in non-native than native essays. (N2)
- The Polish essays tended to use more varied vocabulary. (P1)

Other factors mentioned were accurate word choice (emphasised by N2, but also mentioned by P2) and accurate word grammar (particularly relevant to P1). Correct spelling of words was only mentioned as a criterion

by one rater (P1). For N2, spelling mistakes were not relevant if words were otherwise used correctly.

- Spelling did not play an important role unless students tried to use the words they obviously didn't know. (N2)
- Another criterion P1 took into consideration was spelling.

Two raters (N1 and P2) also mentioned four other criteria, which are not common in rating scales: natural use, fluidity, spontaneity and impression.

- Natural use [of vocabulary] and fluidity (how an essay flows) seems to be most important. (N1)
- The criteria P2 took into consideration [in the assessment of vocabulary] were . . . spontaneity of vocabulary use and impression (P2 looked for vocabulary which would make an impression on her).

Unfortunately, the raters did not elaborate further on these criteria, but it can be inferred from the context that the first rater (N1) referred to idiomaticity, whereas the other judge (P2) implied novelty of expression (as opposed to clichéd language).

2. What was the weight of each criterion in the overall mark for vocabulary?

Three factors seemed to be most important for the raters: collocation/idiomaticity, sophisticated vocabulary and register. This is best summarised by the following quotations:

- An effective use of vocabulary is a combination of sophisticated words and phrases showing someone really knows the language. (N1)
- The best essays were the ones that demonstrated natural and fluent use of advanced vocabulary, appropriate to the register. (P2)

For all the raters the factor which was mentioned first and given prominence in the assessment of lexis was collocation and idiomaticity. This is exemplified by the following quotations:

- N1 looks more for a phrase that shows that someone really knows the language [rather than sophisticated vocabulary].
- In assessing vocabulary, N2 was less interested in individual words and he focused on collocation.
- While evaluating vocabulary P1 focused on collocations (word combinations) and the grammar of a word.

- P1 evaluated highly the use of phrasal verbs.
- Generally, P2 considers accuracy less important. She pays attention to idiomaticity of language (incorrect idiom or collocation) but not to basic grammatical mistakes (such as *he go*).

However, two raters (N2 and P2) expressed awareness that the use of certain set phrases could work to a detriment of an essay:

- N2 noticed that some of the students were trying too hard and used 'borrowed language' (that is, slogan-like language of the media) and their compositions sounded like advertisements. He was also sensitive to 'borrowed ideas' such as references to terrorist threats which resulted in 'borrowed language'. Essays which used 'borrowed language' tended to exploit vocabulary which was more catchy and basic. He marked down such essays.
- The good students did not use clichés and mass-media slogans. (N2)
- P2 was sensitive to big words which were used in a clichéd . . . way.

The four raters elaborated most on their expectations and perceptions related to the use of sophisticated vocabulary. Their opinions were somewhat self-contradictory in this respect. Three raters (N1 and P1 and P2) stated that they did not expect sophisticated words in an essay:

- N1 is against using sophisticated vocabulary for the sake of using sophisticated vocabulary. In journalism the simplest word is the best word (e.g. the word *body* instead of *organism*).
- Sophisticated vocabulary was not important, a high mark could be assigned for an essay using vocabulary which was simple but worked very well. (N1)
- As far as lexical sophistication is concerned, P1 expected vocabulary which would match the register (formal, fairly advanced) but not necessarily very sophisticated.
- P2 does not expect sophisticated words in academic essays.

P2 also expressed the same attitude in relation to technical vocabulary related to mobile phones:

- A large majority of the Polish students had problems with mobile phone terminology. However, P2 did not consider it a serious flaw in the compositions because they were still communicative.
- When marking vocabulary the knowledge of mobile phone terminology was not particularly important. (P2)

Yet, the same three raters admitted that more advanced vocabulary had a positive effect on the mark:

- But sophisticated words can push the grade. (N1)
- P1 tended to mark an essay down if the vocabulary/style was too colloquial.
- Another criterion was the level of formality. The level of formality does not necessarily correspond to a word's frequency band; however, they tend to go very much together. P1 looked for words beyond the first 2000-most frequent words in English. For example, he expected a student to use verbs other than *give* or *take*.
- If P2 had to name the most important criterion [in the evaluation of vocabulary], it would be natural use of sophisticated vocabulary.

N2 also recognised the importance of sophisticated words:

- N2 also marked an essay down when vocabulary was too simple. For example, he did not like when students kept using words *good* or *bad* instead of finding more sophisticated synonyms.
- If an essay was marked as very good it meant that the words were more sophisticated, there was a higher density of sophisticated words. (N2)
- N2 noticed two extremes. On the one hand, students use vocabulary that is too simple. For example, they talk about good and bad things instead of positive and negative consequences.

Although all the judges admitted that lexical sophistication was an important aspect of evaluation, they all expressed their reservations and listed conditions on the use of advanced words so as they could have a positive effect. Sophisticated words boosted the mark when they were used:

- accurately
    - N2 marked down . . . the essays which used 'big words' inaccurately. Although he claims that the latter is seldom the case with in-class writing when students do not have an opportunity to use a thesaurus.
    - N2 noticed two extremes. . . . On the other hand, [students] use sophisticated words inaccurately without knowing what they actually mean (as if making a stab in the dark).
- consistently
    - For example, in one of the essays the student used a combination of two very sophisticated words *taxing* and *onerous* but the rest of the essay was uninteresting in terms of vocabulary. So the important thing was that words should match. (N2)
    - P1 did not appreciate sophisticated words which tended to stand out and did not match the rest of the essay. However, he admits

he tends to be more lenient about penalising unnatural use of sophisticated words than about vocabulary/style which were too informal (e.g. *a pain in the neck*).
    - In some Polish essays very sophisticated words were forced and did match the general lexical level of the text, thus producing a clash. (P1)
    - If sophisticated words are used they should be used consistently. (P2)

- matching the content
    - Sophisticated words can push the grade (provided they fit the content). (N1)
    - A good mark did not mean that a student was capable of using a few sophisticated words such as *taxing* or *onerous*. The sophisticated vocabulary had to indicate more depth into an opinion. (N2)
    - If sophisticated words are used they should . . . match the logic and organisation the essay. (P2)

- matching the register
    - N2 did not like when students kept using words *good* or *bad* instead of finding more sophisticated synonyms. But at the same time he realised that if the essay was written in the conversational style, the words *good* and *bad* fit the context best.
    - As far as lexical sophistication is concerned, P2 expected vocabulary which would match the register (formal, fairly advanced).
    - The best essays were the ones that demonstrated natural and fluent use of advanced vocabulary, appropriate to the register. (P2)

The two Polish judges mentioned one more criterion hedging the positive effect of sophisticated words in students' writing. Advanced words can boost the mark if they are used:

- naturally
    - P1 did not appreciate sophisticated words which tended to stand out and did not match the rest of the essay. However, he tends to be more lenient about penalising unnatural use of sophisticated words.
    - P2 was sensitive to big words which were used in a clichéd and unnatural way. In fact, such use of vocabulary worked to the detriment of the essay as it was difficult to read it with interest and it did not sound natural. It sounded forced and contrived.
    - If P2 had to name the most important criterion, it would be natural use of sophisticated vocabulary. So, the best essays were

the ones that demonstrated natural and fluent use of advanced vocabulary, appropriate to the register.

Naturalness is not a standard criterion occurring in evaluation rubrics and it is rather difficult to define. It can be stipulated from the context that the two raters referred to the consistent use of sophisticated words. P2 could have also alluded to the use of advanced vocabulary which matched the content and register. Interestingly, the criterion of natural use was mentioned by another rater (N1) in relation to all vocabulary (not just advanced words), but in his case the context suggested that he was referring to collocation accuracy.

3. Do you assign points for each criterion to arrive at the final mark?

Only two raters (P1 and P2) addressed this question directly and indicated that the mark was purely subjective rather than based on careful weighing of several criteria. However, it can be inferred from the answers to the previous question about the importance of individual components in the assessment of lexis that the other two raters followed the same procedure:

- Unlike in the previous round of assessment, the mark was not an average of all component marks. It was more impressionistic. (P1)
- Generally marking was based on impression rather than careful weighing of criteria. This mark was more impressionistic than the previous one. (P2)

4. How on the whole did you evaluate the essays?

The raters' general impression about the essays was that they represented the whole range of lexical quality:

- On the whole, N2 noticed two extremes. On the one hand, students use vocabulary that is too simple. For example, they talk about good and bad things instead of positive and negative consequences. On the other hand, they use sophisticated words inaccurately without knowing what they actually mean (as if making a stab in the dark).
- It is difficult to make a general statement about the lexical level of the essays. It varied. There were some very simple and even simplistic essays (tackling the issue from one side only). (P1)
- P2 found it difficult to comment on the general level of vocabulary knowledge. Vocabulary use ranged from barely communicative to quite impressive.

5. Could you tell the three groups of essays apart? Did you see any difference between the levels?

The raters observed that it was easier for them to distinguish between the groups of students when focusing solely on the assessment of vocabulary:

- This time it was easier to tell the groups apart, although again, there were many cases when the difference was not obvious. (N1)

Three raters (N2, P1 and P2) had very similar observations about the differences in vocabulary use between American and Polish students. The vocabulary used by American students tended to be more informal and colloquial. Polish vocabulary was more varied and advanced:

- Native speakers had a tendency to use more colloquial and idiomatic language, non-native students were more aware of the existence of different words. On the whole there is a wider vocabulary in non-native than native essays. (N2)
- There were some differences between native and Polish essays concerning vocabulary. Again, the vocabulary used by the American students tended to be informal. . . . The Polish essays tended to use more varied vocabulary. On the other hand, in some Polish essays very sophisticated words were forced and did match the general lexical level of the text, thus producing a clash. (P1)
- The American essays were worse in terms of register. They were communicative and to the point but very banal. They demonstrated poor language awareness. This awareness was much more visible in the Polish essays. Language awareness means the ability to create a certain effect. This does not mean that the Polish essays always created this effect (not if they used clichéd language). There was much more range in the Polish essays from very simple vocabulary use to quite impressive lexis. (P2)

However, the Polish raters remarked that American students had the advantage of using terminology connected directly to the technical aspects of mobile phones:

- [American students] used a fair amount of 'more advanced' words pertaining to the technical aspects of mobile phones (*static*, *towers*). (P1)
- What caught P2's attention was wrong use of technical vocabulary (mobile phone terminology) by non-native speakers (e.g. *we send SMSs*). A large majority of the Polish students had problems with mobile phone terminology.

N2 commented on collocational accuracy in student writing

- Even native speakers occasionally had problems with collocations. One student wrote *When cell phones first broke into Minneapolis*. (N2)

Although this sentence refers explicitly only to the American students, it implies that Polish learners, on the whole, had more problems with correct phraseology.

N2 suggested that the difference in the use of sophisticated vocabulary was not so much a result of L1 background, but the complexity of ideas expressed by the students:

- Students who used information from a variety of sources tended to use more sophisticated vocabulary. Sophisticated language was a proof of reflection. There was not much difference in this respect between native speakers and Polish students. (N2)

6. Do you have any additional comments?

Two raters offered general comments of the process and results of the assessment focusing of lexis:

- N1 noticed that in general when focusing solely on vocabulary he tended to give higher marks.
- N2 found marking the essays based solely on vocabulary use rather difficult. When he marked the whole batch he noticed that he gave a lot of 3+ and 4+, which showed that in many cases he was undecided. So, he reread the whole batch marking some essays down and some up.

N1 also observed that the use of interesting (sophisticated) vocabulary is determined by the topic:

- It is very difficult to come up with interesting vocabulary in an essay on the mobile phone. In an essay on literature students could easily use much more sophisticated vocabulary. (N1)

Discussion

The interviews highlighted several interesting ideas concerning the assessment of vocabulary use in writing.

The raters' comments reflect the dual role of writing in language teaching and assessment, already discussed in Chapter 3. The judges were aware that writing is a method used to assess the acquisition of language elements, such as vocabulary and grammar. However, they were also conscious of the significance of the writing skill in its own right, and of a need for assessing learners' ability to create communicative and effective discourse. The judges differed subtly in the priority given to each of these functions. For two raters the communicative and discursive aspects of texts overbalanced their linguistic qualities, whereas the two other judges

considered these two sides of writing equally important. This result confirms the model of performance assessment in the 'strong' and 'weak' sense proposed by McNamara (1996). It also implies that the assessment of a text depends at least to some extent to a place a rater takes on the 'strong/weak' continuum and his/or her attitudes related to the significance of the two aspects of writing in the final mark.

Vocabulary of a text is just one of several criteria in the global assessment of writing, as declared by the raters. However, its relative importance in the overall mark depends on a rater's conscious or unconscious position on the 'strong/weak' continuum of performance assessment. The raters who acknowledged the importance of the linguistic aspects in the global mark at the same time asserted the importance of lexical aspects in the holistic assessment, whereas the judges focusing on the communicative value of texts in global evaluation undermined the role of vocabulary in its final results. At the same time, all the raters agreed that vocabulary use is strongly related to and almost inseparable from other aspects of writing, such as content and organisation, on the one hand, and grammar on the other. Good vocabulary and its effective use are and should be a result of complexity of ideas expressed in a text. At the same time, grammatical mistakes may often be a result of incorrect word choice or unawareness of a particular item's word grammar, which was also suggested by two raters. Although the judges did not indicate this, the interdependence or even interpenetration of vocabulary and other aspects of writing may be characteristic for fairly advanced users of language, whose writing was assessed in this study. Such students are usually assigned more ambitious topics allowing a more complex discussion and treatment of the subject. In addition, more advanced students, especially those exposed to large doses of formal instruction rarely make consistently repeated purely syntactic errors (e.g. incorrect use of a tense) and their problems with accuracy are frequently local. This was suggested by one of the raters who claimed that students' grammatical mistakes frequently had lexical roots. This observation explains why most of the rating scales in Chapter 3 do not make a distinction between lexical and grammatical/syntactic errors and include only one criterion of language use covering both categories.

The association of lexis and other aspects of writing, postulated by the raters in the interviews, can be supported by the results of two earlier studies discussed in detail in Chapter 3 (Astika, 1993; Lee et al., 2009). Both papers reported very high (and the highest) correlations between analytic marks for vocabulary and other criteria, in particular content, organisation and language use. The same studies also demonstrated very high correlations between marks for vocabulary awarded in analytic assessment and holistic grades of the same texts. It is worth noting that the data used for assessment in these studies were also produced by fairly advanced learners of English as a foreign language.

Interestingly, the most important criterion in the assessment of vocabulary indicated unanimously by the four raters did not pertain to individual words and their qualities, but to their relations with neighbouring words. Collocation and formulaicity, which both relate to the phraseology of a text, contributed to the "fluidity" and "naturalness" of a text, according to one of the raters. Yet, the relation of phraseology and text (lexical) quality is rather subtle. On the one hand, the raters valued idiomatic language, which means that they appreciated students using phrases built of words frequently co-occurring in language. On the other hand, two raters expressed distaste for "borrowed language", "clichés" and "mass-media slogans", which are nothing else but combinations of words frequently occurring together. One rater went even further and declared her appreciation for "spontaneous" (novel) language use in student writing, which is the opposite of fixed or semi-fixed word combinations.

As far as the use of individual words is concerned, the raters mentioned lexical variety only twice, and only in passing, without any elaboration on the role of this feature in both global and lexical assessment. It is noteworthy that the perceptually more salient opposite of lexical variety, repetition of words (Jarvis, 2013), never surfaced in the interviews (P1 mentioned that American essays were repetitive, but he did refer specifically to vocabulary use). On the other hand, the four raters devoted a lot of attention to the discussion of lexical sophistication. They indicated that although the presence of advanced words was not a necessary condition for their favourable perceptions of a text, it could positively affect both the grade for vocabulary as well as the holistic score.

A particular attention should be given to the raters' comments concerning the use of sophisticated vocabulary. They demonstrate that this criterion does not work in a straightforward manner, 'the more the better'. All the four raters listed several prerequisites for an effective use of advanced words. One condition was an accurate word choice, that is the sophisticated words should fit the message and demonstrate that the writer is fully acquainted with the intricacies of their meanings. However, one rater admitted that such errors were not very frequent when students did not have access to a thesaurus during writing. Sophisticated words used in a text should also be appropriate for the register selected by the writer and the content of an essay. If the whole text is written in conversational style, sophisticated words may not fit. As suggested by the judges, advanced words are also inappropriate to express banal ideas; then they can sound "forced and contrived".

Accurate and appropriate word choice are criteria frequently mentioned in rating scales. However, the raters paid attention to one more criterion which they considered important for effective use of advanced words but which is never explicitly mentioned in the evaluation rubrics reviewed in Chapter 3—consistency. Sophisticated words should not only fit the register and content of a text but also its general lexical level.

Using one or two rare words in a text which is rather neutral and plain in terms of its lexis does not have a positive effect on the grade. The Polish raters called it "unnatural" use of sophisticated vocabulary.

All the four raters observed a marked difference in the use of vocabulary by the Polish and American students. American essays were written in a conversational style and employed simpler and colloquial lexis, whereas the style adopted by Polish students was more formal and the vocabulary of (some of) the Polish texts was more varied and sophisticated, even "fairly impressive". However, such a counter-intuitive disparity between foreign language learners and native speakers may not be a consequence of discrepancies in the Polish and American students' vocabulary knowledge but rather of the different genres chosen by the two groups. The Polish writers approached the task as an academic essay whereas the American writers treated it as journalistic prose. This explanation, suggested by one of the raters, was confirmed by the Polish and American teachers who had collected the data. One rater also observed that the topic may also affect the sophistication of vocabulary in an essay. He remarked that it is difficult to come up with 'interesting' lexis in an essay on the mobile phone. The influence of the characteristics of a task on language assessment and a level assigned to a student was demonstrated in earlier publications, discussed in Chapter 3 (Bouwer et al., 2015; In'nami & Koizumi, 2016) and addressed in the Common European Framework of Reference (Council of Europe, 2001).

6.6 Discussion and Conclusion

Each of the analyses discussed has offered interesting, perplexing and sometimes even contradictory insights into the lexical quality of the analysed texts as well as the instruments and procedures applied in their evaluation. The aim of this section is to pull these ideas together and reinterpret them in a larger context of assessment of lexical proficiency and lexical competence.

The interviews with the raters explained rather puzzling results obtained in the comparison of the indices in the three groups. The reason why almost all indices demonstrated lowest mean values for the essays written by the native students turned out to be a consequence of adopting different writing styles and taking a different line of argumentation by the Polish and American subjects. These differences occurred in spite of a careful study design which ensured that the participants had comparable educational profiles (university students), wrote essays in the same genre (opinion essay), on the same topics (the mobile phone), in the same context (academic writing course) and for the same audience (course instructor). While the Polish writers approached the assignment as an academic essay and tackled the problem from a more general perspective, the American writers interpreted it as calling for journalistic prose and

> Mobile phones are mainly used because they don't have to be stationary. You can go practically anywhere and are still able to use it. Many would say that they cause more problems than they do good, but that's your own opinion.
>
> Some of the benefits of having a mobile phone is that you don't have to be in your home when you use it. You can go to work, the mall, grocery store, wherever you want and still have communication with others. They also have instant access to 911 if there was an emergency or you saw an accident.

Figure 6.2 A Fragment of an Essay Written by a Native Student (source text code: una2f220)

> The invention of a device called a mobile or a cellular phone has proved to be a landmark discovery of our century. A number of people claim that it has not only enhanced our lives but also become an irreplaceable appliance in everyday mundane activities. However, as other modern inventions, this one has sharply divided people on those who are diehard supporters of this concept, and those who regard it as a curse of our times. The point is, that an appearance of a mobile phone in our everyday life has exerted a huge impact on the way we live.

Figure 6.3 A Fragment of an Essay Written by a Year 4 Student (source text code: u4a1fn34)

referred mainly to their personal experiences in supporting their opinion (cf. Leńko-Szymańska, 2006, who examined a larger sample of Polish and American essays on the mobile phone using the content-analysis methodology). This discrepancy can be illustrated by the fragments of two essays written by a native speaker and a Polish Year 4 student respectively (Figure 6.2 and Figure 6.3).

The two styles adopted by the Polish and American students may be equally valid and justified choices. An opinion essay is characterised in the following way in one of the oldest and most popular writing manuals targeted at grades 6 through 12 in the United States.

> A familiar sort of essay of opinion is the "letter to the editor". It is a written statement of the writer's belief about an arguable subject, supported by evidence, and written to convince. Actually, the essay itself is the culmination of a process all students are normally engaged in—an interest in controversial subjects, an examination of the pros and cons, and finally the formation of judgements about them.
>
> (Warriner & Griffith, 1977, p. 316)

At the same time, the opinion essay is defined in the following way in a handbook addressed to upper-intermediate EFL students produced by Express Publishing, a popular British publisher of EFL materials in Poland.

> Opinion essays are formal in style. They require your opinion on a topic which must be clearly stated and supported by reasons. It is necessary to include the opposing viewpoint in another paragraph.
> (Evans, 1997, p. 71)

The two quotations are symptomatic of varying approaches to this genre spread in their respective contexts. The American definition is not specific about the style in which the opinion essay should be written, but it defines it as a journalistic genre. On the other hand, the advice offered to EFL students does not mention academic writing per se, but it is very specific about the style particularly characteristic of argumentative and academic writing. On a more general level the two approaches are reminiscent of the claim made by Contrastive Rhetoric (Connor, 1996), according to which, the differences between native and non-native production may occur not only on the microlinguistic level associated with overuse, underuse or misuse of specific language elements (such as vocabulary), but also at the macrolinguistic level, and be related to the ways discourse is structured by both groups of language users.

The two styles—argumentative/academic and journalistic—sanction the use of very different vocabulary. As in was mentioned in Chapter 4, journalistic style manuals advise using simple words to make a text more accessible for a wide audience, whereas argumentative/academic texts require advanced, academic vocabulary and specialist terminology. At the same time describing one's own everyday experiences necessitates the use of more concrete and specific words than making generalisations. The individual indices computed for the essays in this study captured exactly these differences in style and content between the two groups. They revealed that the American compositions employed more frequent, concrete, imaginable, interconnected, specific and polysemous lexis than the Polish essays and tended to use a larger proportion of very frequent collocations. This difference was so pronounced that eight selected indices classified jointly the native texts as a separate category almost with a perfect precision.

This finding indicates that lexical indices are very sensitive to differences in style, genre and content. Also the topic can influence vocabulary choices in a profound way, as observed by one of the raters in the interview. Thus, observed differences between the mean values of lexical indices generated for different texts may not be a consequence of discrepancies in lexical proficiency of their authors, but rather their rhetoric choices in writing. These choices may reflect cultural constrains in which

texts were created. This observation is not new (cf. Salsbury et al., 2011, p. 16) but it is seldom explicitly addressed in research utilising such metrics. It challenges the validity and applicability of the indices for making judgements about lexical proficiency of fairly heterogeneous groups even if writing task conditions may seem strictly controlled.

In contrast to the indices, human raters appear to be able to interpret stylistic differences in texts. The marks awarded by the raters to learner and native essays did not reflect the pattern produced by the automatic gauges. Even though the raters commented on the superiority of Polish learner's vocabulary use over the American students, none of the raters seem to have penalised the American students for using simpler vocabulary. Such flexibility in evaluation was expressed by the judges in the interviews. Its existence was also confirmed by the results of the holistic assessment. Even though vocabulary use was reported as one of the criteria in the global evaluation, the Polish raters evaluated the American compositions as equally good as the Year 4 essays. One Polish rater even demonstrated in the regression analysis to use double standards and interpret the same lexical characteristics (productivity, diversity and sophistication) in the Polish and the American essays in a different way. The native-speaking raters were more severe in the holistic assessment of the native compositions, but they also did not mark the American essays lower than Year 1 texts. In addition, in the assessment which focused only on the lexical aspects of the essays, the raters evaluated the native texts as comparable to Year 1 and Year 4 compositions, with a statistically significant inferiority in relation to Year 4 essays only in the case of one rater (P2).

Another lexical characteristic of the native compositions spotted by the raters, which failed to be captured by automatic indices, was the use of technical vocabulary related to mobile phones. This observation can be exemplified by the following sentence:

1. Well it seems that there is no signal in that particular area of your plan. (una2m253)

An earlier analysis of a larger sample drawn from the same pool of compositions (Leńko-Szymańska, 2006) identified the following keywords in the American mobile phone corpus (Table 6.26). They were grouped into two themes frequently tackled by the native writers in their essays. These themes included instances of mobile-specific terminology.

Most of these technical terms appearing in the American texts are in fact very frequent English words with several senses, including special technical meanings. The use of these items as technical terms is a proof of the depth of lexical command, at least within this topic area, and—according to the Polish raters—boosted the marks awarded to the native essays. Yet, these terms cannot be picked up by the indices, as automatic

Table 6.26 Technical Keywords in the Native Essays

Theme	*Key Items Pointing to the Themes*
COST	minutes, plan, distance, service, plans, charges, free, month, roaming, rates, cards, cost, long
TECHNICAL NETWORK PROBLEMS	service, area, reception, static, tower

Source: Leńko-Szymańska (2006, p. 146)

measures do not distinguish between various meanings of individual words. This problem was recognised in more recent tools describing learner vocabulary. The English Vocabulary Profile[11] is a database of over 7000 items which assigns proficiency levels not to individual words but to their different meanings (Capel, 2010, 2012). For example, it lists the word *reception* with three senses, each assigned a different band: "the place in the hotel or office building" (B1), "a formal party" (B2) and "the way people react" (C1). None of these, however, captures the meaning with which the word is used in the American essays, as exemplified in the sentence:

2. Most cell phones run on digital reception, though there are some towers that are still analogue. (una2f205)

Unfortunately, there has been no automatic instrument available so far which would capture the occurrence of frequent vocabulary with more advanced or technical meanings in a text and consequently no index has been proposed to tap this aspect of vocabulary use. At the moment only human raters can record such nuances.

In addition to the differences in style, one more factor can explain the counter-intuitive poor results of the American essays, as tapped by the lexical indices. The Polish students, similarly to other foreign language learners, are likely to be aware of the dual function of writing in language assessment, i.e. a tool used to evaluate not only their writing ability but also their linguistic proficiency. In consequence, they may make a conscious effort to present their full linguistic potential in writing, while the native students may not feel such a need. The fact that such a strategy was indeed employed by the learners was confirmed by the raters in the interviews. They pointed out that a few Polish students seem to have peppered their texts with individual advanced words which did not match the general lexical level of the text and expressed rather trivial ideas. This can be illustrated by the excerpt in Figure 6.4.

Thus, various analyses conducted in this study have jointly revealed that comparing vocabulary use in essays written by mixed clusters of students (here native and Polish EFL students) may resemble comparing apples

> Another advantage is that a mobile phone can turn out unrivalled for mothers and fathers who worry about their children that live far away from home or commute to school. If a child does not come back home at the hour he or she used to, frantic with worry mother has an opportunity to get in touch with her beloved offspring and check whether something bad happened.

Figure 6.4 A Fragment of an Essay Written by a Year 1 Student (u1a1fn11)

with oranges. However, they also demonstrated that when the essays are written by fairly homogenous groups of writers—in terms of their cultural background, rhetoric approach or familiarity with the topic—the results of the assessment performed by both lexical indices and human raters may prove to be more meaningful for comparing their lexical proficiency. Several automatic measures used in this study indeed showed an improvement (increase or decrease) in essays written by Year 4 students when compared to compositions produced by Year 1 participants. These performing indices tapped three dimensions of lexical proficiency: diversity, sophistication and productivity. More advanced learners wrote on average longer essays using more content word types and they used more varied and sophisticated vocabulary, including academic lexis. On the other hand, none of the other measures capturing lexical density, word psychological properties, sense relations and formulaicity demonstrated a significant difference between the two groups of learners. Yet, a closer look at their mean values reveals that out of all 31 indices tapping these aspects, only three changed in the direction opposite to the predicted one. These changes are too small to have a statistical significance, but the fact that almost 90% of them demonstrates anticipated effects proves that these effects cannot be due to chance. It should be noted that the selected eight indices were also fairly successful in classifying the learners into two proficiency groups in the regression analysis, even though admittedly they are not as effective as in distinguishing between learners and native students.

The same tendencies were observed in the assessment performed by the human raters. The mean values of their marks in the holistic assessment, as well as of the marks assigned by three of them in the assessment of vocabulary were higher for Year 4 than Year 1. This improvement was also confirmed by the Polish raters in their interviews. They both remarked that the difference between the two learner groups laid mainly in the quality of their language, rather than in other aspects of their writing.

The interviews with the raters also highlighted another important characteristic of lexical assessment. The judges reported that they found it difficult to separate lexical proficiency from other elements of linguistic proficiency, grammar in particular, as well as from other aspects of

effective writing such as content, rhetoric and organisation. This observation confirms the fact that communicative language ability—i.e. lexical, grammatical, textual, pragmatic and strategic competences (Bachman, 1990; Bachman & Palmer, 1996) develop in parallel in a typical learner. It also implies that all these other traits, including grammar, are related to vocabulary use. That is why the raters reported that they found it difficult to mark the essays based solely on the use of vocabulary.

The results also indicated that the text lexical features could jointly account for 15% to 26% of the variance in the holistic marks assigned by the judges. At the same time the results revealed a complex relationship of the holistic marks with the text lexical qualities tapped by the individual automatic indices. In spite of the fact that the measures of lexical productivity, diversity, sophistication, sense relations and formulaicity were associated with the holistic grades, only lexical productivity seemed to contribute consistently to the scores in a significant way in the regression analysis. The remaining features showed effects only for individual raters, sometimes opposite to the predicted ones, or showed no significant effects at all. An even more complex relationship was noted in the case of the lexical grades. Even though the raters were asked to focus specifically on vocabulary, the lexical qualities tapped by the indices predicted the score of only two raters. In their cases the lexical characteristics had a more significant effect on the overall results (explaining over 30% of the variance in the grades), but here again only productivity consistently contributed to the marks. The grades for vocabulary awarded by the other two judges could not be predicted from the values of the lexical indices.

An explanation of these complex relationships can again be found in the interviews with the raters. The judges mentioned in particular the influence of three features on their marks—formulaicity, lexical sophistication and diversity, yet they also maintained that their effects are filtered by additional conditions of word use such as accuracy, appropriateness for the register, appropriateness for the expressed ideas and the topic, consistency with the text's general lexical level (naturalness) and avoidance of clichéd language. It can be argued that some of these conditions, in particular accuracy and appropriateness represent another dimension of lexical proficiency omitted from the model proposed by Bulté et al. (2008) and addressed only in earlier studies on measuring vocabulary use in learner production (Linnarud, 1975, 1986; Arnaud, 1984; Engber, 1995). Unfortunately, lexical accuracy and appropriateness cannot be measured automatically and require manual coding of errors in an essay. This process is particularly tricky in the case of advanced learners as it is sometimes difficult to distinguish between inaccurate/inappropriate and creative uses of a word (cf. Lewandowska-Tomaszczyk et al., 2000; Leńko-Szymańska, 2002). This problem became evident in analysing the category of bigrams absent from the reference corpus (Analysis 2). The other conditions of word use mentioned by the raters pertain to a larger

construct of writing ability, and represent strategic competence. They also cannot be easily quantified, although such attempts are being made in the field of artificial intelligence. Latent Semantic Analysis (Landauer et al., 1998; Martin & Berry, 2007; Crossley et al., 2010b), discussed in detail in Chapter 4 is an example of such an endeavour. This area of automatic measurement, however, is still in its infancy.

The fact that most lexical indices correlated with one another with varying strengths proves that lexical proficiency is not an amalgamation of independent traits, but that each of its dimensions simultaneously covers elements of another one. Jarvis (2013) included word rarity (sophistication) in the definition of lexical diversity. It can be argued that lexical elaborateness and lexical productivity are also part of lexical diversity because lexically varied texts entail the use of synonyms and other semantically related items, and require a production of larger quantities of words in general. By the same token, lexical accuracy is related to bigram-based measures of formulaicity as it covers, among other problems, collocational errors. As pointed out by Bulté et al. (2008), individual indices do not tap single components of lexical proficiency. This interdependence of individual aspects of lexical proficiency was also evident in the regression analyses. Even though the lexical indices could jointly predict the authorship of the essays with a fairly satisfying accuracy and account for a certain amount of variance in holistic and vocabulary marks, the effects of individual measures were small and rarely statistically significant.

The analysis of the scores produced by the Vocabulary Levels Tests sheds additional light on lexical proficiency as a performance-based reflection of lexical competence. The fact that the results of all three assessment procedures—(1) the test scores (which tap vocabulary knowledge more directly), as well as (2) the lexical indices and (3) the raters' marks—registered a growth between Year 1 and Year 4 confirms that lexical competence and lexical proficiency indeed develop over time, even in more advanced learners. This conclusion is further supported by the observed relationships between the results of the productive test and the individual indices tapping various components of lexical proficiency as well as the raters' scores. The fact that these associations are weak indicates that both the indices and the marks are also affected by other characteristics of texts.

All the observed effects discussed in this section call for a reappraisal of the three approaches to lexical assessment and for an updated model of lexical proficiency and lexical competence, which will be presented in the Conclusion.

Notes

1. In the case of Year 4 students the distribution of the essays into sets was motivated by the availability of Vocabulary Levels Tests completed by the same informants.

2. Technically, they were replaced with a full stop to stop bigram extraction.
3. The reference list representing American English available in *TAALES* is based on the *Brown* corpus (Francis & Kucera, 1964) or genre-specific sections of the *COCA*. Since the *Brown* corpus is a small and fairly old corpus and the *COCA* results reflected genre differences, the decision was to choose the BNC as reference.
4. Special thanks to Michał Dąbrowski, a student at Polish-Japanese Academy of Information Technology in Warsaw for his invaluable help with these computation procedures.
5. For the sake of clarity of presentation only significant results are reported in the large tables in this chapter.
6. Technically, this is the index that expresses the proportion of the bigrams whose MI<3, and the association is negative.
7. Several indices were normally distributed in each of the three groups, even though they did not demonstrate normal distribution for the whole sample (cf. the previous section).
8. For the Welch test, an adjusted omega squared formula was used (https://statistics.laerd.com/premium/spss/owa/one-way-anova-in-spss-18.php).
9. As with the comparisons between the indices, the ANOVA tests are employed in spite of the fact that all but two indices in the three groups are not normally distributed (Lix et al., 1996; Maxwell & Delaney, 2004).
10. A linear regression can be performed even if the predictor variables are not normally distributed if other assumptions are met (Kleinbaum, Kupper, Muller, & Nizam, 1998; Thomas Lumley, Diehr, Emerson, & Chen, 2002).
11. http://vocabularypreview.englishprofile.org/staticfiles/about.html (accessed 22 December 2018)

Conclusions

7.1 Three Approaches to the Assessment of Lexical Proficiency—Reappraisal

The main aim of this volume has been to review three approaches to the assessment of lexical proficiency. More specifically, these approaches included indirect, discrete-point vocabulary tests and well as tasks eliciting larger samples of learners' written or oral production whose 'lexical texture' (Linnarud, 1986) can then be evaluated by trained human judges or measured with the help of a variety of statistical metrics. The characteristics of each of these approaches were reviewed in Chapters 2, 3, 4, and the analyses and comparisons of their results reported in literature were presented in Chapter 5. The results of a new and comprehensive study juxtaposing the three approaches applied to assessment of the same three groups of subjects—an upper-intermediate and an advanced group of L2 English learners as well as a group of equivalent native speakers of English—were detailed in Chapter 6. This study produced some very unexpected and puzzling results which shed a new light at the concept of lexical proficiency and its assessment.

In accordance with the hypotheses formulated before the study and with basic intuitions concerning language learning, more advanced learners achieved significantly better results in the assessment of their lexical proficiency carried out by means of the three approaches. The results in all but one section of the receptive and productive vocabulary tests were significantly higher for Year 4 students than Year 1 students. The essays written by Year 4 students were also evaluated significantly higher for both general quality as well as its lexical content by four independent raters, irrespective of the fact that on the whole there were rather weak or at most moderate associations between the scores awarded by individual judges. The assessment carried out with the statistical measures of lexical proficiency also produced similar results. Eight different metrics, related to lexical productivity, lexical diversity and frequency-based lexical sophistication also discriminated well between the two groups of learners. The remaining 36 indices used in the study (except for three relating

to sense relations and phraseology) showed the expected increase or decrease, but the differences were too small to be statistically significant.

However, the counter-intuitive results were noted for the native subjects. The group was not asked to complete the vocabulary tests, as they had been designed specifically with L2 learners in mind. The tests cover the vocabulary up to the most frequent 10,000 words in English, whereas a modest estimate of the lexicon of an average educated native speaker gives the number of 20,000 word families (Nation & Waring, 1997). Yet, in the assessment performed by human raters, the American essays did not come out as better than the writing produced by L2 learners, either in terms of their general quality, or only its lexical texture. In the holistic assessment, the native raters evaluated the compositions as comparable with upper-intermediate learners' production. The Polish raters were more generous and gave the native essays marks similar to advanced learners' texts. In the assessment focusing on vocabulary, the raters placed native composition in between Year 1 and Year 4 groups, with no statistically significant differences with either of the learner groups (with one exception). The statistical indices applied to the American texts produced even more intriguing results. Except for two metrics of productivity and the gauge of lexical density, all the remaining 15 purely lexical indices indicated that the lexical qualities of the native texts were at a lower level than in both groups of L2 learners, and for all but one of these indices, these differences were statistically significant. The phraseological gauges presented a more ambiguous pattern which is much harder to interpret, but several gauges relating to three types of collocations—very rare word pairings, strongly associated collocations and common collocations—produced results contrary to the hypotheses.

The explanation of such an unexpected outcome was found in the interviews with the raters. They pointed out that the American students used a different genre for addressing the assignment question than the Polish learners, which had an enormous influence on the vocabulary choice in the native and Polish compositions. The former were more 'chatty' and employed simpler lexis, which was, however, consistent with the adopted style, whereas the latter tackled the question in more general terms and were written in academic style, which resulted in the use of a larger number of academic and infrequent words.

This observation provides an important insight into assessment of lexical proficiency, which was already embarked upon in the model of language use proposed by Bachman and Palmer (1996) and the paradigm of vocabulary ability proposed by Chapelle (1994)—both discussed in Chapter 1, but which is frequently overlooked in more recent theoretical models of L2 vocabulary or in empirical studies. Language use—and thus vocabulary use—always takes place in context and is not only the result of the language user's competence, but is also affected by contextual constrains and the language user's more general cognitive and affective

variables, such as familiarity with the topic, a goal, attitude and motivation to complete the assignment well. Strategic competence, which is the central component in Bachman and Palmer's model, assesses the demands of the task. Proficient language users, including native speakers and fairly advanced learners, have already developed substantial strategic competence, and have a fairly vast repertoire of language resources—including vocabulary and phraseology—at their disposal. The interaction of these two elements allows such users to choose from a large stock of available words and phrases, the ones that they find most appropriate for the task at hand. The users are capable of varying their vocabulary choice to suit the genre that they want to apply and the effect they want to create, such as humour, irony, showing involvement, sounding eloquent, etc. Thus, vocabulary use is determined not only by lexical competence, but also by pragmatic constrains, and it is coordinated by strategic competence.

Lexical proficiency was defined in Chapter 1 as a behavioural—i.e. observable and measurable—construct which reflects the L2 learner's ability to use L2 vocabulary for communication. It was also pointed out that lexical proficiency is frequently assumed to be a manifestation of the learner's underlying declarative and procedural knowledge of L2 vocabulary. However, the study demonstrated that as the language user's strategic competence increases, lexical proficiency becomes less and less a direct reflection of the breadth, depth and accessibility of his or her lexical resources and becomes increasingly a reflection of his or her awareness of various pragmatic aspects of word choice. This was already indicated by Chapelle (1994), who included not only lexical competence (i.e. vocabulary knowledge and fundamental processes) but also the context and metacognitive strategies of vocabulary use in her framework of vocabulary ability. However, her paradigm of lexical proficiency has not been taken up by other researchers or applied in empirical studies related to vocabulary use, which ignore the broader context and learners' individual choices to handle the topic in their analyses (see Chapter 5 for examples).

Another conclusion from the review of related literature, as well as from the new study described in this volume indicates that while the three approaches to the assessment of vocabulary proficiency tend to discriminate between different groups of subjects, there is a weak or at best moderate relationship between their results. Except for the metrics of lexical productivity, which have already been widely demonstrated to produce highest correlations with essay quality scores (e.g. Ferris, 1994; Jarvis, Grant, Bikowski, & Ferris, 2003; Leńko-Szymańska, 2015), the grades awarded by the raters in the holistic assessment and the assessment of vocabulary showed rather weak correlations with text lexical characteristics captured by individual indices. While the effects for the measures of lexical productivity, diversity and frequency-based sophistication were fairly stable among the raters, the effects for the remaining indices seemed

random. The regression models taking into account eight indices in order to predict raters' grades awarded solely to vocabulary explained 35% of variance in the scores produced only by two out of four raters. While the predictive power of these models was fairly high, it also indicated that the remaining 65% of variance was a result of other features than those covered by different indices.

Interviews with the raters again clearly pointed out what these other features could be. One of the significant areas of influence of human perceptions of lexical proficiency is accuracy. Accuracy is considered one of the key characteristics of language proficiency in the CAF model discussed in detail in Chapter 1. However, the paradigms developed specifically to describe lexical proficiency, also described in Chapter 1, do not take this feature into account. In addition, until now few statistical algorithms have been proposed in order to measure this aspect of proficiency, especially lexical proficiency (see Chapter 4 for a review), and none of them is automatic. The reason for this—also discussed in detail in Chapter 4—is that lexical errors are extremely difficult to identify and classify. An additional challenge is establishing error gravity: some errors are more disturbing than others and have a larger influence on a rater's negative perception. The raters also emphasised the importance of a feature which they referred to as natural use of vocabulary, by which they meant that words should be selected appropriately for the genre, purpose and the general lexical content of the text. This observation suggests that accuracy is a much broader concept, encompassing not only a lack of errors, but also suitability for context understood in very broad terms. At the moment, there is no fully established measure which can cover this aspect of vocabulary use. Thus, the fact that lexical accuracy and appropriateness are not measured in the assessment performed by statistical indices, but clearly taken into account in the assessment performed by human raters, can be a reason for rather weak correlation among these two types of assessment.

Another conclusion that can be drawn from the comparison of the behaviour of various indices and the interviews with the raters is a lack of adequate measures referring to all the complexity aspects of lexical proficiency. Lexical diversity and lexical sophistication seem to be well established characteristics of lexical proficiency which have repeatedly been demonstrated to discriminate well between students at various stages of language development or L2 learners and native speakers[1], provided certain conditions are met (i.e. texts represent the same genre and students show the same level of motivation). They indeed show that more advanced users of a language use a wider range of words as well as a larger proportion of less frequent words in a language to express themselves. In addition, the studies reviewed in Chapter 5, as well as the study reported in this book, demonstrated that learners at more advanced stages of language development write texts which contain a larger number of

less specific, concrete, imaginable, meaningful and polysemous words. However, it is not clear which aspect of lexical proficiency the new indices based on psychological word properties and meaning relations in fact capture. Crossley and his colleagues (various dates, e.g. Crossley et al., 2011a, 2011b) claim that they reflect the depth and access qualities of L2 lexicon and support their claim with numerous psycholinguistic experiments. Nevertheless, it does not seem clear why using a larger number of concrete, imaginable, meaningful and polysemous words is a proof of less in-depth knowledge of individual words or their accessibility. It seems more reasonable to claim that the proposed indices are alternative metrics of lexical sophistication defined through other features than word frequency.

While many different statistical metrics have been proposed to tap lexical diversity and lexical sophistication, the aspect of lexical complexity which seems to have been neglected in measurement is lexical elaborateness[2] defined by Bulté et al. (2008) as the occurrence of words displaying their more specific, peripheral and less frequent properties in language production. This important quality of lexical proficiency was mentioned by some of the raters and discussed in Chapter 6. One of the judges commented on the use of technical terms by native speakers. A further analysis demonstrated that these technical terms proved to be very frequent items of English with very specific meanings in the context of mobile technology and industry. Databases of words which differentiate between various meanings of lexical items and rank them in order of their frequency or a typical order of acquisition have already been compiled (cf. English Vocabulary Profile[3]). However, so far no automatic tool has been proposed which would be capable of sense disambiguation of words used in L2 learners' production. If such a tool existed, the elaborateness of a text's lexis could be measured by, among other metrics, a ratio or a mean expressing the level/peripherality/frequency of individual senses rather than word forms or lemmas. Such an attempt was made by Leńko-Szymańska (2015); however the ratio-based index was computed only for small samples of texts produced by language learners, as both sense disambiguation and coding of individual words had to be performed manually.

In their definition of lexical elaborateness as referring to specific, peripheral and less frequent properties of words, Bulté et al. indicated that these features do not have to relate only to the semantic component of word knowledge, but can also pertain to collocational, grammatical or pragmatic elements. This important characteristic of lexical proficiency was particularly emphasised by the raters. They valued "fluidity" but also the absence of slogans and clichéd language. Several phraseological indices based on t-score and MI value analysed in the reported studies captured exactly these properties of word use. In essence, t-score taps frequent word combinations made of frequent items. The higher the

score, the more 'usual' the combination is as far as its grammatical and collocational properties are concerned. On the other hand, word combinations with higher MI value capture rare but strongly associated word pairings. The category of bigrams absent or very infrequent in a corpus taps the use of novel word pairings. Thus, this family of indices can be proposed not only to describe the formulaicity of a text but at the same time the elaborateness of the author's lexical proficiency. In addition, the highly complex measure called Latent Semantic Analysis (discussed in Chapter 4) may tap the collocational but also pragmatic aspects of lexical elaborateness.

The research on the families of measures described previously: metrics based on individual senses of words, or gauges capturing the types of a word's lexical, grammatical and pragmatic relations with other words, is still in its infancy, but they show a promising avenue of tapping lexical elaborateness, which for now is not adequately measured by existing statistical indices.

As the interviews have demonstrated, raters are capable of perceiving and taking into account these subtle nuances of word choice which are not adequately covered by statistical measures. At the same time the judges admit that they find it difficult to separate the assessment of vocabulary from other aspects of production such as grammar, content, rhetoric or organisation. That is why analytic marking of a text is more challenging than its holistic evaluation, in particular if only one aspect of production is to be assessed (e.g. vocabulary) and the others disregarded completely. The raters also revealed different attitudes towards evaluating L2 production. Some judges put emphasis on the communicative properties and effectiveness of a text while the others give equal importance to its linguistic qualities, including vocabulary. Admittedly, in the study the raters did not have assessment rubrics at their disposal and they had not been trained for this particular assessment task. Yet, the influence of rater's beliefs on the final marks has also been researched in studies with a more rigorous design as far as assessment criteria are concerned and its existence has been confirmed (see Chapter 3 for details). Thus, the raters' marks, even though encompassing more fine-grained details concerning vocabulary use tend to be subjective and the more detailed they are, the less reliable they become. That is why asking raters to evaluate various aspects of lexical proficiency separately may be counterproductive.

In a broader perspective, the interviews confirmed the perceptual validity of the cognitive linguistic models of language discussed in Chapter 1. They indicate that it is impossible to view lexis as a self-contained module of language, clearly separated from other linguistic elements such as phonology/orthography, grammar or pragmatics and even more general cognitive components such as topical knowledge. These and other aspects of language and cognition permeate one another, which becomes particularly evident at advanced levels, when language users' production

becomes more refined and nuanced in terms of its complexity, accuracy and fluency. The use of words is intrinsically linked with the use of morphemes, collocations, idioms, fixed and semi-fixed phrases, partially filled phrase frames or even fully productive grammatical structures. They are all employed jointly to convey a precise meaning and to achieve a particular effect in an actual context. A higher proficiency involves an awareness of this rich network of interdependence of various linguistic and general cognitive components and its manifestation in language production.

Finally, the discussion in the volume proposed that solving traditional discrete-point lexical tests also involves some strategic competence and affective states, but since the context of vocabulary use is more limited and strictly controlled, learners' use of various strategies is more comparable and predictable. That is why lexical tests allow for the assessment of the test-taker's lexical competence in a more direct, thorough and reliable way than performance-based assessment. At the same time discrete-point tests give little information about the test-taker's ability to use the same words at his or her own will in real-life situations, which is the ultimate goal of language instruction and thus assessment.

The present discussion clearly points to the merits and drawbacks of each of the three approaches to the assessment of lexical proficiency. Indirect discrete vocabulary tests give us a better picture of a learner's underlying lexical competence but they fail to demonstrate how he or she is capable of using vocabulary in uncontrolled situations for communication purposes. In performance assessment, on the other hand, lexical competence is filtered through strategic competence and contextual constrains and it is hard to separate each individual construct. Human raters are capable of perceiving this interdependence and taking it into account in their evaluation, yet their scores blur the distinction between lexical proficiency and other aspects of language production, and are characterised by a degree of subjectivity. Statistical indices are fully objective, but they do not tap all aspects of lexical proficiency equally well. They fail to account for the characteristics of the context situation of language use and if applied to texts which are not strictly identical in terms of topic and genre they may produce distorted results. That is why validating one method against the other, as is frequently described in relevant literature (cf. Chapter 5), has its serious limitations. The three approaches should not be treated as alternative but complementary methods of assessment of lexical proficiency, each contributing its unique information to the final mark. This idea is suggested in the following quotation:

> Proficiency, as I am defining it here, refers to a measure of one or more of the four language skills at a particular moment in the learner's learning trajectory as operationalized through some kind of task. In the case of the receptive skills, a proficiency test will typically measure the degree of input comprehension and only *indirectly* lexical

> and grammatical knowledge. In the case of the productive skills, a test will (usually, and among other things) measure lexical richness, complexity, and accuracy. However, a measure of lexical richness, complexity, and accuracy in a proficiency test (e.g. the IELTS) will not be a reliable measure of the learner's lexicon or current interlanguage. This is so because a measure of proficiency will *perforce include* strategic behaviour as defined earlier. Even in a simple task, . . . , the outcome measure (proficiency in writing) will have been affected by the strategies adopted in relation to the linguistic knowledge that the writer possessed at that particular time. Thus, when we consider the relationship between strategic behaviour and proficiency we have to measure *linguistic knowledge* first. This is a crucial first step. Indeed, without a measurement of linguistic knowledge, strategic behaviour and *proficiency* risk getting seriously confounded.
>
> (Macaro, 2014, p. 59)

Uniting the three approaches to testing lexical competence has already been employed in the *TOEFL IBT* exam. Its reading section includes several questions addressing explicitly the understanding of individual words in context. The writing section is evaluated by one human rater and an automatic scoring system called *e-Rater*®.[4] The assessment rubrics for the independent task include the criteria relating to word choice, idiomaticity and lexical accuracy. *E-Rater*® includes several indices related to lexical sophistication, topic-specific word usage, collocational diversity, incorrect word forms, repetition of words and inappropriate words or phrases (Ramineni et al., 2012, p. 34). However, in all the three cases the test-taker's lexical proficiency is assessed as a component of a larger construct of language proficiency and a separate score for vocabulary is not produced.

7.2 An Extended Model of Lexical Competence and Lexical Proficiency for Assessment Purposes

The earlier discussion has also pointed to the need for an extended model of lexical proficiency which will reconcile the paradigms discussed in Chapter 1 as well as several observations made in the study. This model is presented in Figure 7.1.

The proposed model indicates that lexical proficiency is an idealised construct within the blueprint of language representation and use since it cannot be separated from other linguistic and non-linguistic components of cognition. As all language is transmitted lexically, it is impossible to evaluate vocabulary knowledge and ability on its own. For example, rich topical vocabulary and phraseology are not only a proof of the size and organisation of lexical competence use but also of a good topical knowledge. Many grammatical or pragmatic mistakes have their lexical

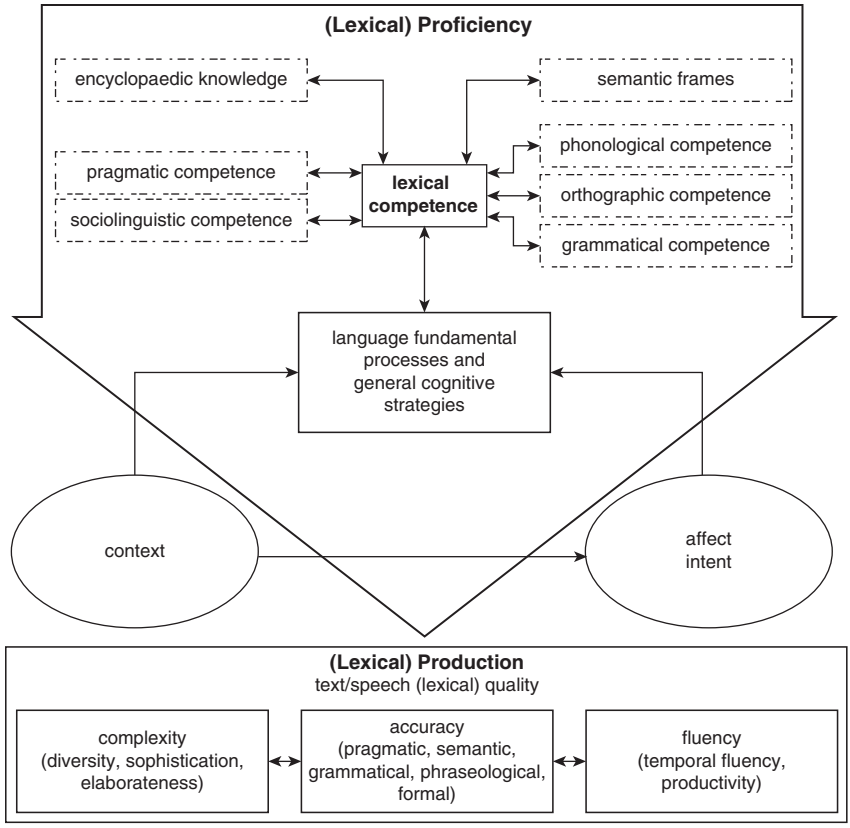

Figure 7.1 A Model of Lexical Proficiency

roots such as unfamiliarity with a word's grammatical behaviour or its pragmatic use constrains. Discourse organisation is managed through a skilful application of well-tried lexical phrases. Lower-level skills such as mechanics of writing also involve mastering correct spelling of individual words. Finally, metacognitive procedures orchestrate an appropriate co-selection of vocabulary and other linguistic elements in order to suit the situation and achieve a desired effect. That is why lexical competence is not a self-contained network of lexical units. It is interwoven with other linguistic systems representing phonological, orthographic, grammatical, sociolinguistic, pragmatic and other competences as well as with broader cognitive structures such as semantic frames or encyclopaedic knowledge. A proficient language user not only needs to know many lexical items (the size component of the lexicon), but these items ought to be richly interconnected with other lexical items (the width component) and

with other linguistic information such as morphological, grammatical or pragmatic specifics (the depth component). With recurring exposure and repetition, the sequences of cognitive processes which link these lexical and non-lexical elements as well as their triggering contexts become proceduralised chains, which increases their accessibility. The interconnected items become gradually entrenched in the cognitive organisation to the point of becoming a single unit. In this way the lexicon gradually transforms into the construction and the boundaries between lexical proficiency and other linguistic and more general cognitive proficiencies become blurred.

The model also highlights the claim that lexical proficiency represents an ability rather than knowledge. It refers to a capacity for language use (comprehension and production) which involves access, retrieval and application of lexical items stored in one's memory. Lexical proficiency draws heavily on lexical competence—i.e. the knowledge of words and their various characteristics—but also includes a range of cognitive and metacognitive procedures which govern lexical selection. A proficient language user not only needs to possess a rich and interconnected word store, but also a fully operational array of strategies which enable him or her to evaluate a situation and its constrains, set goals for language use and match all of these with relevant lexical resources.

Finally, this model also captures the idea that lexical proficiency manifests itself in lexical production. Lexical production can be defined as an act of employing vocabulary in context characterised by the features of complexity, accuracy (including appropriacy) and fluency. It offers a distorted reflection of lexical competence as it discloses only those aspects which are pertinent for a given event of language use, i.e. the context and the language user's intention and emotions at this particular moment. Its separate perceptual features make it hard to evaluate it as a single construct. Although the correlations among the indices tapping various aspects of vocabulary production are moderate and high (cf. Chapter 6)—indicating that these components related to one another—it cannot be assumed that all language users will demonstrate the same profiles of lexical use. Even the same learners will show different strengths and weaknesses in different situations. The three aspects: complexity, accuracy and fluency co-exist in a shaky equilibrium and on different occasions one can take precedence over the others. Thus, more fluent speech or writing may contain more lexical errors. In a similar way fewer lexical errors can be a result of a lower lexical complexity of the output. Finally, the same vocabulary can be appropriate in one context but not in another. As in the case of lexical proficiency, aspects of lexical use are hard to separate from other aspects of language production such as formulaicity, grammatical complexity, organisation and content.

This is not to say that the concepts of the lexicon, lexical competence, lexical proficiency and lexical production are invalid as they do not

reflect the complex nature of linguistic representation, processing and use. Admittedly, they are idealised constructs, but they play a crucial role in language description on the perceptual, theoretical and applied levels. They reflect basic human conception of language and its components. Lexis features in all theoretical models of language, even when formal approaches undermine its importance in the system, and functional and cognitive approaches question its distinctiveness from other linguistic and non-linguistic components of cognition (cf. Chapter 1). Words are the essence of lexicography and they form an indispensable element of language teaching. However, when accounts of real language use and its underlying proficiency and competence need to be made, this simplification starts to crack and becomes inadequate. For example, rule-based machine translation systems based on dictionaries and algorithms simulating grammatical rules have proved to be inefficient. They have given way to phrase-based statistical systems and more recently to neural machine translation systems which are probabilistic and represent linguistic information as a rich network of possible contexts. In the same way, the assessment of advanced language learners focusing on the evaluation of individual language systems and ignoring the importance of non-linguistic factors becomes problematic. That is why we need different approaches to assessing lexical proficiency, each highlighting its different aspects. When applied together they can give a more truthful picture of the L2 learner's lexical ability.

Notes

1. Interestingly, in the study described in this volume the older and simpler indices tapping these characteristics seem to perform this task better or equally well than the more recent and more complex ones.
2. In Bulté et al.'s (2008) model this feature was referred to as complexity.
3. www.englishprofile.org/wordlists
4. www.ets.org/toefl/ibt/scores/understand/

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Aitchison, J. (2003). *Words in the mind. An introduction to the mental lexicon* (3rd ed.). Malden, MA: Blackwell.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 658–662.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Ander, S., & Yıldırım, Ö. (2010). Lexical errors in elementary level EFL learners' compositions. *Procedia—Social and Behavioral Sciences*, *2*(2), 5299–5303. https://doi.org/10.1016/j.sbspro.2010.03.864

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Anthony, L. (2014, July). *A view to the future in corpus tools development*. Presented at the 11th Teaching and Language Corpora Conference (TALC 11), June, Lancaster.

Anthony, L. (n.d.). *AntWordProfiler*. Retrieved from www.laurenceanthony.net/software/antwordprofiler/

Arnaud, P. J. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing papers from the international symposium on language testing* (pp. 14–28). Colchester: University of Essex.

Arthur, B. (1979). Short-term changes in EFL composition skills. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 330–342). Washington, DC: TESOL.

Astika, G. G. (1993). Analytical assessments of foreign students' writing. *RELC Journal*, *24*(1), 61–70. https://doi.org/10.1177/003368829302400104

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Needham Heights, MA: Allyn & Bacon.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, *11*(1), 17–34. https://doi.org/10.1017/S0272263100007816

Barnwell, D. P. (1996). *A history of foreign language testing in the United States from its beginnings to the present*. Tempe, AZ: Bilingual Press.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word level and University word level vocabulary tests. *Language Testing*, *16*(2), 131–162.

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Berlin, B., Breedlove, D. E., & Raven, P. (1974). *Principles of Tzeltal plant classification*. New York, NY: Academic Press.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Bialystok, E., & Sherwood-Smith, M. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second language acquisition. *Applied Linguistics*, *6*(6), 101–119.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Bloomfield, L. (1933). *Language*. New York, NY: Holt, Rinehart, and Winston.

Bogaards, P. (2000). Testing L2 vocabulary knowledge at a high level: The case of the Euralex French tests. *Applied Linguistics*, *21*(4), 490–516. https://doi.org/10.1093/applin/21.4.490

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Brown, R. (1958). How should a thing be called? *Psychological Review*, *65*, 14–21.

Brumfit, C. J., & Johnson, K. (Eds.). (1979). *The communicative approach to language teaching*. Oxford: Oxford University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

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*, *18*(3), 277–298. https://doi.org/10.1017/S0959269508003451

Bybee, J. (1998). The emergent lexicon. In *CLS 34: The panels* (pp. 421–435). Chicago: Chicago Linguistics Society.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, *1*(1). https://doi.org/10.1017/S2041536210000048.

Capel, A. (2012). Completing the *English Vocabulary Profile:* C1 and C2 vocabulary. *English Profile Journal*, *3*(1). https://doi.org/10.1017/S2041536212000013

Carlsen, C. (2012). Proficiency level—A fuzzy variable in computer learner corpora. *Applied Linguistics*, *33*(2), 161–183.

Carroll, J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic perspective* (pp. 46–69). Oxford: Oxford University Press.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187.

Chapelle, C. A., & Brindley, G. (2002). Assessment. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 267–288). London: Arnold.

Chen, Y-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49.

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Connor, U. (1996). *Contrastive rhetoric*. Cambridge: Cambridge University Press.

Cook, V. (2002). Background to the L2 user. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 1–28). Clevedon: Multilingual Matters

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language*, *63*(3), 544–573. https://doi.org/10.2307/415005

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research*, *36*(3), 206–217. https://doi.org/10.1080/00220671.1942.10881160

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary

research and pedagogical applications. *System*, *41*(4), 965–981. https://doi.org/10.1016/j.system.2013.08.002

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., & Salsbury, T. (2011). The development of lexical bundle accuracy and production in English second language speakers. *IRAL—International Review of Applied Linguistics in Language Teaching*, *49*(1), 1–26. https://doi.org/10.1515/iral.2011.001

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010a). The development of polysemy and frequency use in English second language speakers: Polysemy and frequency use in English L2 speakers. *Language Learning*, *60*(3), 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010b). The development of semantic relations in second language speakers: A case for latent semantic analysis. *Vigo International Journal of Applied Linguistics*, *7*, 55–74.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi.org/10.1177/0265532211419331

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 105–134). Amsterdam: John Benjamins.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*(1), 182–193. https://doi.org/10.5054/tq.2010.244019

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The Development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*(8), 895–901.

Dale, E., & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*, 11–20.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–32). Cambridge: Cambridge University Press.

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234–244). Cambridge: Cambridge University Press.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150–164). Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. https://doi.org/10.1093/applin/24.2.197

Davies, A. (1989). Communicative competence as language use. *Applied Linguistics*, *10*(2), 157–170. https://doi.org/10.1093/applin/10.2.157

Davies, M. (2008). *The corpus of contemporary American English (COCA): 560 million words, 1990-present*. Retrieved from https://corpus.byu.edu/coca/

Davies, M. (n.d.). *Word frequency data*. Retrieved from www.wordfrequency.info/

De Boer, F. (2014). Evaluating the comparability of two measures of lexical diversity. *System*, *47*, 139–145. https://doi.org/10.1016/j.system.2014.10.008

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, *3*(1), 59–80.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory*. Amsterdam & Atlanta: Rodopi.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL—International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Eichholz, G., & Barbe, R. (1961). An experiment in vocabulary development. *Educational Research Bulletin*, 1–7.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Chichester: Wiley.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Evans, V. (1997). *Successful writing. Upper-intermediate. Student's book*. Swansea: Express Publishing.

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart: Institut Für Maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212–1248). Berlin: Mouton de Gruyter.

Eyckmans, J., van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behaviour in yes/no vocabulary tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 59–76). Cambridge: Cambridge University Press.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*(2), 414–420. https://doi.org/10.2307/3587446

Fillmore, C. J. (1977). Scenes-and-frames semantics. In A. Zampolli (Ed.), *Linguistic structures processing* (pp. 55–81). Amsterdam: North-Holland.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations in the L2 lexicon. *EUROSLA Yearbook*, *6*, 121–145.

Flahive, E. D., & Snow, B. G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 171–176). Rowley, MA: Newbury House.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299–324. https://doi.org/10.1017/S0272263100015047

Francis, W. N., & Kucera, H. (1964). *Brown corpus*. Retrieved from http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written English*. (TOEFL Research Report No. 64). Princeton, NJ: Educational Testing Service.

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, *18*(2), 173–181. https://doi.org/10.1016/j.asw.2013.02.001

Gajek, E. (2006). Zasoby i narzędzia internetowe w językoznawstwie korpusowym. In A. Duszak, E. Gajek, & U. Okulska (Eds.), *Korpusy w angielsko-polskim językoznawstwie kontrastywnym: teoria i praktyka* (pp. 311–327). Kraków: Universitas.

Gardner, D. (2013). *Exploring vocabulary: Language in action*. London & New York: Routledge.

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: An introductory course*. London: Taylor & Francis.

Gilhooly, K. J., & Logie, R. H. (1980). Meaning-dependent ratings of imagery age of acquisition, familiarity and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, *12*, 428–450.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224. https://doi.org/10.1016/S1364-6613(03)00080-9

Gougenheim, G. (1964). *L'élaboration du français fondamental (1er degré): étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.

Grabe, W. (2008). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Granger, S., & Paquot, M. (1998). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins.

Gries, S. T. (1998). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 3–25). Amsterdam: John Benjamins.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 21–33). New York: Palgrave Macmillan.

Gruszczyński, W., & Ogrodniczuk, M. (Eds.). (2015). *Jasnopis, czyli mierzenie zrozumiałości polskich tekstów użytkowych*. Warszawa: Wydawnictwo ASPRA-JR.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire: essai de méthodologie*. Paris: Presses universitaires de France.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Gyllstad, H. (2012). *Validating the vocabulary size test. A classical test theory approach*. Presented at the Ninth Annual Conference of EALTA, May, Innsbruck. Retrieved from www.ealta.eu.org/conference/2012/posters/Gyllstad.pdf

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—challenges and potential solutions. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 11–28). EUROSLA Monographs Series 2. Amsterdam: Eurosla.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective* (2nd ed.). Oxford: Oxford University Press.

Halliday, M. A. K., & Matthiessen, C. (2014). *Halliday's introduction to functional grammar* (4th ed.). Milton Park, Abingdon, Oxon: Routledge.

Hamp-Lyons, L. (1991a). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5–15). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Gravenhage: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. New York: Springer.

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology*, *4*(2), 686–713.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London and New York, NY: Routledge.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, *18*(1), 87–108.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Howarth, P. A. (1996). *Phraseology in English academic writing, some implications for language learning and dictionary making* (Reprint 2013). Berlin and Boston: De Gruyter.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*(4), 663–667.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J. H., Alderson, J. C., & Schoonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In I. B. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*. (pp. 11–20). EUROSLA Monograph Series 1. Amsterdam: Eurosla.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Urbana, IL: The National Council of Teachers of English.

Hyltenstam, K. (1992). Non-native features of near-native speakers: On the ultimate attainment of childhood L2 learners. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 351–368). Amsterdam: Elsevier.

Hymes, D., H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics. Selected readings* (pp. 269–293). Harmondsworth: Penguin.

Imiołczyk, J. (1987). *Prawdopodobieństwo subiektywne wyrazów: podstawowy słownik frekwencyjny języka polskiego*. Warszawa: PWN.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341–366. https://doi.org/10.1177/0265532215587390

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.

Jacobs, H. L., Zingraf, S., A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

James, C. (1998). *Errors in language learning and use: Exploring error analysis*. London: Longman.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–43). Amsterdam: John Benjamins.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403. https://doi.org/10.1016/j.jslw.2003.09.001

Johnson, W. (1944). Studies in language behavior. A program of research. *Psychological Monographs*, *56*(2), 1–15. https://doi.org/10.1037/h0093508

Jones, R. L. (1985). Second language performance testing: An overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15–24). Ottawa: University of Ottawa Press.

Katz, J., & Fodor, J. (1963). The structure of a semantic theory. *Language*, *39*, 170–210.

Kincaid, J. P., Fishburne, L. R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated readability index, fog count, and Flesch reading ease formula) for navy enlisted personnel* (pp. 8–75). Millington, TN: Naval Education and Training Support Command. Chief of Naval Technical Training.

Kleinbaum, D. G. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and multivariable methods* (3rd ed.). Pacific Grove, CA: Duxbury Press.

Kyle, K., & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*(1), 123–134.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal*, *75*(4), 440–448. https://doi.org/10.2307/329493

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Basingstoke and Hampshire: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33. https://doi.org/10.1177/003368829402500202

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. Studies in honor of William E. Rutherford* (pp. 265–272). Amsterdam: John Benjamins.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lee, Y-W., Gentile, C., & Kantor, R. (2009). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, *31*(3), 391–417. https://doi.org/10.1093/applin/amp040

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417.

Leńko-Szymańska, A. (2002). Lexical problem areas in the advanced learner corpus of written data. In B. Lewandowska-Tomaszczyk (Ed.), *PALC'2001. Practical applications in language corpora* (pp. 505–519). Frankfurt am Main: Peter Lang.

Leńko-Szymańska, A. (2006). The curse and the blessing of mobile phones— a corpus-based study into American and Polish rhetorical conventions. In A. Wilson, D. Archer & P. Rayson (Eds.), *Corpus Linguistics Around the World* (pp. 141–153). Ansterdam-New York, NY: Rodopi. https://doi.org/10.1163/9789401202213_012

Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics*, *19*(2), 225–251.

Leńko-Szymańska, A. (2015). The English vocabulary profile as a benchmark for assigning levels to learner corpus data. In M. Callies & S. Goetz (Eds.), *Learner corpora in language testing and assessment* (pp. 115–140). Amsterdam: John Benjamins.

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 127–148). EUROSLA Monographs Series 2. Amsterdam: Eurosla.

Lewandowska-Tomaszczyk, B., Leńko-Szymańska, A., & McEnery, T. (2000). Lexical problem areas in the PELCRA learner corpus of English. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *Practical applications in language corpora (PALC'99). Papers from the international conference at the university of Lódź, 15–18 April 1999* (pp. 303–312). Frankfurt am Main: Peter Lang.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85–102. https://doi.org/10.1016/j.jslw.2009.02.001

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL—International Review of Applied Linguistics in Language Teaching*, *49*(3), 221–240. https://doi.org/10.1515/iral.2011.013

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109–126). EUROSLA Monographs Series 2. Amsterdam: Eurosla.

Linnarud, M. (1975). *Lexis in free production: An analysis of the lexical texture of Swedish students' written work*. Swedish-English Contrastive Studies, report no. 6. Lund University.

Linnarud, M. (1977). Some aspects of style in the source and the target language. *Papers and Studies in Contrastive Linguistics*, *7*, 85–94.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmö: C.W.K. Gleerup.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, *66*(4), 579–619. https://doi.org/10.2307/1170654

Lorenz, G. R. (1999). *Adjective intensification: Learners versus native speakers: A corpus study of argumentative writing*. Amsterdam: Rodopi.

Lu, X. (2012a). *Lexical Complexity Analyzer*. Retrieved from www.personal.psu.edu/xxl13/downloads/lca.html

Lu, X. (2012b). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Maas, H-D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift Für Literaturwissenschaft Und Linguistik*, *2*(8), 73–79.

Macaro, E. (2014). Reframing task performance: The relationship between tasks, strategic behaviour, and linguistic knowledge in writing. In H. Byrnes & R. Manchón (Eds.), *Task-based language learning. Insights from and for L2 writing* (pp. 53–78). Amsterdam: John Benjamins.

Magee, S., & Rundell, M. (1996). *The role of the corpus-based 'Phrasicon' in English language teaching*. Presented the 2nd Teaching and Language Corpora Conference (TALC 96). July, Lancaster.

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01164

Malvern, 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., & 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., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke and Hampshire: Palgrave Macmillan.

Martin, D., & Berry, M. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Earlbaum Associates.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing*, *11*(1), 29–51.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International*, *66*, 12.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McCarthy, P. M., Watanabe, S., & Lamkin, T. A. (2011). The Gramulator: A tool to identify differential linguistic features of correlative text types. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 312–333). Hershey, PA: IGI Global.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, *15*(3), 323–338. https://doi.org/10.1093/llc/15.3.323

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

McNamara, D. S., & Graesser, A. C. (2011). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global.

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

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. C. (2005). *Coh-Metrix* (Version 1.4). Retrieved from http//: cohmetrix.memphis.edu

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, *13*(3–4), 221–246. https://doi.org/10.1017/S0261444800008879

Meara, P. (1990). Some notes on the Eurocentres Vocabulary Tests. In J. Tommola (Ed.), *Foreign language comprehension and production* (pp. 103–113). Turku: Finnish.

Meara, P. (1992). *EFL Vocabulary Tests*. Swansea: Centre for Applied Language Studies, University of Wales. Retrieved from www.lognostics.co.uk/vlibrary/meara1992z.pdf

Meara, P. (1994). *LLEX: Threshold Level Vocabulary Tests*. Swansea: Centre for Applied Language Studies, University of Wales.

Meara, P. (1996a). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–52). Cambridge: Cambridge University Press.

Meara, P. (1996b). The vocabulary knowledge framework. Retrieved from *Vocabulary acquisition research group virtual library*, www.lognostics.co.uk/vlibrary/meara1996c.pdf (10 October 2016).

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109–121). Cambridge: Cambridge University Press.

Meara, P. (2005a). Designing vocabulary tests for English, Spanish and other languages. In C. S. Gomez, G. Butler, C. Gómez González, M. de los Ángeles, & S. M. Doval-Suárez (Eds.), *The dynamics of language use: Functional and contrastive perspectives* (pp. 271–285). Amsterdam: John Benjamins.

Meara, P. (2005b). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, *26*(1), 32–47. https://doi.org/10.1093/applin/amh037

Meara, P. (2006). *Y_Lex: The Swansea Vocabulary Levels Test* (Version 2.05). Swansea: Lognostics.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

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.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80–87). London: CILT.

Meara, P., & Jones, G. (1990). *The Eurocentres Vocabulary Size Tests. 10K*. Zurich: Eurocentres.

Meara, P., & Milton, J. (2003). *X_Lex: The Swansea Vocabulary Levels Test*. Newbury: Express.

Meara, P., & Miralpeix, I. (2015). *V_YesNo* (Version 1.0). Swansea: Lognostics.

Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. *Angles on the English Speaking World*, *4*, 85–96.

Medgyes, P. (1992). Native or non-native: Who's worth more? *ELT Journal*, *46*(4), 340–349. https://doi.org/10.1093/elt/46.4.340

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Palgrave Macmillan.

Mihalcea, R., & Moldovan, D. (2000). Semantic indexing using WordNet senses. In *Proceedings of the ACL-2000 workshop on recent advances in natural language processing and information retrieval*: Held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics—Volume 11 (RANLPIR '00) (Vol. 11, pp. 31–45). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1117755.1117760

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 47–58). Cambridge: Cambridge University Press.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics*, *107–108*, 17–34.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol: Multilingual Matters.

Mullen, K. (1980). Evaluating writing proficiency. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 160–170). Rowley, MA: Newbury House.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*, 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 55–88.

Nation, I. S. P. (n.d.). *Range*. Retrieved from www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher*, *31*(7), 9–13, Appendix.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge, NY: Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Noble, C. E. (1952). An analysis of meaning. *Psychological Review*, *59*, 421–430.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1), 1–25.

Palmer, H. E. (1917). *The scientific study and teaching of languages*. London: Harrap.

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*(1), 9–29.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–200). Cambridge: Cambridge University Press.

Park, S-K. (2015). The interplay of task, rating scale, and rater background in the assessment of Korean EFL students' writing. *English Teaching*, *70*(2), 55–82. https://doi.org/10.15858/engtea.70.2.201506.55

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–126). London: Longman.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, *14*(1), 61–69. https://doi.org/10.2307/3586809

Pęzik, P. (2012). Towards the PELCRA learner English corpus. In P. Pęzik (Ed.), *Corpus data across languages and disciplines* (pp. 33–42). Frankfurt am Main: Peter Lang.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-Rater® scoring engine. *ETS Research Report Series*, *2009*(1), i–35.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-Rater® scoring engine for the TOEFL® independent and integrated prompts. *ETS Research Report Series*, *2012*(1), i–51.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*, 12–25.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum Associates.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–210). Cambridge: Cambridge University Press.

Reppen, R. (2009). Exploring L1 and L2 writing development through collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 49–59). New York: Palgrave Macmillan.

Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) second release LDC2005T35. DVD*. Philadelphia, PA: Linguistic Data Consortium.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. https://doi.org/10.2307/3585941

Rumelhart, D. E., McClelland, & PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching*. Amsterdam: John Benjamins.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343–360. https://doi.org/10.1177/0267658310395851

de Saussure, F., Bally, C., Sechehaye, A., & Riedlinger, A. (1916). *Cours de linguistique générale*. Lausanne and Paris: Payot.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., & McCarthy, M. (Eds.). (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211–236. http://dx.doi.org.atoz.han.buw.uw.edu.pl/10.1177/0265532207086782

Schwarz, M. (1995). Accessing semantic information in memory: The mental lexicon as a semi-module. In R. Dirven & J. Vanparys (Eds.), *Current approaches to the lexicon* (pp. 63–72). Frankfurt am Main: Peter Lang.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a Foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Stansfield, C. W. (1985). A history of dictation in foreign language teaching and testing. *The Modern Language Journal*, *69*(2), 121–128. https://doi.org/10.1111/j.1540-4781.1985.tb01926.x

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Swan, M. (1995). *Practical English usage* (2nd ed.). Oxford: Oxford University Press.

Taylor, D. S. (1988). The meaning and use of the term 'Competence' in linguistics and applied linguistics. *Applied Linguistics*, *9*(2), 148–168. https://doi.org/10.1093/applin/9.2.148

The BNC Consortium. (2007). *The British National Corpus* (version 3, BNC XML Edition). Retrieved from www.natcorp.ox.ac.uk/

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*(2), 307–336.

Thorndike, E. L. (1921). *The teacher's word book*. New York, NY: Teachers College, Columbia University.

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness in semi-spontaneous French speech. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 133–149). Cambridge: Cambridge University Press.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Oxford: Lawrence Erlbaum.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). Amsterdam: John Benjamins.

Tweedie, F., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352. https://doi.org/10.1023/A:1001749303137

Uden, J., Schmitt, D., & Schmitt, N. (2014). Jumping from the highest graded readers to ungraded novels: Four case studies. *Reading in a Foreign Language*, *26*(1), 1–28.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443–452). Cambridge: Cambridge University Press.

van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83. https://doi.org/10.1177/026553220001700103

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (Vol. 10, pp. 173–189). Amsterdam: John Benjamins.

Vidakovic, I., & Baker, F. (2010). Use of words and multi-word units in skills for life writing examinations. *Cambridge ESOL: Research Notes*, *41*, 7–14.

Waring, R. (1997). A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata*, (1), 53–68.

Waring, R. (1999). *Tasks for assessing receptive and productive second language vocabulary* (PhD Thesis). University of Wales, Swansea, UK.

Waring, R. (2002). Scales of vocabulary knowledge in second language vocabulary assessment. *Kiyo, The Occasional Papers of Notre Dame Seishin University*, *46*(1), 35–41.

Warriner, J. E., & Griffith, F. J. (1977). *Warriner's English grammar and composition, 4th course* (4th ed.). New York, NY: Harcourt Brace Jovanovich.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. New York, NY: Prentice Hall.

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners*. Beijing: Foreign Language Teaching and Research Press.

Wesche, M. B., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review*, *53*, 13–39.

West, M. (1953). *General service list*. London: Longman, Green & Co.

White, E. M. (1984). Holisticism. *College Composition and Communication*, *35*(4), 400–409. https://doi.org/10.2307/357792

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305–319. https://doi.org/10.1177/026553229301000306

Wilkins, D. A. (1972). *Linguistics in language teaching*. London: Edward Arnold.

Wilks, C., Meara, P., & Wolter, B. (2005). A further note on simulating word association behaviour in a second language. *Second Language Research*, *21*(4), 359–372. https://doi.org/10.1191/0267658305sr251oa

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary (Version 2). *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6–11.

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

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning*, *67*(S1), 6–11. https://doi.org/10.1111/lang.12223

Witalisz, E. (2007). Vocabulary assessment in writing: Lexical statistics. In Z. Lengyel & J. Navracsics (Eds.), *Second language lexical processes. Applied linguistic and psycholinguistic perspectives* (pp. 101–116). Clevedon: Multilingual Matters.

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawai'i Second Language Teaching and Curriculum Center.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Xue, G. Y., & Nation, I. S. P. (1984). A University Word List. *Language Learning and Communication*, *3*(2), 215–229.

Yamamot, Y. (2011). Bridging the gap between receptive and productive vocabulary size through extensive reading. *The Reading Matrix*, *11*(3), 226–242.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Index