EUROSLA MONOGRAPHS SERIES 1

COMMUNICATIVE PROFICIENCY AND LINGUISTIC DEVELOPMENT: intersections between SLA and language testing research

EDITED BY

INGE BARTNING
University of Stockholm

MAISA MARTIN
University of Jyväskylä

INEKE VEDDER
University of Amsterdam

EUROPEAN SECOND LANGUAGE ASSOCIATION 2010

Eurosla monographs

Eurosla publishes a monographs series, available open access on the association's website. The series includes both single-author and edited volumes on any aspect of second language acquisition. Manuscripts are evaluated with a double blind peer review system to ensure that they meet the highest qualitative standards.

Editors

ISBN 978-1-4466-6993-8

First published by Eurosla, 2010
Graphic design and layout: Pia 't Lam

An online version of this volume can be downloaded from eurosla.org

Table of contents

Foreword

This book is the first volume of the new EUROSLA Monograph Series. It presents work by the SLATE network (Second Language Acquisition and Testing in Europe, see http://www.slategroup.eu.

The SLATE network shares an interest in combining knowledge of communicative proficiency, as expressed by the can-do scales of the Common European Framework of Reference for Languages (CEFR), with research results with respect to the degree of control of various linguistic features in language productions judged to be at a given CEFR level.

The contributions included in this book all share this common goal. Despite this common goal the studies presented in the book differ from each other in many aspects. There are many target languages (L2; Dutch, English, Finnish, French, Italian, Norwegian, Spanish) and more than 20 source languages (L1). Learners are mostly adults but also young people and children. Learning environments include both formal and informal contexts. Some studies span all CEFR levels from A1 – C2, some concentrate on comparing two adjacent levels. The grammatical descriptions, underlying the definitions of the linguistic features chosen for the studies, comprise both rule-based and usage-based approaches. Parameters of development include vocabulary size, various grammatical features, as well as pragmatic or textual characteristics. Some use the CAF triad (Complexity, Accuracy, Fluency) as the way to track linguistic development, some focus on other areas.

The main focus of the present book, despite of this variety of data and approaches, is on levels of linguistic proficiency as generally understood in testing research, and developmental stages as often posited in SLA research. European languages are in very different positions as to having established stages of second language acquisition. Innumerous studies have addressed the acquisition of English, and there is a fair amount of information on French, while less widely spoken languages like Finnish and Norwegian are only now being examined from this angle. This provides researchers with different starting points, with their advantages and disadvantages. When the overall order of acquisition of various grammatical structures is more or less known, as is the case for English, the study of the development of structures across the CEFR levels can

be based on previous studies. There are thus good reasons to make predictions about which features might prove to be good indicators of development – criterial or diagnostic features, depending on the point of view of the researcher. When there is no such previous knowledge, the choice of the features to be examined must be based on the intuitions of experienced language teachers or on the knowledge of development of other languages, even if these theories have not been tested against the language in question.

In such a vast area of research, only a few issues can be touched upon in one volume. The main foci of the studies described in the various chapters are briefly presented below, to familiarize the reader with what is to follow. The SLATE network and its aims are described in the Introductory Chapter by Charles Alderson (Lancaster University), Jan Hulstijn and Rob Schoonen University of Amsterdam). The introduction is followed by eight chapters presenting results from different SLATE-related projects across Europe. In the final section the first conveners of SLATE, Charles Alderson and Jan Hulstijn, evaluate the SLATE work to date, as presented in this book. All the contributors are briefly introduced in the end of the book.

In the first chapter Riikka Alanen, Ari Huhta and Mirja Tarnanen (University of Jyväskylä) discuss a number of issues relevant to task design and assessment attempting to combine the goals and practices of task-based SLA research, on the one hand, and language testing, on the other. The authors illustrate this by discussing theoretical and methodological decisions underlying the Finnish research project *Cefling,* set up to study the linguistic features of the proficiency levels described by the CEFR scales. In the *Cefling* project, written L2 data (roughly 2400 texts) have been collected from 7^{th} – 9^{th} graders studying Finnish and English as L2 at school. The overall aim of the chapter is to discuss the issues involved in developing tasks and assessment procedures, as well as the construction of L2 corpora that fit the requirements that SLA and language testing research place on language tasks. The chapter also offers an illustrative analysis of task variability performance and a discussion of how quantitative and qualitative analysis of task performance can help researchers to evaluate task design.

Maisa Martin, Sanna Mustonen, Nina Reiman, and Marja Seilonen (University of Jyväskylä) also report results from the Finnish *Cefling* project. The underlying theoretical principle of the *Cefling* project is a usage-based and a cognitively oriented view on language learning. The aim of the chapter is two-fold: to show how the development of particular linguistic structures can be tracked across CEFR levels and to find evidence of potential co-development of different domains. The structures under investigation are locative cases, and transitive and passive constructions in L2 Finnish. Language proficiency was meas-

ured by using three parameters: frequency, accuracy, and distribution across the six CEFR levels. The corpus on which the study is based, which represent the written production of adult learners (informal messages, formal messages and argumentative texts),.was derived from the National Proficiency Certificates exams for L2 Finnish,.

Folkert Kuiken, Ineke Vedder (University of Amsterdam), and Roger Gilabert (University of Barcelona) investigate the relationship between communicative aspects of L2 writing, as defined in the descriptor scales of the CEFR, and the linguistic complexity of L2 performance. The main goal of the *CALC* study (Communicative adequacy and linguistic complexity in L2 writing) is to provide evidence of learner performance, both in functional and linguistic terms (grammar, lexis, accuracy) at a particular CEFR level (A2-B1). A second aim is to contribute to the description of interlanguage by analyzing the use of particular linguistic features that typically characterize L2 performance at a given proficiency level. The investigation was based on a corpus of 200 short essays written by L2 learners of Italian, Dutch, and Spanish. Communicative adequacy was assessed by means of the ratings of individual raters on a Likert scale ranging from 1 to 6. Linguistic complexity was measured both by means of the overall scores of a Likert scale and by means of general measures.

The chapter by Angeliki Salamoura and Nick Saville (University of Cambridge ESOL Examinations) discusses the findings of the *English Profile Programme*. The goal of the project is to provide a set of 'Reference Level Descriptions' (RLD) for each of the six CEFR levels and to identify the so called 'criterial features' of English.. Criterial features, the use of which may vary according to the level achieved, can serve as a basis for the estimation of a learner's proficiency level. The authors illustrate how hypotheses formulated from theories and models of SLA and a corpus-informed approach can be used to investigate L2 learner data in order to develop the RLD's for English. The analyses are based on data from the *Cambridge Learner Corpus* (CLC), containing subsets of examination tests and representing 130 different L1s. In the chapter an overview is given of a number of criterial features of English for the CEFR levels A2-C2 and some preliminary results of verb co-occurrence frames, relative clauses and verbs expressing spatial information are presented.

Similarly to the contributions by Alanen *et al.*, Martin *et al.* and Kuiken *et al.*, the study of Fanny Forsberg and Inge Bartning (University of Stockholm) aims at matching communicative competence as proposed in the CEFR scales with the development of linguistic proficiency. The main goal of the chapter is the investigation of linguistic features of French L2 (CEFR-levels A1-C1), in terms of morpho-syntax, discourse organisation and the use of formulaic sequences. In earlier research on acquisitional orders in oral L2 French the lin-

guistic phenomena under investigation in the chapter discussed in the book were already found to be 'criterial' for French. In the study written data were collected from 42 Swedish university students of L2 French. The first results show that morpho-syntactic accuracy measures yield significant differences between the CEFR-levels up to B2. Also the use of lexical formulaic sequences increases at higher CEFR-levels, but with significant differences only between A2, B2 and C2.

The chapter by Gabriele Pallotti (University of Reggio Emilia) describes a project aimed at 'bringing interlanguage analysis to school'. In the project 10 kindergarten, 7 primary and 2 middle school classes in Italy were involved, including 120 NNS children and 40 NS children, aged between 5 and 13. A number of elicitation procedures were used, comprising film and picture-story retellings, static pictures description and semi-guided interviews. Also teachers were involved in data collection. The chapter shows some examples of interlanguae analysis and focuses upon different stages of the research process, including data collection, transcription, coding, scoring, and quantitative and qualitative analysis. The author concludes that carrying out a systematic interlanguage analysis as it is done in SLA research is highly time consuming and very impractical in most teaching and testing contexts. If CEFR scales are to be related to acquisitional sequences in teaching and testing contexts, it is necessary to find ways in which the latter can be assessed in a reasonable amount of time and without specialized skills, while preserving the procedure's validity with respect to current SLA theorising and methodology.

Cecilie Carlsen (University of Bergen) reports on an empirical investigation of the use of acquisition of cohesion and coherence in written texts, particularly the use of cohesive devices. The study was based on a corpus of Norwegian L2-texts written by immigrant learners of Norwegian with different language backgrounds. A number of 36 different connectives were selected. Following the predictions of the CEFR concerning the increasing range and higher control of cohesive devices across the six levels, the author hypothesizes that although cohesive devices occur at all CEFR levels, texts at higher levels will contain a broader range and a higher degree of control of cohesive devices than lower level texts. A qualitative analysis of the type of cohesive devices which were used showed that additive connectives were employed correctly by almost all the level groups whereas adversative and causal connectives demonstrated a higher error rate.

The main issue discussed in the chapter of James Milton (University of Swansea) concerns the relationship in L2 between vocabulary size and linguistic proficiency in terms of CEFR levels. The study, building on data from English, French and Greek, proposes tools for diagnosing learners' vocabulary

proficiency level in order to examine the number of words that L2 learners at each CEFR level typically know. As an example of a useful tool instrument for testing vocabulary size, the author discusses Meara's XLex test of passive receptive vocabulary. The XLex test estimates the knowledge of the learner of the most frequent 5000 lemmatised words. By means of this test the results of the study show that progressively higher vocabulary scores are generally associated with progressively higher CEFR, despite a certain degree of individual variation

The target of the present volume, set at the SLATE meeting in Jyvaskyla in summer 2009, was both to disseminate research results and to develop future directions for addressing the SLATE research agenda, described in the Introduction Chapter. Although the majority of the studies reported in this book are still going on, the editors of the book firmly believe that publishing these first stimulating results of the work which has already been done will be useful to provoke discussion and to exchange ideas. Let the work go on!

The editors of this book wish to thank all contributors to the book: the authors, the reviewers who commented on earlier drafts of the chapters, Gabriele Pallotti, editor of the Eurosla Monograph Series, Robert Cirillo for proofreading the chapters, Pia 't Lam for graphic design and layout, and Kati Penttinen and Veera Tomperi for helping us with the reference sections and many other details.

July 2010

Inge Bartning, Stockholm
Maisa Martin, Jyväskylä
Ineke Vedder, Amsterdam

Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them?

Jan H. Hulstijn, J. Charles Alderson and Rob Schoonen
University of Amsterdam / Lancaster University / University of Amsterdam

The papers in this volume were written by European researchers loosely organized in the SLATE group: Second Language Acquisition and Testing in Europe (http://www.slate.eu.org/). The group was formed after two meetings, held in 2006, at the University of Amsterdam, organized by Jan Hulstijn, Charles Alderson and Rob Schoonen[1]. This introductory chapter describes the rationale for the SLATE group's ambitions.
The first meeting, sponsored by the European Science Foundation (Ref: EW05-208-SCH) and held on 23-25 February 2006 at the University of Amsterdam, was called *Bridging the gap between research on second-language acquisition and research on language testing*. It brought together 19 researchers, based at eight universities in seven European countries, working in the fields of second language acquisition and language testing. The follow-up meeting, which was sponsored by the Netherlands Organisation for Scientific Research (NWO) (grant 254-70-100) took place 1-2 December in the same year and at the same venue. This meeting, attended by 20 researchers, affiliated to ten universities in eight European countries, was called *Stages of second-language acquisition and the Common European Framework of Reference for Languages*. The titles of these two SLATE meetings characterize the aim and the focus of the SLATE group. After a brief description of the Common European Framework (CEFR), this chapter shows how the SLATE group envisioned research, linking developmental stages in SLA with L2 proficiency levels as defined by the CEFR.

1. The Common European Framework (CEFR)

The Common European Framework of Reference for Languages (Council of Europe, 2001, in English and French; thereafter translations into 31 languages so far, including Arabic, Friulian and Japanese) contains proposals for formulating

1 An earlier meeting had taken place in 2004, also in Amsterdam. Some of the people present at the two SLATE meetings in 2006 had also attended the 2004 meeting.

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 11-20

functional learning targets for language learning, teaching and assessment. Throughout Europe (and beyond), the CEFR has become a major point of reference for language in education, with both the ambition and the potential of bringing common standards and transparency across Europe. It has also become the point of reference for the formulation of objectives of foreign-language learning curricula and the certification of foreign-language proficiency skills of citizens continuing their educational or professional careers in other European countries.

The CEFR presents language and language learning within its social context, sees language users and learners as "social agents", and in general advocates a notional/ functional, "sociolinguistic" approach to language use and therefore to (second/foreign) language development. The framework, distinguishing six "common reference levels" (Chapter 3), adopts a multifaceted approach to the concept of language proficiency. Acquisition of an L2 is said to be a matter of development along what is called a horizontal and a vertical dimension (Council of Europe, 2001, pp. 16–17).

The *horizontal dimension* specifies the "language activities" in which language users engage (pp. 44–57) in terms of (1) context of language use (e.g., in the personal and professional domain), (2) communication themes (e.g., travel and health), and (3) communicative tasks and purposes (e.g., making enquiries from employment agencies and reading safety instructions). Chapter 4 of the CEFR contains 40 scales specifying a large number of forms of oral and written language use (pp. 58–84), including several scales of strategic competence. The horizontal development also comprises dimensions of more general "communicative language competences" (p. 108), subdivided – as in the language proficiency model of Canale and Swain (1980) – in linguistic, sociolinguistic and pragmatic competences. Chapter 5 contains 13 scales for these competences (pp. 110–129).

The *vertical dimension*, outlined in Chapter 3 and applied to all the descriptor scales in Chapters 4 and 5, consists in "an ascending series of common reference levels for describing learner proficiency" (p. 16). The authors caution that "any attempt to establish 'levels' of proficiency is to some extent arbitrary, as it is in any area of knowledge or skill. However, for practical purposes it is useful to set up a scale of defined levels to segment the learning process for the purposes of curriculum design, qualifying examinations, etc." (p. 17).

The development of communicative language competence can thus be seen both at the level of expanding one's range of communicative activities and at the level of performing them in increasingly more complex and sophisticated ways.

It is important to emphasise that the 2001 version of the CEFR itself did not suddenly appear out of nothing. It is the natural result, an organic development, of the work in modern languages of the Council of Europe over three decades, from the initial development of notional-functional syllabuses (Wilkins, 1976,)

and *The Threshold Level* (van Ek, 1975), the latter being widely translated and revised for use in numerous other languages (*Niveau Seuil*, *Kontaktschwelle*, etc.), and then the later elaboration of further levels of the CEFR including a new version of *Threshold* (van Ek & Trim, 1998a), *Breakthrough* (Trim, 2001), *Waystage* (van Ek & Trim, 1998b), *Vantage* (van Ek & Trim, 2001) and the other levels published in the 1980s and 1990s. It can be argued that behind the CEFR of 2001 (and the two draft versions of 1996 and 1998) lies 30 years of experience in developing and implementing curricula, syllabuses and teaching materials at different "levels" of foreign language development. What is, however, in considerable doubt is whether the extensive use of such previous theoretical and practical perspectives was accompanied by empirical research into the constructs and claims of the curricula and their associated notions of level and development. What is perhaps most significant about the CEFR is that, unlike its predecessors, it was accompanied by numerous scales of foreign language development, which were the result of significant empirical research.

Although the bulk of the CEFR is the so-called Descriptive Scheme, which takes up the nine main chapters of the CEFR, most impact, based on the admission of the Council of Europe itself as a result of surveys it has recently conducted (Council of Europe, 2005), seems to be due to the various scales that were developed in parallel to the Descriptive Scheme. This is doubtless in part due to the fact that much of the Descriptive Scheme is couched in rather dense, academic and at times frustratingly opaque terms of little appeal to the likely average readership, whereas the scales are not only shorter and in more accessible language, they are also couched in 'Can-do' terms, since they indicate what learners at given "levels" are believed to be able to do. Indeed, the evidence is that the vast majority of those who claim familiarity with or knowledge of (not the same thing) the CEFR are referring to the scales, or at the very least, to the labels by which the six major levels are identified: A1, A2, B1, B2, C1, C2. In point of fact, as is clearly shown whenever CEFR familiarisation activities are conducted with language educationists (as recommended by the Manual for linking examinations to the CEFR, Council of Europe, 2009), detailed knowledge of even these six levels is frequently defective and uninformed. Given the complexity of the CEFR itself (not to mention the broad field of foreign language learning and use) it is not surprising that even those closely associated with the development of the CEFR and its implementation in various contexts do not claim complete knowledge or understanding of all that it contains.

A total of some 56 scales are contained within the CEFR publication, and these cover the macro skills (called "Communicative Activities" in the 1996/8 version of CEFR) of Reception (both written and spoken), Interaction, and Production (the macro skill/ activity of Mediation is not accompanied by any scales). In addition, scales of Strategies are subdivided into the same three macro

activities/skills, and scales of "Communicative Language Competence" are divided into Pragmatic (with separate scales of spoken fluency, flexibility, coherence and precision) and Linguistic (Range – both general and vocabulary range - and Control - grammatical accuracy, vocabulary control, phonological control and orthographic control).

Although the scales presented in the 2001 publication are in some sense a new compilation, it is important to be aware that their origin is very heterogeneous, and indeed Appendix B to the CEFR details this: a total of 30 sets of scales, from the USA, Canada, Australia, the UK, and Europe, dating from 1974 to 1993, including well-known scales like

- The Foreign Service Institute (FSI) and the Interagency Language Roundtable (ILR) Proficiency Ratings,
- The Australian Second Language Proficiency Ratings,
- The American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines,
- The IELTS Band Descriptors for Speaking and Writing,
- The British Languages Lead Body National Language Standards
- The University of Cambridge/ Royal Society of Arts Certificates in Communicative Skills in English,

as well as lesser known scales like

- The Hebrew Oral Proficiency Rating Grid,
- Goteborg University Oral Assessment Criteria
- Fulcher's Fluency Rating Scale.

Thus, in a very important sense, the CEFR scales themselves are *not* new: they are based upon decades of experience in building, using and, presumably, refining scales in the light of experience.

Secondly, it is also important to remember that the original scales were subjected to extensive analysis and deconstruction into pools of classified descriptors, and critical scrutiny, review, revision, adjustment and empirical testing before they resulted in their current form (see Appendix to 1996 CEFR draft 1996, p. 152ff, and CEFR 1998 draft p. 163ff for details). In practice, well over a thousand descriptors were taken from the source scales, filtered for overlap, clarity and focus, sorted by large groups of language teachers (basing themselves on their experience as both learners and teachers) into the categories the descriptors purported to describe, then critiqued for their clarity, accuracy and relevance and sorted into bands of proficiency. Many of these descriptors were applied to learners in the classes of participating teachers.

This exercise was repeated in a second round, this time with teachers of French and German as well as English, who inspected versions of the scales in

all three languages. The resulting draft scales and descriptors were then subject to Rasch analysis to determine their probabilistic rank order of "difficulty". Those descriptors that 'fit the model' i.e. which could be scaled mathematically, were then calibrated, and those that could not be scaled were rejected (this was particularly the case with draft scales for socio-cultural competence).

The production of the scales was thus an extensive empirical exercise (resulting, in part, in a PhD thesis by North, 1996; see also North & Schneider, 1998). It is fair to say that the resultant scales are probably the best researched scales of foreign language in the world, although perhaps not (yet) the most widely used – that award probably goes to one of the source scales, the FSI, ILR, ACTFL family of scales.

It is important to separate consideration of the scales and the Descriptive Scheme of the CEFR from the use of either. It is clear that the CEFR and especially the scales have had and will continue to have enormous influence. Many ministries of education now require that their national exams be at one or more of the levels of the CEFR. Many institutions (including many universities and examination boards) claim that their exams are at particular levels of the CEFR. However, very few of these have produced any empirical evidence whatsoever that this is indeed the case. Although the Council of Europe has issued a manual for linking exams to the CEFR (Council of Europe, 2009; see also Figueras, North, Takala, van Avermaet, & Verhelst, 2005) and is currently encouraging the development of case studies in the use of the CEFR (Figueras & Noijons, 2009), there is no doubt that most claims of links lack appropriate empirical support. Similarly, many institutions, especially higher education colleges and universities, claim that their curricula are aligned to the CEFR and their graduates are at a given level of the CEFR. Again, there is virtually no empirical evidence that this is the case (see survey by Alderson, 2007a). Finally, some authors have pointed out that the CEFR levels are neither based on empirical evidence taken from L2-learner performance, nor on any theory in the fields of linguistics or verbal communication (Alderson, 2007b; Hulstijn, 2007), while some language testing experts have raised concerns with respect to the suitability of the CEFR for language testing purposes, arguing that the CEFR, in its current form, lacks specificity and consistency in its definitions and is insufficiently supported by empirical research to allow immediate implementation (Weir, 2005; Alderson et al., 2006).

Nevertheless, the scales have been used as the basis of developments in numerous contexts, even before they were published in their 2001 format. The DIALANG project (http://www.dialang.org/; Alderson, 2005; Alderson & Huhta, 2005) was one of the first language testing projects to use the CEFR scales as the basis of the specifications for their tests in 14 European languages, and the project translated these scales from English into the other 13 languages

through a careful quality control process, as well as making statistical comparisons during piloting to ensure the equivalence of the translations (Alderson, 2005; Kaftandjieva & Takala, 2002 report on these equivalences). In addition, since DIALANG contained tests of Grammar and Vocabulary, attempts were made to construct scales for those constructs for all 14 languages.

2. Calls for CEFR related acquisition research

Although there is still much work to be done to validate CEFR scales and their links to curricula and other statements regarding the development of communicative proficiency, we must equally recognise that especially SLA, but language assessment as well, have to date operated with notions of development and levels of development which have all too frequently been hopelessly imprecise. SLA, for example, has frequently simply taken groups of learners at supposedly different levels of ability, conducted cross-sectional research and claimed that the results show development. Yet the levels have been woefully undefined, often crudely labelled "intermediate" or "advanced", or "first and second year university students" – which means little if anything in developmental terms – and which cannot therefore be interpreted in any meaningful way. It is our belief that, whatever its shortcomings, the CEFR has introduced a notion of levels of development that is far better – if only because it can be challenged – than the vague terms (not measures) used to date.

It should also be borne in mind that the CEFR is intended to be relevant to all languages, and is therefore not language specific, which means that most of the descriptor scales (those appearing in Chapter 4) concern the development of language-independent communicative competencies. Chapter 5 contains a few scales on the development of linguistic areas such as phonology, lexicon and grammar, but these are among the most problematic ones. The need for such scales to be language-independent, and thus be applicable to languages as different as Spanish, German and Finnish, makes them appear little more than a list of generic statements about growing accuracy and/or complexity in each linguistic domain.

A number of projects have begun to attempt to identify the features of particular languages at different CEFR levels - Référentiel de Français Langue Étrangère, Profile Deutsch (Glaboniat, Müller, Rusch, Schmitz, & Wertenschlag, 2005), English Profile (http://www.englishprofile.org/), but a number of theoretical and empirical questions remain concerning the componential structure of language proficiency at various CEFR levels. The CEFR authors acknowledge that language users may be placed at different levels of different scales, possessing "uneven profiles": "Progress is not merely a question of moving up a vertical scale" (Council of Europe, 2001: 17). Furthermore, what the CEFR does not

indicate is whether learner performance at the six functional levels as defined in Chapter 4 actually matches the linguistic characteristics defined in Chapter 5, and, more specifically, *which linguistic features (for a given target language) are typical of each of the levels*.

According to the SLATE group, the question of which linguistic-communicative profiles could, and which profiles do, exist constitutes an empirical issue deserving to be investigated properly. Answers to these questions are essential if the CEFR is to be successfully implemented across Europe. Furthermore, as the CEFR is intended to be relevant to all and any language, it does not provide information for specific target languages, much less for specific first-second language combinations. In short, the CEFR is not yet capable of successful implementation. Additional research, linking functional and linguistic information, is urgently needed.

3. SLATE's overarching research question

At the first SLATE meeting in 2006, the following research question was seen as central to a collaborative enterprise: *Which linguistic features of learner performance (for a given target language) are typical at each of the six CEFR levels?* Initial investigation of this question can be conducted in at least two ways:

- Learner performance data that have been elicited and analysed in previous SLA research can be additionally rated with functional rating scales based on the scales presented in Chapter 4 of the CEFR (Council of Europe, 2001).
- Responses of learners who took language proficiency tests related to the CEFR (i.e., data already collected by language testers) can also be rated with functional CEFR-based scales and analysed linguistically.

However, for a more thorough investigation of this research question, new, international, cross-linguistic studies need to be conducted using a *common* research design, tasks, procedures, and analyses.

4. SLATE's more specific research questions and research goals

During the first SLATE meeting, the following points were mentioned as research questions and issues of *potential* relevance.

1. What are the linguistic profiles at every CEFR level for the two productive language skills (speaking and writing) and what are the linguistic features typical of the two receptive skills (listening and reading) at every CEFR level?

2. To what extent do common or different profile features exist across the seven target languages investigated by researchers in the SLATE group (Dutch, English, Finnish, French, German, Italian and Swedish)? Do the profiles differ along language-family lines - Finnish versus the two Romance languages represented in SLATE (French and Italian), and the three Germanic languages (Dutch, English and German)? To what extent do the profiles reflect learners' L1?
3. What are the limits of learners' performance of tasks at each of the CEFR levels? It is important not only to investigate what learners typically *do* at each of the CEFR levels (which requires elicitation of performance in rather "open" task formats), but also what learners *can* and *cannot do* (which requires the administration of tasks of a "closed" format, with and without time pressure, and the measurement of both accuracy and reaction times). In other words, in order to find features typical of a given CEFR level, we need to explore borderline features and to identify features shared by two or more levels and to construct appropriate scales.
4. Which linguistic features, emerging from our profiling research, can serve as successful tools in the diagnosis of learners' proficiency levels and of weaknesses that require additional attention and training? The investigation of this question is of considerable practical importance for all stakeholders (learners, teachers, curriculum and test designers and other language educationists). One desired outcome of investigation into this question would be the development of diagnostic tools as well as information which could be included in learners' language portfolios (Alderson, 2005).
5. Are there commonalities and differences between the linguistic profiles of foreign-language learners (learning the target language in the formal setting of a school curriculum or a language course) and those of second-language learners (learning the target language in the context of everyday use, with or without formal instruction)? This question is not only of theoretical importance but also of practical importance. For example, a feature such as (un)successful use of subject-verb agreement might have a predictive value different for foreign language learners from the value for second language learners.

At the second SLATE meeting (December 2006), additional targets were formulated. Some of these deserve to be reproduced here (quoted from the report of the second meeting):

- A framework for the linguistic analysis of learner performances at different CEFR levels.

- Suggestions for improvements of the CEFR and/or for extending it for use with young learners, L1 speakers, LSP (language for specific purposes) students, and the like.
- New knowledge that will help to design new teaching materials, curricula and diagnostic assessment instruments relating to the linguistic features that characterise different CEFR levels.

The papers in this volume, written by members of the SLATE group, resulted, either directly or indirectly, from the 2006 meetings in Amsterdam and later meetings in Aix-en Provence and Jyväskylä, demonstrating the fruitful nature of the issues raised and discussed.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007a). The challenge of diagnostic testing: Do we know what we are measuring? In J. Fox et al. (Eds.), *Language Testing Reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Alderson, J. C. (2007b). The CEFR and the need for more research. *The Modern Language Journal, 91*, 658–662.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct project. *Language Assessment Quarterly, 3*(1), 3–30.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2005). *Survey on the use of The Common European Framework of Reference for Languages (CEFR): Synthesis of results*. Strasbourg: Council of Europe.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Languages: Learning, teaching, assessment. A Manual*. Strasbourg: Council of Europe.

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: CITO/ EALTA

Figueras, N., North, B., Takala, S., van Avermaet, P., & Verhelst, N. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing, 22*(3), 261–279.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile Deutsch*. Frankfurt: Langenscheidt.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: a validation study. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment. Case studies* (pp. 106–129). Strasbourg: Council of Europe.

North, B. (1996). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement* (Unpublished doctoral dissertation). London: Thames Valley University.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*, 217–262.

Trim, J. L. M. (2001). *Breakthrough*. Unpublished manuscript.

Van Ek, J. A. (1975). *The Threshold Level*. Strasbourg: The Council of Europe.

Van Ek, J. A., & Trim, J. L. M. (1998a). *Threshold 1990*. Cambridge: Cambridge University Press. (Original work published 1991).

Van Ek, J. A., & Trim, J. L. M. (1998b). *Waystage 1990*. Cambridge: Cambridge University Press. (Original work published 1991).

Van Ek, J. A. & Trim, J. L. M. (2001). *Vantage*. Cambridge: Cambridge University Press.

Weir, C.J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 301–320.

Wilkins, D.A. (1976). *Notional syllabuses*. Oxford: Oxford University Press.

Designing and assessing L2 writing tasks across CEFR proficiency levels

Riikka Alanen, Ari Huhta and Mirja Tarnanen
University of Jyväskylä

With the advent of the Common European Framework of Reference (CEFR) for the learning, teaching and assessment of modern languages, there have been renewed calls for the integration of the research perspectives of language testing and second language acquisition across Europe. The project Cefling was set up in 2006 with this purpose in mind. In the project our aim is to describe the features of language that L2 learners use at various levels of language proficiency defined by the CEFR scales. For this purpose, L2 Finnish and L2 English data were collected from young and adult L2 learners by using a set of communicative L2 writing tasks. In the course of the project, the different understandings of what the purpose of an L2 writing task is needed to be reconciled not only in the minds of researchers but also in research design. In what follows, we will discuss the issues involved in designing and assessing L2 tasks for SLA and language testing purposes by using the design and assessment procedures in the project as a case in point. We will also present some of our findings to illustrate how statistical procedures such as multifaceted Rasch analysis can be used to examine task difficulty.

1. Introduction

Until quite recently, there have been relatively few empirical studies combining the research perspectives of language testing and second language acquisition. Beginning in the 1990s (see e.g. Bachman & Cohen, 1998), the number of such studies has steadily grown although it has remained fairly small, in particular in task-based research. With the advent of Common European Framework of Reference, CEFR, (Council of Europe, 2001) for learning, teaching and assessment of modern languages, there has been an increasing interest in setting up studies combining both research perspectives across Europe. Integrating the two research perspectives is not without difficulty, however; almost inevitably, compromises must be made. In this chapter, we will discuss a number of issues relevant to task design and assessment in research approaches attempting to combine the goals and practices of SLA research in task-based research, on one hand, and language testing, on the other hand, by using the theoretical and

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 21-56

methodological decisions taken in one such research project as an illustration. The project in question is called *Cefling – The linguistic basis of the Common European Framework levels: Combining second language acquisition and language testing research*; it is a project set up to study the linguistic features of the proficiency levels described by the CEFR scales (see Martin, Mustonen, Reiman, & Seilonen, this volume). The chapter ends with an illustrative analysis of task variability learner performance, and a discussion of how quantitative and qualitative analysis of task performance can help researchers to evaluate task design.

SLA research is, of course, primarily interested in the development of L2 proficiency, complexity, accuracy and fluency, in particular. Language testing, on the other hand, is occupied with the development of reliable and valid measures for assessing communicative language ability or language proficiency. It is primarily interested in how successful the items used in language testing are, and, depending on the type and goals of the language test, also in the communicative adequacy of tasks (see Pallotti, 2009).

Task emerges as a key notion linking both SLA research and language testing practice. It is a key unit both in L2 data elicitation and measurement. In SLA and language teaching, task is regarded as a programmatic or even curricular unit, a type of meaning-based activity L2 teaching should be focused on or organized around (see e.g. Bygate, Skehan, & Swain, 2001; Ellis, 2003; Samuda & Bygate, 2008; Van den Branden, 2006; Van den Branden, Bygate, & Norris, 2009). It is also a key unit in performance based assessment of L2 proficiency. As Brindley (1994/2009) defines it, task-based language assessment (or task-centered as it was called then) is

> the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge. (p. 437)

In task-based assessment, task performance can be assessed according to its communicative adequacy, i.e., on how well the learner is able to use language to accomplish task requirements. Communicative adequacy is commonly evaluated by rating scales; yet, as Pallotti (2009) notes, there are surprisingly few studies attempting to look at complexity, accuracy and fluency in terms of such scales in task-based SLA research. De Jong, Steinel, Florijn, Schoonen, & Hulstijn (2007) and a number of other studies in this volume (e.g. Gilabert, Kuiken, & Vedder; Martin et al., both this volume) are among the first to approach the issue from this particular perspective.

The notions that are of particular importance for this chapter are *task*, and how *L2 performance* on a particular task are perceived and operationalized in

SLA research and language testing. In a number of ways, some of the issues fundamental to task-based language assessment are central to this chapter, as well. To slightly modify the list originally devised by Norris (2002, p. 337), in this chapter we will discuss the following issues: 1) Why are participants asked to perform communicative L2 writing tasks in the first place? 2) What exactly do researchers want to know on the basis of their task performances? 3) How can tasks be selected or designed, and performances judged, so that researchers can know these things? and 4) What are going to be done with judgments of participants' task performances, once they have been elicited?

In what follows, we will first briefly describe the characteristics of language testing practice and language testing research relevant for SLA research. It is useful to be aware of what sensible language testing *practice* entails because that is crucial for the quality of whatever data elicitation and collection instruments an SLA researcher is using. Awareness of what goes on in language testing *research*, for its part, is also useful for ensuring the quality of the instruments but here the contribution of testing relates more to the conceptual level of the entire measurement process (cf. Norris & Ortega, 2003, p. 720). In the latter part of the chapter, we will discuss issues relevant to task design and assessment, and finally, show, by using the judgment data from the Cefling project as an illustration, what can be done to analyze and evaluate the participants' task performance.

2. SLA research and language testing: goals, purposes and practices

L2 development is a complex, dynamic process; however, it is mostly investigated through L2 products, slices of L2 performances elicited at certain points of time under a set of specific circumstances. There are a number of aspects in L2 development that have been the focus of study over the years, including particular linguistic structures (e.g. negation, question formation, tense and aspect). Increasingly, since the 1990s, the notions of complexity, accuracy and fluency have come to be used to define and describe L2 performance and L2 proficiency, or CAF for short, from the SLA perspective (see the special issue of *Applied Linguistics* on CAF in SLA research edited by Housen and Kuiken, 2009).

Compared with SLA research, language testing has a large number of very different purposes, and a number of different decisions can be made on the basis of the results of assessments. It is beyond the scope of this article to give a detailed account of language testing but it is useful to be aware of some of its main purposes and how they might relate to SLA.

The key question that determines the quality and trustworthiness of language assessment instruments is the degree to which testing (or assessment more generally) adheres to professional guidelines on good practice that have been

developed both for measurement in general – for example *The standards for educational and psychological testing* (American Educational Research Association, 1999) – or language assessment – e.g., *EALTA Guidelines for good practice in language testing and assessment* (European Association for Language Testing and Assessment, 2006) or *ILTA Guidelines for practice* (International Language Testing Association, 2007). Such guidelines strive to ensure that assessment is carried out reliably, validly and fairly, without which test results would be meaningless. In the context of SLA research, poorly designed data collection instruments would cause findings to lack interpretability and generalizability (Norris & Ortega, 2003, p. 717). Thus, following the principles of test design and validation developed in such fields as language testing, educational measurement and psychological testing will help SLA researchers make sure that they can depend on the data that they gather with their instruments.

There is no clear-cut way of categorizing purposes and types of decisions made on the basis of assessment but one simple division can be made between assessment related to specific language courses or curricula and proficiency testing that is detached from any teaching (see e.g. Bloom, Hastings, & Madaus, 1971; Huhta, 2008; Millman & Greene, 1993; Weigle, 2002). The latter is often carried out by examination organizations that certify learners' level of proficiency (e.g., Educational Testing Service, Cambridge ESOL, and the Goethe Institut are examples of such organizations). One specific type of proficiency testing relates to research in applied linguistics. Here, the aim of assessment is to enable applied linguists to gather information about learners' language skills as reliably and validly as possible (Huhta & Takala, 1999). As Douglas (1998) notes, when a language test elicits linguistic features of performance, it functions as an "SLA elicitation device" (p. 141). In an early discussion of the issues connecting language testing and SLA research, Byrnes (1987) points out how "data from proficiency testing should be able to provide information about the interrelationship between posited developmental stages and variational features, particularly in instructed SLA" (p. 48).

There are two main strands of research combining the SLA and language testing perspective (see also Hulstijn, Schoonen, & Alderson, this volume). On the one hand, researchers can collect L2 performance data from existing language tests and examination systems and analyze them for a number of linguistic features (e.g. Banerjee, Franceschina, & Smith, 2004; Norris, 1996, as cited in Norris & Ortega, 2009; Salamoura & Saville, this volume). On the other hand, they may choose a task-based approach and design a set of communicative tasks and rate them for communicative adequacy and at the same time analyze the linguistic features of learner performance. In what follows, we will discuss task-based approaches to SLA research and language testing by taking the solutions developed in the Cefling project as a case in point.

3. Tasks in SLA research and language testing

One of the key choices that needs to be made in any research attempting to combine both SLA and language assessment perspectives concerns the elicitation and measurement of L2 performance. In SLA research, performance data have always been collected in naturalistic conditions, that is, in the context of real language use. Alongside this strand of research, there is also a strong tradition of research relying on analytic tests and discrete point data collection instruments such as structured exercises and completion tasks (see e.g. Hulstijn, 1997). Beginning in the mid-1990s, a new, task-based research tradition investigating the multiplicity of cognitive and interactive factors on task performance (see e.g. Bygate, Skehan, & Swain, 2001; Robinson, 2001; Skehan, 1998; Skehan & Foster, 1997) began to emerge. This line of research has produced a number of studies and hypotheses about the influence of features such as task complexity or task difficulty on L2 development.

Task has been defined in a number of ways. In this chapter, it is defined as "an activity which requires learners to use language, with emphasis on meaning, to attain an objective" (Bygate, Skehan, & Swain, 2001, p. 11). A task typically involves holistic language use: "through engaging with the task, learners are led to work with and integrate the different aspects of language for a larger purpose" (Samuda & Bygate, 2008, p. 8). The learner's performance on the task can be assessed by focusing on specific features (such as grammatical accuracy or lexical complexity, or fluency or any number of features specified in the ALTE evaluation grids, for example). However, an essential feature of task is that it has a goal and an outcome: there is an objective that learners have to complete, and to do that, they have to use language (cf. Brindley, 1994/2009). How successful and efficient learners are in achieving the task's goal is called communicative adequacy: Pallotti (2009) notes that communicative adequacy is a dimension that most task-based studies in SLA have rarely looked at. For a learner to achieve the purpose of the task, i.e., to be communicatively adequate, it is not necessary for him or her to use correct, target-like language (Skehan, 2001, p. 167).

As Pallotti (2009, p. 597) goes on to point out, in open tasks, adequacy can be assessed by using qualitative ratings such as the CEFR scales. This is an approach adopted by the Cefling project (see also Gilabert et al., this volume). In the Cefling project, a key role of language testers was to ensure that the tasks and language proficiency ratings needed in the project were designed according to good language testing practice and that the quality of the ratings and data collection instruments was empirically ascertained.

There are, of course, all kinds of language tests, ranging from discrete point multiple choice tests to tests imitating real life communicative situations, and

their emphasis is not always only on meaning or even carrying out a particular communicative task. However, one of the central aims for language testing practice and research has been the development of tests and test items that measure language proficiency or communicative language ability as reliably and validly as possible (see e.g. Bachman, 1990; Bachman & Palmer, 1996).

The nature of language abilities has received considerable attention over the decades and is one of the main contributions of language testing research to applied linguistics (see e.g. Bachman & Cohen, 1998). Whether such abilities-oriented approaches to L2 proficiency can be used successfully to predict real-life task performances is another matter: scholars like Skehan (2001), for example, remain rather skeptical of whether the "codifying nature of the underlying competence-oriented models" (p. 167) can be used to make such predictions and have preferred to create alternative models of test performance (see also McNamara, 1996). On the other hand, language testers are well aware of the effect that different contextual factors, test taking strategies, test-taking processes and characteristics of the tasks and measurement instruments in general have on test performance, and studies focusing on such features are considered important in both language testing and SLA research. Yet, one might argue that one of the key purposes of most language testing research – when it relates to large-scale, high-stakes tests in particular – is designing tests and tasks that are as impervious as possible to such contextual factors; after all, the aim of such tests is to be able to generalize the test performance to other contexts.

Such motives may also reflect on the way L2 proficiency and L2 performance are conceptualized in language testing: preferably, both should be as stable as possible because in that way their measurement is easier, more reliable, generalizable and valid. Yet, from an SLA perspective, tasks need to be such that they elicit L2 performance that is variable enough. In usage-based approaches to SLA, in particular, L2 proficiency and L2 performance are considered inherently dynamic (e.g. de Bot, Lowie & Verspoor, 2005; Larsen-Freeman, 2002, 2009). The degree to which L2 performance is regarded as relatively stable and/or more or less systematically influenced by task structure or cognitive features varies from approach to approach. In the case of the Cefling project, what both SLA and language testing researchers share is a common understanding of L2 proficiency as something based on (or even having its origins in) communicative L2 use. Evaluating learner performance on communicative L2 tasks emerges as the common factor that both SLA research and language testing share an interest in.

Different types of tasks elicit different L2 performance. From the SLA perspective, to rely on just one kind of task in L2 data elicitation may lead into serious error, or at least yield only incomplete findings on the nature of L2 development, whether it is CAF, DEMfad (Martin et al., this volume), or a particu-

lar linguistic structure that is the focus of research. Similarly, from a language testing perspective, a range of different tasks should be used if the researcher wants to obtain a generalizable picture of learners' (writing or other) skills: there should be sufficient and valid evidence about learners' proficiency unless one is interested only in learners' ability to do one or two particular tasks (see e.g. Norris, Brown, Hudson, & Bonk, 2002). Finally, after L2 performances have been collected, the tasks should also be scrutinized to check whether they were functioning the way they were supposed to – in terms of their difficulty or complexity, for example – or whether they elicited the type of language that was expected.

In what follows, we will highlight some of the issues to be considered before and after L2 data elicitation when designing tasks for both SLA and language testing purposes by using data from Cefling as a demonstration. We will pay particular attention to the issues related to task design and assessment.

4. L2 writing proficiency from the SLA and language testing perspectives

Research on the development of L2 writing proficiency – L2 writing is used here as a cover term for both second and foreign language writing – has a fairly long history (see e.g. Grabe, 2001; Matsuda, 2005). A number of measures have been developed to capture various aspects of L2 development in writing, including complexity, accuracy and fluency. Wolfe-Quintero, Inagaki, & Kim (1998) reviewed a number of constructs and measures used to operationalize them. Recently, Norris and Ortega (2009) closely analyzed the notion of (linguistic) complexity and its measurement, taking a critical view of some of the suggestions made by Wolfe-Quintero et al. (see also Ortega, 2003; Pallotti, 2009). In sum, it appears that researchers are gradually beginning to take into account the multidimensional and multicomponential nature of L2 proficiency and development both in designing research as well as interpreting the findings: the constructs and measures which seem particularly adapted for capturing the growth of proficiency across, for example, the beginning stages of L2 writing proficiency, may not be as suitable for the more advanced levels of writing, or for L2 speaking, for that matter, or vice versa. Language testing has, of course, for a long time looked at L2 proficiency as multicomponential, although this conceptualization has been used more for improving test construction than for understanding L2 development.

There are a number of studies that have looked at L2 writing proficiency by examining data from the already existing language tests such as IELTS or Cambridge ESOL examinations (see e.g. Banerjee et al., 2004; Salamoura & Saville, this volume). However, not much attention has been paid to the effect

of task type, task complexity or task structure on learner performance in studies of task-based L2 writing from the SLA perspective, in contrast to L2 speaking. In one of the first studies of this kind, Kuiken and Vedder (2008) examined the effect of task complexity on syntactic complexity, lexical variation, and accuracy in the written productions of low-proficiency and high-proficiency L2 Italian and L2 French learners. Students' L2 proficiency was assessed by using a separate cloze-test; as a task, students had to write a letter to a friend helping them choose a holiday destination. No indication of significant interaction between task complexity and lexical variation and syntactic complexity was found in the study; however, an increase in task complexity led learners to produce a text which was more accurate. As Kuiken and Vedder (2008) conclude, this finding can be interpreted as a function of an increased control of the L2 system that a more complex task may require from learners rather than support for any of the existing models of task performance.

In fact, as Kuiken and Vedder (2008) note, the relationship between task type or task complexity and L2 writing performance is not at all clear: as an example, they point to a study by Hamp-Lyons and Mathias (1994) showing that, contrary to expectations, students' L2 performance was lower on personal and expository writing tasks, which were judged as easier by experts, and better on argumentative and public tasks, which were rated more difficult.

Various types of L2 writing tasks have been used in research as data collection instruments. However, it appears that at least in the U.S., there is a difference between the types of writing tasks most commonly used in foreign language and ESL writing classes. In her review, Reichelt (1999) notes that in the former, the focus is typically on creative and expressive, non-academic writing while the latter has tended to involve essays or compositions or other argumentative or descriptive texts commonly used in academic contexts. That L2 writing tasks of personal or expressive nature – essays on hobbies, family life, friends, holidays, and personal letters – are preferred in foreign language classrooms is probably true for other countries as well.

Some of the studies on L2 writing development include studies which have an explicit link to language testing (whether any of these studies can be regarded as task-based in the sense that today's research uses the term is unclear). Valdés, Haro, & Echevarriarza (1992) analyzed the ACTFL scale for writing in detail and then examined short essays written by novice, intermediate and advanced L2 Spanish learners attending a college-level language program. Their findings suggested that the students' L2 proficiency interacted with their L1 writing skills so that more proficient L2 learners were able to write more competent and coherent essays. Henry (1996) investigated the early L2 Russian writing development of novice and intermediate L2 learners by analyzing their essays for fluency, syntactic fluency and accuracy and by contrasting them to the

ACTFL writing proficiency descriptors. For various reasons, neither study used the ACTFL proficiency scales to rate the students' texts; instead, either years of study (Valdés et al., 1992) or a type of global assessment (Henry, 1996) was used to determine the writers' proficiency level. A separate attempt was made by Henry (1996), however, to evaluate learner performances as to their communicative adequacy by asking the judges to decide whether the essays would be understandable (p. 315).

As Skehan (2001) notes, in task-based SLA research using rating scale measures is not typical; rather, researchers have tended to use various operationalizations of constructs such as CAF. Similarly, Pallotti (2009) notes that there are few studies using qualitative ratings to evaluate the communicative adequacy of tasks. Studies reported by Wigglesworth (1997, 2001) are an early exception: Wigglesworth studied variability in L2 speaking performances across five different tasks. Tasks targeted at different proficiency levels were rated with a rating scale used to assess the L2 English proficiency of adult immigrants to Australia. Wigglesworth examined task variability by looking at learners' performance on tasks and their evaluations of task difficulty, and conducted multifaceted Rasch analyses on the data by using the statistical modeling program FACETS (e.g., Linacre, 2010) (see also McNamara, 1996). Her findings reveal a complex interaction of a number of factors such as task structure and task conditions (interlocutors' actions and familiarity with the topic).

In sum, there are few studies so far attempting to use specific proficiency scales to assess the level of task-based L2 writing performances and then utilizing those judgments as independent variables in order to determine particular and general linguistic features typical for those levels. In their review of prototype task-based performance tests called ALP (see Norris, Brown, Hudson, & Yoshioka, 1998), Norris et al. (2002) found, among other things, that careful simulations of L2 communication tasks could effectively elicit a wide range of L2 performances, and that average performance patterns based on the rating scales reflected expected differences among the ability levels of participants. Findings such as these support the idea that ratings of learner performances on a set of communicative tasks can be used as an indication of learners' ability to accomplish such tasks, and that task-independent ratings can be used as an indication of learners' general abilities in performing the range of test tasks (Norris et al., 2002, p. 415).

In the Cefling project (see Martin et al., this volume), L2 Finnish and L2 English data were collected from young and adult L2 learners by using a set of communicative L2 writing tasks. Learner performances were rated by using two scales, the CEFR (young learners) and the National Certificates (adult learners) examination scales for writing, and the Finnish *National Core Curriculum for Basic Education* (2004) scales (young learners). The data collected in the project

were used to build an L2 Finnish and L2 English learner corpus for the analysis of linguistic features (within the limits presented by the data set). In the future – work on corpora is still in progress — the data collected in the project will give researchers a chance to look at not only the linguistic features of the CEFR levels but also shed light on the linguistic basis of the ratings.[1]

5. Designing and selecting communicative L2 writing tasks

Designing and selecting tasks that are relevant for both SLA and language testing research is particularly challenging. Tasks should elicit L2 data that can be analyzed for differences in linguistic features, while it should also be possible to rate task performances based on the communicative adequacy of those performances. The latter dimension imposes conditions of its own on the type of tasks that can be used in data elicitation: from the outset, learners' level of L2 proficiency either constrains or supports their ability to carry out L2 tasks successfully. In language testing, this has been taken into account by using different tests and tasks for e.g. beginning, intermediate and advanced learners.

In research combining both SLA and language testing perspectives, there are a number of solutions to this problem: for example, one can simply ask all learners, regardless of their age or proficiency level, to do all types of task, or, one can try to match tasks with the test taker's ability. In Cefling, an attempt was made to combine both approaches: all learners were asked to do a set of four different tasks (with one of the tasks having an alternate version); at the same time, the type of tasks that the participants were most likely to have encountered and the level of their L2 proficiency was carefully estimated in advance so as to make the tasks as suitable for them as possible.

Since our intention in the project was to collect data for research purposes from learners from all proficiency levels, from both adult and young learners, our task was challenging indeed. However, what both helped and constrained us in the design and selection of L2 tasks was the existence of a data set collected from adult L2 learners available from the National Certificate (NC) language examination system. The NC is based at the Centre for Applied Language

1 It is important to note that initially, only those learner performances were included in the corpora that received CEFR ratings that were sufficiently reliable (this was done because the ratings were used as an independent category variable in the study). Tracking the performances of individual language learners across different tasks, no matter how they were rated, will be possible, however, at later stages of research.

Studies at Jyväskylä; it has stored L2 data in digital format from test takers in several languages in its data base since 2004. The data base is available for researchers through a web portal at the Finnish Social Science Data Archive (http://www.fsd.uta.fi/english/index.html). The tasks used to collect these data served as a starting point – whatever data were going to be collected from teenaged language learners had to match the existing data base as to the type and nature of the task.

A considerable amount of time and effort went into designing or selecting tasks that would reflect a variety of features relevant for the development of communicative L2 writing. It was also necessary to take into account the scales that were going to be used for L2 writing assessment: the CEFR and the Finnish National Core Curriculum scales. The NC tasks were designed for adults, and since cognitive, interactive and learner factors are influenced by age and experience gained through activity in a wide variety of social contexts (see Vähäpassi, 1982; Weigle, 2002), they were not necessarily suitable for younger L2 learners writing in school context.

The issues that needed to be taken into account included the proficiency level the task was aimed at, the topic and domain of the tasks, as well as the genre and functions of the language we expected the tasks to generate. In many ways, the decisions made in the project concerning the nature and type of tasks reflect a striving for communicative authenticity and adequacy of tasks; yet, an attempt was also made to make sure that tasks would elicit particular linguistic structures (e.g. locative expressions, verb forms, relative clauses, questions, negation).

It was also felt that the tasks should be communicative and that they should have some measure of authenticity in terms of text types and processes needed in completing the tasks. In Finland, it seems that users mostly engage in communicative L2 writing outside the classroom (e.g. on the Internet) (Luukka et al., 2008).

Based on such considerations, a set of tasks was designed and extensively piloted by administering tasks to 7th graders in a number of schools. Piloting the tasks was an essential part of the process and yielded much valuable information. The final set of tasks consisted of five communicative tasks representing a variety of text types, functions and register; most tasks belonged to the personal domain. Task 1 was an email message to a friend, Task 2 was an email message to a teacher, Task 3 was a complaint to an internet store, Task 4 was an opinion piece, and Task 5 was a story (see Table 1 for the features of the tasks and Appendix 1 for the tasks themselves). For logistical reasons, Tasks 1 and 2 were alternates: it was felt that both teachers and students were easily able to fit four tasks in their lessons during one term but no more than that. In the end, the participants did either Tasks 1, 3-5 or Tasks 2-5.

One last issue we want to raise has to do with the nature of prompts used in data elicitation. Quite often, prompts for L2 writing tasks of this type include target language data: the instructions may be in English, or the task includes newspaper articles or letters to the editor or other such items in the target language. However, from an SLA perspective, in order to obtain a reliable description of what the linguistic repertoire of learners from each proficiency level is, it may be better to exclude such prompts (see e.g. Grant & Ginther, 2000). On the one hand, it is difficult to say what effect, if any, such recycled fragments of L2 might have in the statistical analysis of data; in the worst case, it could potentially seriously distort the analysis of such dimensions of L2 performance as CAF. Be that as it may, at least for the L2 English tasks, a decision was made to ensure that the task prompts contained as little L2 input as possible.

Table 1. The domain and register and the functions and linguistic structures that the tasks in Cefling were expected to elicit.

Tasks	Domain, register and functions
Task 1 Email message to a friend	• *Domain and register*: personal life, informal • *Functions*: apologizing, argumentation or expressing obligation/necessity (answering to *why* question, negation), requesting, giving information • *Linguistic structures*: questions, negation, tense, locative expressions
Task 2 Email message to a teacher	• *Domain and register*: personal life, school, informal or formal • *Functions*: argumentation or expressing obligation/necessity, asking for information • *Linguistic structures*: questions, negation, tense, locative expressions
Task 3 Email message to an internet store	• *Domain and register*: personal, public, formal • *Functions*: introducing oneself, complaining, requesting correction, suggesting solution • *Linguistic structures*: questions, negation, aspect
Task 4 Opinion	• *Domain and register*: personal, everyday life, school, informal or formal • *Functions*: expressing an opinion, arguing for or against • *Linguistic structures*: tense, aspect, locative expressions
Task 5 Story	• *Domain and register*: personal life, informal • *Functions:* describing and narrating, argumentation or expressing an opinion, liking or dislike • *Linguistic structures:* tense, agreement

6. Designing the assessment procedure

In much of language assessment, the focus is very much on finding out what L2 learners can and cannot do communicatively, i.e., their ability to do things with the language. The main interest in many types of language assessment is to place learners at certain levels of L2 proficiency. In the kind of tasks used in the project, this always involves human raters since they are the ones who decide – based on a set of descriptors – how well L2 learners succeed in completing the tasks (see Weigle, 2002). Such a ratings process is always subjective in the sense that it depends on how raters understand, interpret, and put into practice the scales and their descriptors. Usually, but not always, raters are also given a set of benchmarks, a set of performances that serves as prototypical examples of specific performance levels.

The selection of scales and training raters in how to use them is crucial for a research project using the proficiency levels as independent variables, in particular. Two issues about rating scales should be mentioned at this point. In our view, none of the CEFR scales is a proper rating scale comparable to most scales specifically designed for rating purposes (see Alderson, 1991). Using CEFR scales for rating is therefore a challenging exercise and it is in principle uncertain to what extent particular CEFR scales actually enable reliable rating to take place even though the careful design of these scales gives cause for some optimism (North, 1996/2000). In comparison, the Finnish curriculum scale for writing appears more 'rater friendly', with references to linguistic features and to deficiencies in learners' performance. However, no published research appears to exist yet on how well the CEFR writing scales or the Finnish curriculum scales actually work for rating purposes. One of the aims of the Cefling project was to investigate how the scales function as an evaluation tool of learner performances.

The second issue is the effect on the ratings of the linguistic features of the rated performances. On what features do raters base their ratings? Paying too much attention to linguistic features could introduce circularity in the reasoning: proficiency levels are determined on the basis of linguistic features, and these features are, in their turn, used in defining the levels. The choice of use-oriented CEFR scales was intended to minimize this danger: they focus on communication with practically no references to specific linguistic features. The use of several raters instead of just one may also address this problem, as it is likely to reduce the effect on the ratings of very linguistically-oriented raters. While it makes sense to try to minimize linguistically-influenced rating of performances when the aim is to place learners on proficiency levels on the basis of their ability to use language, we believe it is ultimately impossible to totally remove the effect of linguistic features from any rating of language performances.

Two proficiency scales – the CEFR scale and the Finnish National Core Curriculum scale – were selected for placing learners' performances in the Cefling study. The CEFR scale used in the Cefling project is a combination of six CEFR scales for writing (see Appendices 1 and 2). The Finnish National Core Curriculum (NCC) scale is an adaptation of the CEFR scale for the purpose of defining targets for learning, teaching and assessment in the foreign language curricula for primary and secondary education in Finland (see Appendix 3). It is the official reference scale that teachers in the Finnish schools should apply when assessing their pupils' foreign and second language proficiency. There are no language-specific versions of the NCC scale but the same scale is used for target-setting and assessment in all foreign and second languages covered by the national curriculum. As regards content, the NCC differs from the CEFR scales in that it has no genre-specific level descriptors for different text types. Importantly, the level descriptors of the NCC scale make explicit references to general linguistic characteristics such as accuracy and complexity, vocabulary and structures, while the CEFR scales do not (see Hildén & Takala, 2007).

Good rating scales are necessary but not sufficient for ensuring reliable and valid rating of learner performances. Design of the entire rating process, training of the raters and selection of benchmark examples are also important. There are a number of recommendations in language testing literature as to how such a process should unfold (e.g. Alderson, Clapham, & Wall, 1995). Of course, it also helps to have experience in organizing large-scale rater training, and to have raters who have previous rating experience in using similar scales. For example, the training of raters in Cefling was a multi-stage process that consisted of one full session, a brief update meeting and self-study. The training process was also used to create the final, official benchmarks that were used in the final phase of task assessment.

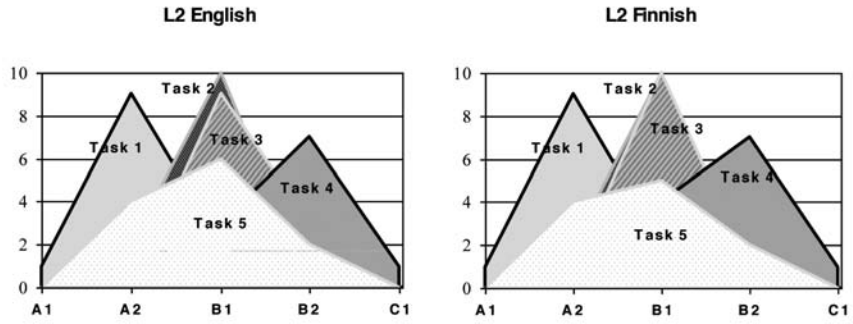
7. Evaluating task difficulty based on rating data

In this section, we will examine the tasks in more detail, focusing on the 7th-9th graders' actual performances on L2 English tasks in terms of the ratings they received on the tasks. The final data collection began in the autumn term of 2007 and lasted well into the spring term of 2008. A total number of 3427 L2 Finnish and L2 English performances were collected from 7th to 9th graders from schools. A number of scripts (N = 1789) were selected for assessment by a team of trained raters (N= 9 for English, N = 11 for Finnish). Because of the large number of texts, rating was divided among the team members so that each script was rated by four (L2 English) or three (L2 Finnish) raters. Incomplete

ratings were not considered a major problem since multifaceted Rasch analyses (see below) are capable of handling incomplete rating designs as long as there are enough links between raters and tasks in the design. The rating of the performances was based on the learners' original handwritten scripts, which were photocopied and delivered to the raters together with the instructions, scales, and the benchmarks.

To begin with, the raters' perception of tasks was placed under scrutiny. Before the actual rating of learner performances, the suitability of the tasks for L2 learners at various CEFR and NCC levels was assessed by 12 of the raters. For each L2 Finnish and L2 English task, the raters were asked to indicate the level they thought the task would be most and second-most suitable for, as well as the lowest and highest proficiency level that the task could be used for. Figure 1 shows the level the tasks were best suited for in both L2 English and Finnish according to the raters. The frequency distributions are very similar for both languages. The raters (N=12) rated Task 4 as most suitable for B2, while Task 1 was considered as the easiest for both languages – most raters considered it as the most suitable for A2. Tasks 2, 3 and 5 were considered best suited for B1, with Task 5 more skewed towards A2.

Figure 1. Perceived task difficulty by raters (n=12) for L2 English and Finnish across all tasks: the level the task is most suited for.



In evaluating the interaction of tasks and judgements of learners' proficiency level, various statistical analyses were used. Figure 2 shows how the L2 English and L2 Finnish task performances were rated. The box plots show the arithmetic means of ratings of L2 English and L2 Finnish performances and their variance across the five tasks.

Figure 2. The ratings given for L2 English (on the left) and L2 Finnish (on the right) performances across the five tasks (T1-T5); dots represent outliers in the data set.

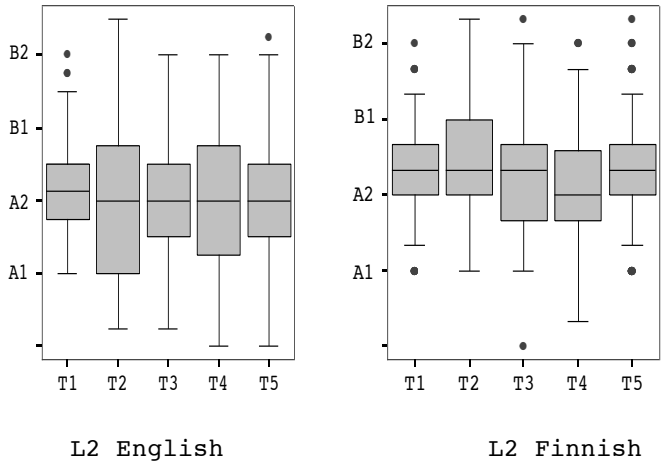
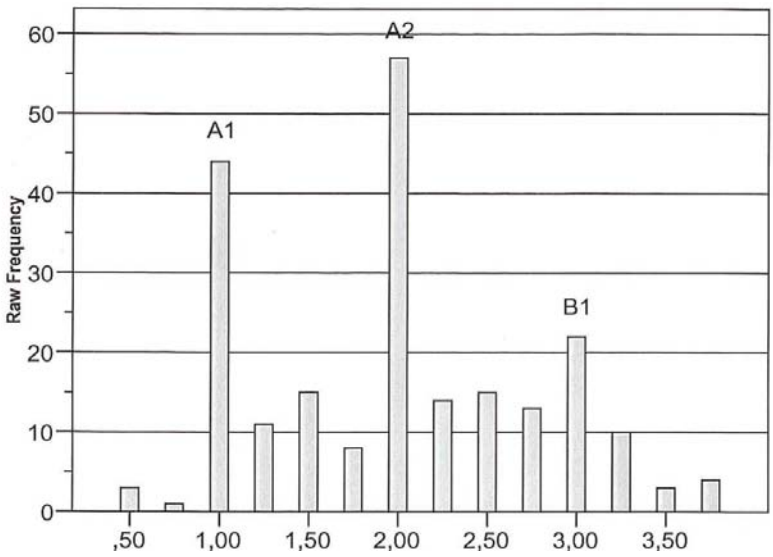


Figure 3. The median of median ratings on Tasks 1 – 5 in L2 English.



Looking at these results alone, it appears that the level of task performances in both L2 English and L2 Finnish was most often A2, with L2 Finnish learners receiving slightly higher ratings apart from Task 4. The distribution of the medi-

an ratings for performances in L2 English support this (see Figure 3). A2 is the most frequent median rating, while A1 appears to form a second peak in the distribution.

Figure 2 also reveals that there was greater variance among L2 English learner performances in terms of the ratings they received than among L2 Finnish learners. To gain a more in-depth understanding of task variability, multifaceted Rasch analyses were also calculated for the L2 performances by using FACETS (Linacre, 2010). A Rasch analysis takes into account a number of facets, or elements, in test performance, to check how the test taker's ability – in this case L2 learners' proficiency – interacts with other factors such the raters' relative severity or leniency, which makes it possible to analyze task difficulty from multiple perspectives. In language testing research, this method has been used, for example, to study different types of performance tasks (McNamara, 1996), the rating of L2 oral discussion tasks (Bonk & Ockey, 2003), or the interaction of teacher, peer- and self-assessment on the rating of EFL writing tasks (Matsuno, 2009). Wigglesworth (2001) used it to analyze task variability in oral L2 performance data.

The facets included in the analysis were the ratings for L2 learners, the raters, and the tasks. Figures 4 and 5 show the order of task difficulty in L2 English and L2 Finnish, from the easiest down. Zero stands for the centre of the range of item (task) difficulty (default origin). The higher negative value indicates a higher difficulty level for the task.

Table 2. The results of the Rasch analyses (in logit points) run on Tasks 1-5 in L2 English and Finnish.

Logit points	L2 English	Logit points	L2 Finnish
.38	Task 3	.29	Task 2
.00	Task 5	.25	Task 1
.02	Task 1	.20	Task 3
-.10	Task 2	-.29	Task 5
-.27	Task 4	-.46	Task 4

Table 2 shows the order of task difficulty in L2 English and L2 Finnish, from the easiest down. In a multifaceted Rasch analysis, a common measurement scale, called a logit scale, is created that allows the facets of interest, such as learners' proficiency, task difficulty, and raters' severity, to be placed on the same scale and, thus, directly compared. The logit scale is an interval scale, i.e., the distance between the scale points is exactly the same across the scale, unlike the CEFR and NC rating scales, which are ordinal scales in which the distance

between, say, A1 and A2 is probably not the same as the distance between B1 and B2 or C1 and C2. Measuring language proficiency, task difficulty, and later severity on an interval, logit scale thus provides us with more precise information of these phenomena.

Figure 4. A multifaceted Rasch analysis of L2 English performances across Tasks 1-5.

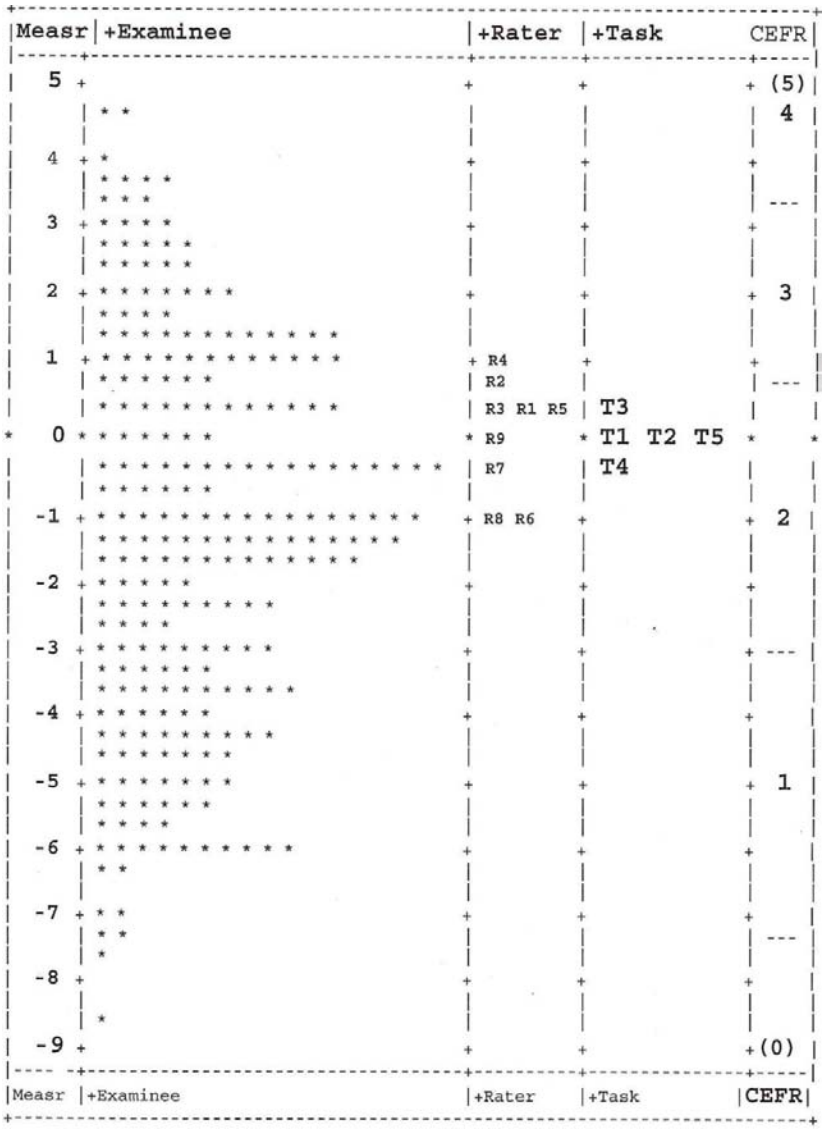
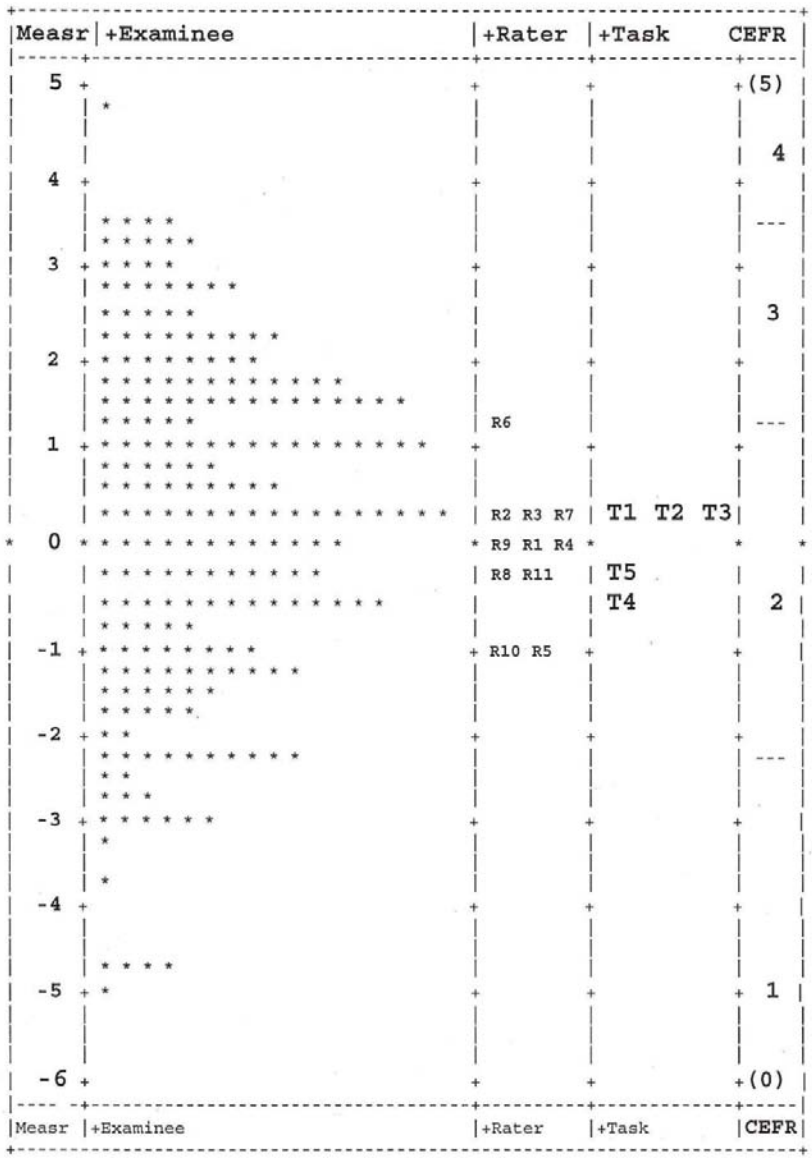


Figure 5. A multifaceted Rasch analysis of L2 Finnish performances across Tasks 1-5.



As Figures 4 and 5 also show, the tasks do not differ very much in terms of difficulty because the minimum and maximum logit values only differ by .65 points for English and .75 for Finnish. The analysis also reveals that there is both within group and between group variation across task performances. Most

notably, Task 2 seemed to be more difficult for L2 English learners than it was for L2 Finnish learners while Task 4 seemed to be difficult for both groups of L2 learners. However, the differences between tasks are rather small.

In sum, the mean L2 ratings and the results of multifaceted Rasch analyses both support the conclusion that the tasks designed for Cefling were quite successful in targeting the expected proficiency levels of the young language learners participating in the project. The statistical analysis of L2 learners' performances reveals that both L2 Finnish and L2 English learners received an average rating of approximately A2 in the 7th - 9th grades (aged 12-16) when the data were collected. Both in L2 English and L2 Finnish, the level of tasks corresponds to the borderline between A2 and B1. The relatively low variation in learner performances and task difficulty reflects the fairly narrow range of proficiency levels elicited by tasks. There could have been a greater number of higher level (B1 and above) performances in the data. The low number of such performances was likely due to the relatively young age of the participants and their limited experience as language learners. Adult performances included in the NC data base will likely yield more varied data.

8. Conclusion

In this chapter, we have been concerned with issues relevant for designing and assessing L2 tasks for SLA research purposes. What should be the most appropriate tasks when collecting evidence of the development of L2 writing from particular proficiency levels?

Task-based SLA research has traditionally been concerned with the effect of task features (structure, complexity, planning time etc.) on L2 performance and development. To gain reliable and valid results, task-based SLA research aims at controlling a number of such features and studying their effect on L2 performance (Norris & Ortega, 2009). An overall generalizable assessment of learners' level of L2 proficiency in terms of communicative success or adequacy has usually not been a major concern; sometimes, various a priori measures or categories have been used to estimate participants' level of language proficiency (course level, years of study), sometimes, it has been assessed by using L2 data elicited during the task performance. Language testing – performance-based testing in particular – has slightly different concerns: first, that the tasks should be designed with a particular proficiency level in mind; second, that the linguistic performance should be grounded within a particular L2 construct that can be assessed by using a reliable and valid rating scale; and third, more than one task and one rater should be used to elicit data. One of the advantages of projects like Cefling is that they have a rating system based on learner performances across several tasks.

A research design that uses several tasks and several raters allows researchers to be more certain about learners' level of proficiency. Rating learners' performances across a number of tasks with reference to e.g. the CEFR proficiency scale allows SLA researchers to define learners' proficiency with more precision and a firmer basis than by relying on the number of years studied or courses taken, for instance. Using more than one task and one rater follows sound measurement principles: several tasks cover learners' proficiency better than one task and this is likely to result in a more generalizable picture of the learner's proficiency – unless one is interested in performance on certain specific tasks only. Measurement instruments always introduce some method effect – or to put it differently, a learner's performance is always a combination of his or her skills and the effects of the task (and a host of other factors). Applying several tasks reduces the effect that any one task format has on the learner's performance. The use of several raters functions basically in the same way although this is usually discussed in terms of reliability: the effect of any one, possibly 'unusual', rater on the outcome is reduced when several are used. Furthermore, if there are enough raters, a rater who is too idiosyncratic can be removed from further data analysis by using the multifaceted Rasch analysis program FACETS.

The use of several tasks and raters also gives SLA researchers more options in how they define the data from which they draw conclusions about language learning. It is possible to include in the analyses only those learners or task performances whose rating fulfils specific quality criteria and leave out from the analyses those learners whose rating is considered too unreliable or otherwise problematic. That is, if one wants to describe the development of linguistic features across different proficiency levels, such as the CEFR levels, it is possible to use only those learners who have been successfully (reliably, validly) placed on specific levels. As a consequence, the validity of the findings about language development improves.

There are different ways to decide which methods represent particular proficiency levels reliably. The one that is used in the first Cefling analyses of linguistic features and their relationship to different proficiency (CEFR) levels is based on direct observation of rater agreement. For the linguistic analyses conducted in Cefling, only those samples of writing were chosen on which the majority of the raters had reached sufficient agreement: three out of four raters in English and two out of three in Finnish rated the texts as belonging to the same proficiency level. An additional criterion was also used: the remaining rater should not deviate from the others by more than one CEFR level up or down. If these criteria were not fulfilled, the sample was not included in the L2 data set to be used for linguistic analyses. At the moment, about 63% of the rated performances in English and 92 % of the rated performances in Finnish are included in the data set for linguistic analyses.

As long as the data set is large enough, it is possible to study the effect of applying other criteria for keeping or rejecting L2 writing samples on SLA data analysis. As is probably the case in most other comparable studies that use more than one rater to assign proficiency level to learners and their performances, it is possible to assign a proficiency level to all samples by using various statistical measures – such as the mean or median rating – and thereby include them in the analyses. More sophisticated analyses such as FACETS can also be used to create a value for each learner that is similar to the mean, for example, but one that also takes into account the difficulty of the tasks he or she has taken and the severity or leniency of the raters who happened to rate them. These 'ability values' could then be related to proficiency scales (such as the CEFR scale) by studying what the analyses tell us about the relationship between the analyzed learners, tasks, raters and the scale(s) used in the rating. Possibly some standard setting procedures would also be needed to confirm the translation of such an ability value scale into the scale in question (see e.g. Kaftandjieva, 2004).

Very little is known about the effect of applying different criteria for the inclusion or exclusion of data on the results of linguistic analyses. The extensive amount of L2 data collected in research projects such as Cefling or others with similar design enables researchers to check whether the linguistic analyses will remain the same when all of the cases – instead of only some 60-70% of the samples – are included and their CEFR level is determined in one of the possible ways described above.

Finally, the analysis of linguistic features – CAF or various linguistic structures – is needed for researchers to be able to tell whether tasks were successful in eliciting the kind of data they were expected to. The ultimate aim of SLA research is to shed light on the nature of L2 development and the dynamic processes that underlie L2 proficiency. In the case of Cefling, linguistic analyses are still in progress.

The linguistic analysis of the data is doubly important since – as we are acutely aware – the assessment of communicative L2 performances cannot be wholly separate from the linguistic features such as complexity, fluency, or an increasing accuracy of a given linguistic structure in the same performances. So how to live with the potential circularity built in our research design? This methodological conundrum has implications for our understanding of L2 development as well. The ultimate aim of the projects combining SLA and language testing perspectives may be to discover the linguistic features that characterize proficiency levels, regardless of how they are determined. Nonetheless, the question is whether the findings will be more like patterns and probabilities of occurrence rather than a list of features, yet to be discovered, that with a 100% certainty are always present at a particular level. And if such features are discov-

ered, what will they tell us about the nature of the assessment process and ratings? What do raters base their judgments on?

At the moment, such linguistic features remain to be fully discovered, but based on the findings presented in this volume it seems more likely that such overarching features, if present, will be rather general, on the order of nouns and verbs or words expressing (agentive or other) relations. These features will be present at level A1 with an increasing fluency and complexity of relations, and at higher levels of proficiency there will be an increasingly target-like use of L2 repertoire, to varying degrees. As Norris and Ortega (2009) point out, a finer understanding of the multidimensional and multicomponential nature of the development of L2 proficiency is not only desirable but also absolutely necessary on both theoretical and methodological levels. A conceptualization of L2 tasks as a unit of activity with multiple dimensions and components – such as task completion, linguistic accuracy, situational or discourse-pragmatic appropriateness – that learners can carry out with a varying degree of success from both communicative and linguistic perspectives is both a challenge and a necessity for future SLA and language testing research.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Macmillan.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, American Psychological Association & National Council on Measurement in Education.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Cohen, A. D. (1998). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Banerjee, J., Franceschina, F., & Smith, A. M. (2004). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports*, *7*, 249–309.

Bloom, B. S., Hastings, J. T., & Madaus. G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Brindley, G. (1994/2009). Task-centred language assessment in language learning. The promise and the challenge. In J. Norris, M. Bygate, & K. Van den Branden (Eds.), *Task-based language teaching. A reader* (pp. 435–454). Amsterdam/Philadelphia: Benjamins.

Bygate, M., Skehan, P., & Swain, M. (2001). Introduction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 1–20). Harlow, England: Longman/Pearson Education.

Byrnes, H. (1987). Proficiency as a framework for research in second language acquisition. *The Modern Language Journal*, *71*, 44–49.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

De Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition. An advanced resource book*. London: Routledge.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. (2007). The effects of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 53–63). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

Douglas, D. (1998). Testing methods in context-based second language research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 141–155). Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

European Association for Language Testing and Assessment (2006). *EALTA Guidelines for good practice in language testing and assessment*. Retrieved from http://www.ealta.eu.org/guidelines.htm

Grabe, W. (2001). Notes toward a theory of second language writing. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 39–57). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*, 49–68.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal*, *80*, 309–326.

Hildén, R., & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen, & V. Kohonen (Eds.), *Foreign languages and multicultural perspectives in the European context/Fremdsprachen und multikulturelle Perspektiven im europäischen Kontext* (pp. 291–300). DICHTUNG, WAHRHEIT UND SPRACHE. LIT-Verlag.

Housen, A., & Kuiken, F. (Eds.). (2009). Complexity, accuracy and fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics, 30*(4).

Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky & F. Hult (Eds.), *Handbook of educational linguistics* (pp. 469–482). Malden, MA: Blackwell.

Huhta, A., & Takala, S. (1999). Kielitaidon arviointi. In K. Sajavaara & A. Piirainen-Marsh (Eds.), *Kielenoppimisen kysymyksiä* (pp. 179–228). Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Hulstijn, J. H. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition*, *19*, 131–143.

International Language Testing Association (2007). *ILTA Guidelines for practice*. Retrieved from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133

Kaftandjieva, F. (2004). *Standard setting. Section B of the Reference Supplement to the preliminary version of the Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*, 48–60.

Larsen-Freeman, D. (2002). Language acquisition and language use from a chaos/complexity theory perspective. In C. Kramsch (Ed.), *Language acquisition and language socialization* (pp.33–46). London: Continuum.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 579–589.

Linacre, J. M. (2010). FACETS. Version 3.66.3. [Computer software]. Chicago: MESA Press.

Luukka, M.-R., Pöyhönen, S., Huhta, A., Taalas, P., Tarnanen, M., & Keränen, A. (2008). Maailma muuttuu – mitä tekee koulu? Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies.

Matsuda, P. K. (2005). Historical inquiry in second language writing. In P. K. Matsuda & T. Silva (Eds.), *Second language writing research. Perspectives on the process of knowledge construction* (pp. 33–48). Mahwah, NJ: Lawrence Erlbaum Associates.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75–100.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

Millman, J., & Greene, J. (1993). The specification and development of tests of achievement and ability. In R. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 335–366). Phoenix, AZ: Oryx Press.

National Core Curriculum for Basic Education (2004). Helsinki: Finnish National Board of Education.

Norris, J. M. (1996). *A validation study of the ACTFL Guidelines and the German speaking test* (Unpublished MA thesis). Honolulu: University of Hawai'i.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, *19*, 337–346.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and

task difficulty in task-based second language performance assessment. *Language Testing*, *19*, 395–418.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu: University of Hawaii Press.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M.H. Long (Eds.), *Handbook of second language acquisition* (pp. 716–761). Malden, MA: Blackwell.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578.

North, B. (1996/2000). *The development of a common framework scale of language proficiency* (Doctoral dissertation). London: Thames Valley University. (Reprinted 2000, New York: Peter Lang).

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*, 590–601.

Reichelt, M. (1999). Toward a more comprehensive view of L2 writing: Foreign language writing in the U.S. *Journal of Second Language Writing*, *8*, 181–204.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke: Palgrave.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 167–185). Harlow, England: Longman/Pearson Education.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211.

Vähäpassi, A. (1982). On the specification of the domain of school writing abilities. In A. C. Purves & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (pp. 265–289). Oxford: Pergamon.

Valdés, G., Haro, P., & Echevarriarza, M. P. (1992). The development of writing abilities in a foreign language: Contributions toward a general theory of L2 writing. *The Modern Language Journal*, *76*, 333–352.

Van den Branden, K. (Ed.) (2006). *Task-based language education. From theory to practice*. Cambridge: Cambridge University Press.

Van den Branden, K., Bygate, M., & Norris, J. M. (Eds.). (2009). *Task-based language teaching. A reader*. Amsterdam/Philadelphia: Benjamins.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*, 101–122.

Wigglesworth, G. (2001). Influences on performances in task-based oral assessments. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks. Second language learning, teaching and testing* (pp. 186–209). Harlow, England: Longman/Pearson Education.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Technical Report #17. University of Hawai'i: Second Language Teaching & Curriculum Center.

APPENDIX 1
Task prompts in CEFLING (translated into English from the Finnish originals)

Task 1. Note to a friend	You've set up a meeting with your English-speaking friend at a café. However, something has come up and you have other things to do. Send an email message to your friend. • Explain why you can't come. • Suggest a new time and place. Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	You've agreed with your friend that you will meet after school at a café. However, you have other things to do. Send an email message to your friend. • Tell why you can't come. • Suggest a new time and place. Write **in Finnish** in clear characters in the space below. Remember to **begin** and **end** the message appropriately.	more basic syntactic structure more frequent vocabulary more basic morphology
Task 2. Message to your teache	You've been away from school for a week. Soon you'll have an English exam. Your teacher, Mary Brown, speaks only English. Send an email message to the teacher. • Tell her why you've been away. • Ask two things about the exam. • Ask two things about the English lessons that were held during the week. Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	You've been away from school for a week. Soon you'll have a Finnish exam. Send an email message to the teacher. • Explain why you've been away. • Ask two things about the exam. • Ask two things about other events that occurred during the week. Remember to **begin** and **end** the message appropriately. Write **in Finnish** in clear characters in the space below.	more basic morphology

>>>

Task 3. Message to an internet store	Your parents have ordered a PC game for you from a British internet store. When you get the game you notice that it doesn't work properly. You get upset and decide to write an email message to the internet store. In the message, say • Who you are • What your parents ordered • Why you're unhappy (mention at least two defects/problems) • How you would like them to take care of the matter • Give your contact information Remember to **begin** and **end** the message appropriately. Write **in English** in clear characters in the space below.	Your big brother has ordered a PC game for you from an internet store. The game works badly. Write an email message to the internet store and say • Who you are • Why you're writing (mention two problems about the game) • How you would like them to take care of the matter • Give your contact information Write **in Finnish** in clear characters in the space below. Remember to **begin** and **end** the message appropriately.	more basic syntactic structure more frequent vocabulary more basic morphology less detailed contextualization
Task 4. Opinion	Choose one of the topics and write about what you think about the matter. Give reasons for your opinion. 1. Boys and girls should go to different classes at school. 2. No mobile phones at school! Write **in English** in clear characters in the space below (continues on the reverse side).	Choose topic 1 or 2 and write for the school paper about your opinion on the matter. Give reasons for your opinion. 1. No mobile phones at school! 2. Parents get to decide how children use the internet. Write **in Finnish** in clear characters in the space below. Write at least five sentences.	more basic syntactic structure more frequent vocabulary more basic morphology more detailed contextualization
Task 5. Narrative	Tell about the scariest / funniest / greatest experience in your life. Choose one. • Tell what happened (what, where, when, and so on). • Tell why the experience was scary / funny / great. Write **in English** in clear characters in the space below (continues on the reverse side).	Tell about one scary or funny thing that has happened to you. • What happened. • Why the experience was scary or funny. Write **in Finnish** in clear characters in the space below.	more basic syntactic structure more frequent vocabulary more basic morphology less detailed contextualization

APPENDIX 2
CEFLING rating scales (based on the CEFR levels)

	OVERALL WRITTEN PRODUCTION	WRITTEN INTERACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT
A1	Can write simple isolated phrases and sentences.	Can ask for or pass on personal details in written form.	Can write a short simple postcard. Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country, etc. such as on a hotel registration form.	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do.
A2	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'.	Can write short, simple formulaic notes relating to matters in areas of immediate need.	Can write very simple personal letters expressing thanks and apology. Can take a short, simple message provided he/she can ask for repetition and reformulation. Can write short, simple notes and messages relating to matters in areas of immediate need.	Can write about everyday aspects of his/her environment, e.g. people, places, a job or study experience in linked sentences. Can write very short, basic descriptions of events, past activities and personal experiences. Can write a series of simple phrases and sentences about their family, living conditions, educational background, present or most recent job. Can write short, simple imaginary biographies and simple poems about people. Can tell a story or describe something in a simple list of points.

>>>

B1	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence.	Can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision. Can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point he/she feels to be important.	Can write personal letters giving news and expressing thoughts about abstract or cultural topics such as music, films. Can write personal letters describing experiences, feelings and events in some detail. Can write notes conveying simple information of immediate relevance to friends, service people, teachers and others who feature in his/her everyday life, getting across comprehensibly the points he/she feels are important. Can take messages communicating enquiries, explaining problems.	Can write straightforward, detailed descriptions on a range of familiar subjects within his/her field of interest. Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points.
B2	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.	Can express news and views effectively in writing, and relate to those of others.	Can write letters conveying degrees of emotion and highlighting the personal significance of events and experiences and commenting on the correspondent's news and views.	Can write clear, detailed descriptions of real or imaginary events and experiences, marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play. Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples.

>>>

APPENDIX 2 >>>

	OVERALL WRITTEN PRODUCTION	WRITTEN INTERACTION	CORRESPONDENCE & NOTES, MESSAGES, FORMS	CREATIVE WRITING & THEMATIC DEVELOPMENT
C1	Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.	Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.	Can express him/herself with clarity and precision in personal correspondence, using language flexibly and effectively, including emotional, allusive and joking usage.	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
C2	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.	As C1	As C1	Can write clear, smoothly flowing, and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.

APPENDIX 3
The Finnish National Core Curriculum scale for writing

A1.1	Can communicate immediate needs using very brief expressions. Can write the language's alphabets and numbers in letters, write down his/her personal details and write some familiar words and phrases. Can use a number of isolated words and phrases. Cannot express him/herself freely, but can write a few words and expressions accurately.
A1.2	Can communicate immediate needs in brief sentences. Can write a few phrases and sentences about him/herself and his/her immediate circle (such as answers to questions or notes). Can use some basic words and phrases and write very simple main clauses. Memorized phrases may be written accurately, but prone to a very wide variety of errors even in the most elementary free writing.
A1.3	Can manage to write in the most familiar, easily predictable situations related to everyday needs and experiences. Can write simple messages (simple postcards, personal details, simple dictation). Can use the most common words and expressions related to personal life or concrete needs. Can write a few sentences consisting of single clauses. Prone to a variety of errors even in elementary free writing.
A2.1	Can manage in the most routine everyday situations in writing. Can write brief, simple messages (personal letters, notes), which are related to everyday needs, and simple, enumerated descriptions of very familiar topics (real or imaginary people, events, personal or family plans). Can use concrete vocabulary related to basic needs, basic tenses and co-ordinate sentences joined by simple connectors (and, but). Can write the most simple words and structures with reasonable accuracy, but makes frequent basic errors (tenses, inflection) and uses many awkward expressions in free writing.

>>>

APPENDIX 3 >>>

A2.2	Can manage in routine everyday situations in writing. Can write a very short, simple description of events, past actions and personal experiences or everyday things in his/her living environment (brief letters, notes, applications, telephone messages). Commands basic everyday vocabulary, structures and the most common cohesive devices. Can write simple words and structures accurately, but makes mistakes in less common structures and forms and uses awkward expressions.
B1.1	Can write an intelligible text about familiar, factual or imaginary topics of personal interest, also conveying some detailed everyday information. Can write a clearly formulated cohesive text by connecting isolated phrases to create longer sequences (letters, descriptions, stories, telephone messages). Can effectively communicate familiar information in the most common forms of written communication. Has sufficient command of vocabulary and structures to formulate most texts used in familiar situations, even if interference and evident circumlocutions occur. Routine language material and basic structures are by now relatively accurate, but some more demanding structures and phrases still cause problems.
B1.2	Can write personal and even more public messages, describing news and expressing his/her thoughts about familiar abstract and cultural topics, such as music or films. Can write a few paragraphs of structured text (lecture notes, brief summaries and accounts based on a clear discussion or presentation). Can provide some supporting detail to the main ideas and keep the reader in mind. Commands vocabulary and structures required for a relatively wide range of writing. Can express coordination and subordination. Can write intelligible and relatively accurate language, even if errors occur in demanding situations, text organisation and style and even if the influence of the mother tongue or another language is noticeable.

>>>

B2.1	Can write clear and detailed texts about a variety of areas of personal interest and about familiar abstract topics, and routine factual messages and more formal social messages (reviews, business letters, instructions, applications, summaries). Can express information and views effectively in writing and comment on those of others. Can combine or summarise information from different sources in his/her own texts. Can use broad vocabulary and demanding sentence structures together with linguistic means to produce a clear, cohesive text. Flexibility of nuance and style is limited and there may be some jumps from one idea to another in a long contribution. Has a fairly good command of orthography, grammar and punctuation and errors do not lead to misunderstandings. Contributions may reveal mother tongue influence. Demanding structures and flexibility of expression and style cause problems.
B2.2	Can write clear, detailed, formal and informal texts about complex real or imaginary events and experiences, mostly for familiar and sometimes unfamiliar readers. Can write an essay, a formal or informal report, take notes for future references and produce summaries. Can write a clear and well-structured text, express his/her point of view, develop arguments systematically, analyse, reflect on and summarize information and thoughts. The linguistic range of expression does not noticeably restrict writing. Has a good command of grammar, vocabulary and text organisation. May make mistakes in low-frequency structures and idiomatic expressions and style.
C1.1	Can write clear, well-structured texts about complex subjects and express him/herself precisely, taking the recipient into account. Can write about factual and fictional subjects in an assured, personal style, using language flexibly and diversely. Can write clear and extensive reports even on demanding topics. Shows command of a wide range of organisational means and cohesive devices. Has a very wide linguistic range. Has a good command of idiomatic expressions and common colloquialisms. Has an extremely good command of grammar, vocabulary and text organisation. May make occasional mistakes in idiomatic expressions and stylistic aspects.

On Becoming an Independent User

Maisa Martin, Sanna Mustonen, Nina Reiman
and Marja Seilonen
University of Jyväskylä / University of Jyväskylä / University of
Jyväskylä / University of Jyväskylä, University of Eastern Finland

The chapter presents tentative results of the project *Linguistic Basis of the Common European Framework for L2 English and L2 Finnish (Cefling)* for three structures of Finnish: the use of local cases, and transitive and passive constructions. The data consist of 669 texts written by adult learners of Finnish as a second language, rated on a functional CEFR scale at levels A1 – C2. The chapter also presents the DEMfad Model used for the analysis and some tentative results which show that often the frequency of use of a structure increases significantly from level A2 to B1 while the leap in accuracy follows it, with the greatest growth between levels B1 and B2. In addition, some aspects of linguistic complexity and the use of constructions as opposed to rules as the starting point of the analysis are discussed.

1. Introduction

The focus of this book is on the relationships between the communicatively defined levels (as in the Common *European Framework of Reference for Languages*, *CEFR*, Council of Europe, 2001) and the linguistic domains the language users control when they have been assessed to be on one of these levels. The learner is assumed to progress through six stages: Basic User (Breakthrough & Waystage), Independent User (Threshold & Vantage), and Proficient User (Effective Operational Proficiency & Mastery). Two assumptions underlie this research: (1) Language proficiency can be described as progressive, and stages along the progression can be established. (2) The range and quality of linguistic items used at a given level, in comparable tasks, bear at least some similarity across learners in the sense that some growth patterns can be shown. The first assumption is taken for granted in this chapter (see Alanen, Huhta, & Tarnanen, this volume), the second one is under our scrutiny.

The research reported in this chapter is based on a project called *The linguistic basis of the Common European Framework levels: Combining second language acquisition and language testing research*, also known as Cefling

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 57-80

(https://www.jyu.fi/ cefling). Like the SLATE network, the Cefling project attempts to bring together the knowledge acquired in the areas of Second Language Acquisition (SLA) and Language Testing. Cefling focuses on writing only, although the methods developed for the analysis are equally applicable to speaking. The target languages of the project are English and Finnish but only the latter is discussed in this chapter. The studies reported here are piloting many of the problematic issues involved, and most of the results are thus only tentative.

There are two main aims for this article: to show how the development of certain linguistic structures can be followed across CEFR levels, and to find evidence for potential co-development in different domains, as well as for interesting diversions of the development from the linear progression. Three structures of Finnish are targeted here as examples of the development in Finnish and as different realizations of the DEMfad Model (below): The use of locative cases, and the transitive and passive constructions.

The emphasis on communication and the importance of a comprehensive view of language, underlying the CEFR, form a part of the conceptual background of the project. As to how language knowledge develops, the broad underlying framework of the project is a usage-based and cognitively oriented view of language learning: acquisition takes place by encountering a growing number of instances of the second language (L2) from which regularities are extracted by use of the general cognitive mechanisms. The domains to be studied are not defined as rules or items but as constructions. Constructions are here loosely defined as units of language which contain a form and a meaning, both of which can vary within some limits (e.g. Goldberg, 2003).

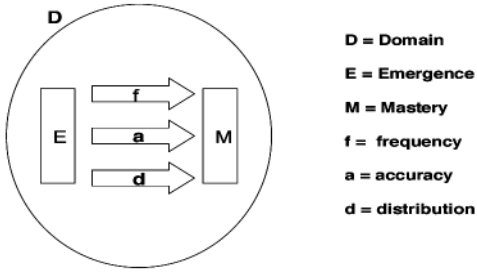
Construction Grammar (see e.g. Fillmore & Kay, 1996; Goldberg, 1995, 2003) has been previously employed to describe several structures of Finnish, e.g. infinitive constructions and the local and dative cases (see e.g. Leino, J., 2003, 2008; Leino P. et al., 2001; Visapää, 2008). However, the construction approach as the linguistic basis of studying second language development has not been previously used for Finnish, and only minimally for other languages (see e.g. Eskildsen, 2008).

1.1. The DEMfad Model

Language proficiency is commonly described as developing in three dimensions: complexity, accuracy, and fluency (the CAF triad). Many measures have been used to track the development of these dimensions (for an overview, see e.g. Housen & Kuiken, 2009; Wolfe-Quintero, Inagaki, & Kim, 1998). In this study, too, three dimensions are tracked, in some ways but not completely comparable to the CAF triad. For the analysis of our data the dimensions are combined in the DEMfad Model, which is intended to make the tracking of the

development comparable across the levels, domains and languages. The structure of the model is shown below.

Figure 1. The DEMfad Model (Franceschina, Alanen, Huhta, & Martin, 2006)



The **domain**s in this model are the areas of developing language skills, such as a construction or a set of constructions or the use of certain linguistic devices in service of semantic functions, a set of vocabulary etc. Thus the definition of a domain here is theory-independent (instead of a construction, a domain could equally well be defined as the application of a rule in the data, if the underlying linguistic framework were rule-based). Obviously some uniform guidelines are needed, particularly to avoid overlaps when the co-development of domains is examined. In this chapter the three domains (local case use, transitive constructions, passive constructions) are considered parallel but separate, with no consideration of potential co-development, so the overlap issue is not discussed here. It will need clarification in future theoretical work on the model as the number of empirical studies based on it grows.

The **emergence** is defined as the first occurrence of some indication of the presence of a domain, e.g. the use of the locative cases (Study 1) emerges when a noun is used with an ending interpretable as a locative suffix. The passive use (Study 3) can similarly be recognized by the presence of a morphological cue. The transitive construction (Study 2) has emerged if a noun or pronoun, verb, and another noun or pronoun follow each other in some order in some context where they can be interpreted to express subject, verb and object (Finnish is an SVO language), regardless of their formal properties.

The construction-based approach solves one of the problematic issues of acquisition criteria (see e.g. Pallotti, 2007): whether an item is actually acquired in a general sense or memorized as a chunk is not of interest in this approach. The first appearance of a recognizable construction, however chunk-like or however far from the target form, is the starting point of the development in the

domain. This is because in the construction-based view the memorized chunk is the basis of acquisition. The development is seen as gradual expansion: the variety of lexicon which can be used in the construction, and the semantic and formal variation within the construction, will grow.

Mastery in the DEMfad model is loosely defined as approximately idiomatic (target-like) use of the domain from the standpoint of frequency and distribution. For accuracy a tentative level of 80%[1] correctness has been established, but this can easily be varied if necessary in the future. For the other two parameters, frequency and distribution, no such mastery level has been set, as the control data from native speakers are being collected and assessed at the time of writing of this chapter, to provide future applications of the DEMfad Model a better point of comparison.

Frequency here is related to the concept of fluency. Fluency is notoriously difficult to define (see e.g. Wolfe-Quintero, Inagaki, & Kim, 1998), in writing as well as in speaking. As the writing tasks were performed in test situations and under time pressure, the number of words written can be seen as one indication of fluency, even if we are keenly aware that there are many other factors involved. Thus the number of words per text is used as an overall measure of fluency. Some of the other aspects of fluency, such as idiomaticity, are discussed as a part of the qualitative analysis of each domain. In a given domain, frequency of occurrence is calculated per 1000 words of running text. Even if the domains – be they constructions or sets of vocabulary – are not comparable in their likelihood of occurrence, and the task effects are great in this area, the patterns of changes in frequency give some indication of the development even across domains.

While the quantitative measures of fluency can be defined without reference to target language, **accuracy** does not exist without a target. The expressions of a given domain are compared to how a native speaker of the same age and education might formulate the same notion, i.e. the grounds of comparison do not always equal the written norm of the target language, particularly in the tasks requiring informal register. Most of the errors, however, are fairly easy to detect as ungrammatical inflections or usages, spelling errors, etc. Obviously, the native writer might commit many of the same errors when not paying attention. Again, the control data from similar groups of native writers will help in making both quantitative and qualitative comparisons between L1 and L2 users.

---

1 The 80% cut-off point is not based on any particular study but chosen simply to give some space for individual variation and random errors committed by both native and non-native language users. For the difficulty of drawing the line, see e.g. Abrahamsson & Hyltenstam (2009).

In the studies reported here no classification of errors is necessary, as the error types are not in focus, just the overall accuracy. In each domain errors are defined by what is required in the domain; for instance in a study of noun inflection the correct form of a case ending would be required for accuracy, while in the domain of the use of local cases (as in Study 2), accuracy is defined by the choice of the case, not by the spelling of the case ending, as long as it is recognizable.

Distribution is the parameter of the DEMfad Model most in need of further work. In the sub-studies of the moment several approaches are experimented with the purpose of finding potential aspects of distribution. The term *distribution* was chosen over the term *complexity* for several reasons. Most importantly, we were not satisfied with the measures of complexity commonly used in the studies of L2 development. The extent of subordination, for instance, does not seem a very useful measure in a language like Finnish, where – apart from relative clauses – subordinate clauses are not syntactically (e.g. word order) or morphologically (e.g. tempus or modus marking) different from main clauses. The L2 learner, who, unlike the L1 learner, already has the mental capacity required for subordination in general, only needs to acquire the necessary conjunctions. Furthermore, lengthy sentences with numerous co- and subordinations are not considered good style even in academic Finnish.

The underlying construction-based framework also sets new demands for tracking the development of linguistic complexity. A construction can grow by the number of different lexical items which can be used within it. It can also grow by the extent to which it can be semantically or syntactically varied, without taking on additional words or morphemes, which are commonly used to measure complexity. The use of a construction may also grow in complexity in a more traditional sense (i.e. in length) as constructions are combined in various ways, as when inserting a construction inside another one (e.g. a necessity construction in a transitive one: *hän ostaa auton* 'he buys a car' > *hänen täytyy ostaa auto* ´he must buy a car').

The term *distribution* at this stage must then be understood as a cover term for several types of phenomena which can be tracked in learner development. Some issues resemble *complexity* in some of the senses in which it has been used in previous SLA research. Some issues come close to *variability*. Many issues discussed under *distribution* do not easily render themselves to quantification. In this chapter we present examples of the types of issues which arise and some ways of dealing with them. In principle we are calling for a more multidimensional view of complexity, as does e.g. Norris and Ortega (2009). In developing the DEMfad Model we also attempt to differentiate and clarify important constructs of second language development, in a way responding to the plea of Pallotti (2009). In this chapter, however, the

problematic issues are brought up more by examples than by theoretical argument, which of course will also be necessary for the future development of the Model.

1.2. Finnish as L2

Unlike English, there is little previous information about the structural development of L2 Finnish. Individual MA level studies exist but as they are based on many different types of data and on diverse theoretical approaches, the results are not comparable between the studies. Nor is the proficiency of the learners rigorously determined; if learners at different levels are compared, the levels are defined by the courses the learners are taking or by the number of years of study etc. For this reason there are no established acquisition orders to be used as a starting point.

It is not possible to know a priori which structures might turn out to yield interesting information about the development of structures across the communicative CEFR levels, so the choices of structures have been based partly on their importance and frequency in written Finnish and partly on where researchers expect to see development based on their teaching- and assessment-based experience. In addition to the domains discussed in this chapter we have the results of some MA theses on the verb *to be* (Kynsijärvi, 2007), on some infinitive constructions (Paavola, 2008), on noun phrases (Ukkola, 2009), and on negation (Martin, 2008). Several other domains are being studied at the moment.

Below the data and methods are described, followed by the sections on the three domains chosen for this chapter. In each section we first present the main characteristics of the domain in question: what is there to be learnt? Then the results of each domain are presented. The overall results are discussed in the last section.

2. Data and methods

The Cefling Project as a whole uses two sets of data for each language: Writing samples from adults taking the National Proficiency Certificate exams and from young learners (12–16, grades 7–9), collected specifically for this project. Similar tasks are used in all sets of data. There are more than 20 different first languages (L1) among the learners of Finnish as a second language (L2). All participants live in Finland and the young learners go to school with Finnish as the language of instruction. Each writing sample has been independently rated to be at a given CEFR level (for the detailed discussion of task design, piloting, and

rating procedures see Alanen et al., this volume). The CEFR scale used for this purpose is purely functional, i.e. the communicative proficiency level has been assessed without attention to the adherence to linguistic norms or the capacity to use certain structures or vocabulary.

The writing tasks for both adults and young learners include three types of texts: an informal message (to a friend etc.), a formal message (a complaint or request to some institution), and an argumentative text (expressing an opinion). In addition, the young learners have written a narrative text.

The distribution of the data across the CEFR levels and the word counts are presented in Table 1. The data set used here includes only the scripts with high inter-rater reliability (see Alanen et al., this volume, for details). In the studies presented here, only the adult data are used.

Table 1. The total number of texts (all tasks) and words in the adult data for L2 Finnish

CEFR Level	Words	Texts	Words/text
A1	4 974	113	44,0
A2	5 702	103	55,4
B1	10 861	126	86,2
B2	9 080	108	84,1
C1	11 550	117	98,7
C2	10 852	102	106,4
Total/Average	**53 019**	**669**	**79,3**

The figures for individual tasks or genres are not given here as only the total results are discussed below. We have aimed at a roughly equal number of texts for each task/genre. The task effects vary between the domains, e.g. passive constructions are more frequent in argumentative texts, as one might expect. Apart from calculating frequencies and percentages, no statistical analyses have been conducted; the amount of data is given here simply to provide background for the reader. The fairly short average length of the texts, about 80 words, is influenced by the tasks: two of the three tasks presented to the adults were messages, in which case a lengthy test performance was not necessarily an advantage. The argumentative texts thus account for more of the growth of the number of words across the levels than do the other texts.

3. Study 1 - The development of the local case phrases in L2 Finnish: from concrete to abstract use

3.1. The Finnish local cases

There are fifteen cases in the Finnish language, eight of which form a subsystem called local cases. Alternative ways of categorizing the cases exist, but this is the approach taken in this study. Functionally, the local case system works like prepositions in Indo-European languages: it is based on the oppositions of directionality and quality. Directionality indicates the difference between expressions of TO (in Finnish the illative, allative, and translative cases), IN/ON/AT (inessive, adessive, and essive cases), and FROM (elative and ablative cases) while quality refers to the nature of the relationships expressed: internal (in English roughly in, into, from/out of X), external (on, onto, off/from X) or general (being/becoming X). Thus either static *being/existence* (on/in/at) or dynamic *direction of the movement* (from on/in/at; to on/in/at) is expressed by one of the cases. (For the classification and the terminology, see Huumo & Ojutkangas, 2006; Jackendoff, 1983.)

In general, and as far as the form, meaning and function of the local cases are concerned, the case system can be understood through locality or spatiality. The spatial domain (in a different meaning here from the DEMfad Model) is the primary one, and the other domains – e.g. action, circumstances, internal states, roles, time and possession – are analogical to the spatial one. According to a number of cognitive theories (see Johnson, 1987; Lakoff, 1987; Lakoff & Johnson, 1980), these expressions are considered to be metaphorical extensions of the spatial relationships and hence more abstract.

The semantics of the local case phrases (static vs. dynamic; direction of the movement; quality of a place) is not particularly transparent to the learner. The system is not thoroughly logical or watertight, either: in non-spatial relationships the use of the cases is more idiomatic, and the movement and direction are often fictive, so that the learner needs to "see the world in a Finnish way". Also, in the spatial domain some verbs require complements in local cases which indicate unexpected meanings, and accordingly, outline the situation differently from the learner's L1 (e.g. the verb *löytää* 'to find' takes the FROM-case in Finnish).

3.2. Research Questions

The study seeks answers to the following research questions: 1) How do the learners use the local case phrases in concrete and in metaphoric domains? 2) How do the static and dynamic uses differ from each other? 3) How do the domains differ from each other?

As the local cases are high in type frequency and have a wide spectrum of meanings and functions, the need for self expression makes them emerge in learner language from early on. However, the frequency, accuracy, and distribution develop from one level to another at a varying rate.

One hypothesis is that the spatial and static expressions are learnt first, as they are cognitively and linguistically simpler than the metaphoric and dynamic ones. This is the view of many cognitive theories (see e.g. Langacker, 1991; Jackendoff, 1983 on *locality hypothesis*) and also Finnish as L2 researchers (see Lauranto, 1997). As to the target-like uses of the cases, however, the dynamic and metaphoric uses of the local cases may be even more frequent than the concrete ones (for the frequencies of the Finnish cases see e.g. ISK, 2004, p. 1179).[2] Thus, in light of the usage-based view of language learning, whereby language is learnt in and through language use (see e.g. Tomasello, 2003), the predictable and consequential hypothesis is that the most common uses – the metaphorical and dynamic ones – of the local cases would be learnt first.

Therefore the overall aim of the study is to examine the basic assumptions about the learning order of local cases and to suggest which cognitive theory seems to be better at explaining the emergence of the local case system in L2 Finnish. The learning order was tested by using the L2 data from Cefling, and the concrete and metaphorical uses were compared across CEFR levels.

3.3. Some Results

Below, the frequency (per 1000 words) and accuracy of the static local cases are presented. The parameter of distribution is here built in the research design: the concrete and metaphoric uses are assumed to develop differently across the CEFR levels.

It turned out that the spatial expressions are most typical of level A1, after which their frequency in the data slightly decreases till level C2 (see the Figures 1 and 2). As the language skills develop, the learner is using fewer concrete expressions, and their frequency becomes more target-like.

The static expressions in the spatial field (*Spat-AT*)[3] are mastered as early as at level A1, even if the form of the case may still falter. In terms of accuracy, these phrases are mastered by level B1 (see Figure 1). The *Spat-TO*-phrases are

2 ISK = *Iso suomen kielioppi* 'The big Finnish grammar', the authoritative descriptive grammar of Finnish.

3 In the abbreviations below *Spat* refers to the spatial, concrete uses of local cases, *Circ* to the abstract, circumstantial uses. AT indicates static uses, TO and FROM dynamic uses.

Figure 1. The frequency and accuracy of spat-AT phrases

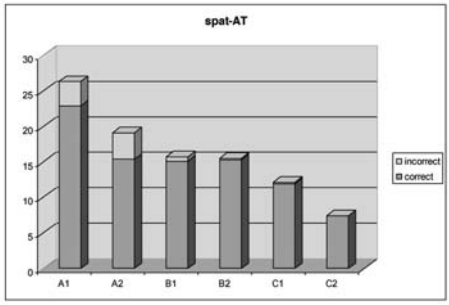
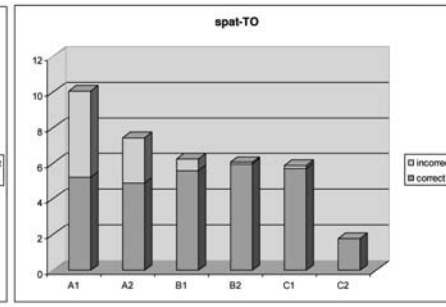


Figure 2. The frequency and accuracy of spat-TO phrases



more complex, both cognitively and morphologically, and hence the number of non-target-like uses is rather high at A1 and A2 levels (see Figure 2).
Meanwhile, the metaphoric uses of the local cases appear to emerge in the reverse order to the concrete uses. The metaphoric uses, *circ-AT* -phrases (*circ* indicating here internal states, action, circumstances, and roles) increase till level B2 (see Figure 3 and Example 1 from level B2).

1) *heillä kaikki on kunnossa*
They-ADESS everything be- PRES-3SG order-INESS
'Everything is in order with them.'

circ-TO -phrases increase till level C2 (see Figure 4). Again, the L2 Finnish learners produce a great number of incorrect TO-expressions, which, in fact, are not mastered until levels B2-C1 (see Figure 4).

Figure 3. The frequency and accuracy of metaphoric, static phrases

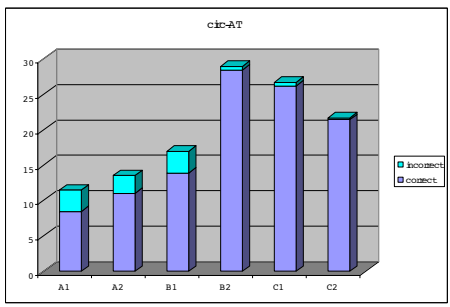
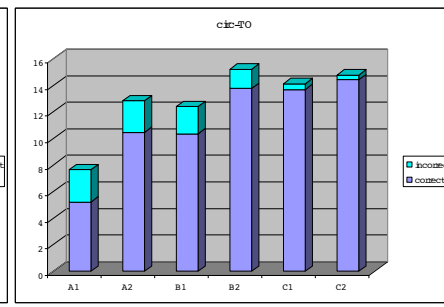


Figure 4. The frequency and accuracy of metaphoric, dynamic phrases



Thus, the developmental path suggested by the data fits the path predicted by one of the cognitively oriented theories and the locality hypothesis: the static, spatial expressions are mastered by level B1 at the latest, whereas it takes more time to learn the metaphoric uses of the same morphosyntactic features. The use of the local cases, particularly in metaphorical domains and in dynamic functions (TO and in particular FROM, which was not presented here but which is notable by its absence till C levels), is a challenge to learners across all CEFR levels.

It should be pointed out that the number of errors or the acquisitional order alone do not fully reveal the logic of the learning process. In addition, a more detailed qualitative analysis is needed to trace the path. The full description of the distributional growth – semantic accuracy, variation of stem words or verbs used in the phrases, etc. – which would provide a more complete picture of the developmental stages, is in progress but remains beyond the scope of this chapter. Characterizing the general tendencies of how the cases are used in concrete and in abstract domains at different CEFR skill levels and in different tasks may also shed light on the sociolinguistic features of the learner language.

4. Study 2 - The development of the transitive construction in L2 Finnish: The Finnish transitive construction

The Finnish transitive construction is an abstracted pattern used in constructing transitive clauses in actual language use. The prototypical construction is the [S [VO]] type: the subject is in the nominative case, but in certain sentence types it may also be in the genitive and partitive cases. In addition, the object takes three different major cases in Finnish: the cases of a total (bounded or resultative event) object are nominative and genitive (example 1), whereas the case of an unbounded (or irresultative, including a negative aspect), indefinite, or partial object is partitive (example 2). Personal pronouns and personal interrogative pronouns have a special accusative form ending in *-t*. Broadly speaking, the case depends on the meaning of the object, or the whole event/action (boundedness/unboundedness). Moreover, in the case of the total object, the choice between the nominative and genitive cases depends on the verb form: for instance, necessive and passive structures require a total object in the nominative case (example 3).

1) *Liisa kirjoitti esseen.* total, positive, resultative/finished action
Liisa write-PST-3SG essay-SG-**GEN**
'Liisa wrote an essay.'

2) *Liisa kirjoittaa esseet**ä**.* unbounded, irresultative action
Liisa write-PRES-3SG essay-SG-**PARTIT**
'Liisa is writing an essay.'

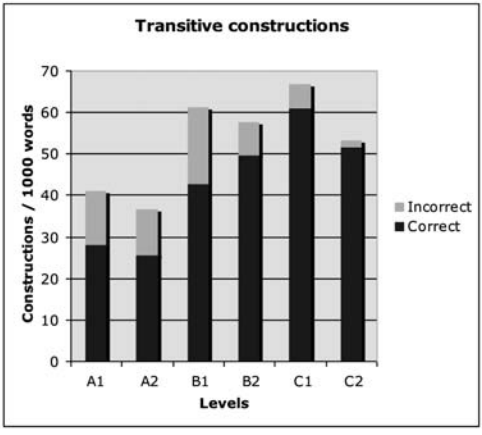
3) *Liisan täytyy kirjoittaa esse**e**.* total, positive, resultative action, **necessive**
Liisa-GEN must-PRES-3SG write-INF1 essay-SG-**NOM**
'Liisa has to write an essay.'

4.1. The development of frequency and accuracy of transitive constructions

The aim of this study is to examine how the frequency and accuracy of Finnish transitive constructions develops in learner language, and, for distribution, what kinds of uses of the transitive construction are typical at each CEFR proficiency level. The transitive construction is a category that is frequent and common in any language. It is semantically open and therefore high in type frequency. In terms of syntactic structure, the selection of the object case in the Finnish transitive construction causes problems, even for advanced learners. In other respects, the prototypical construction is fairly regular and word order is relatively free. Hence the accuracy of the construction is examined using case selection as a criterion of syntactic correctness. The selection of the object cases has been of interest in the area of L2 Finnish, whereas the construction as a broad syntactic unit has not been studied earlier.

Figure 5 presents the frequency of the transitive construction per 1000 words including both correct and incorrect uses across CEFR proficiency levels.

Figure 5. The frequency of the transitive construction / 1000 words across CEFR levels.



As can be seen from Figure 5, the construction emerges in learner language early on and is present at all levels of proficiency. At level A, however, the construction is not as frequent as in the following stages. There is some evidence that learners at level A tend to omit the object in otherwise possible transitive constructions: they use transitive verbs intransitively more often than more advanced learners:

4) *Minä sain uuden valokuvian ja **näytän sinulle**.* (A2)
I get-PAST-3SG new-SG-GEN photograph-PL-PARTIT-GEN and show-PRES-1SG you-ALL
'I got (*a) new photographs and **show you**.'

Regarding accuracy, there is a clear decrease in incorrect object cases between B1 and B2 (see Figure 5). Comparing frequency and accuracy, it can be seen that the apparent growth in the quantity of the construction takes place between A2 and B1, whereas accuracy does not increase until B2. According to the definition of the DEMfad model (mastery: 80% target like occurrences), the use of the object cases is mastered at level B2, in which 86% of the occurrences are target-like. By contrast, at level B1, accuracy is around 70%; almost one third of the uses are incorrect. It can be assumed that an increase in linguistic means may cause problems in accuracy at B1.

4.2. Qualitative change across CEFR levels

The increase of accuracy may indicate development to a certain extent, but it does not explain how the structure and its use change qualitatively. In other words, the quantitative approach can only reveal general tendencies of linguistic development: it does not allow us to draw any far-reaching conclusions about the learning process of a category as broad as a syntactic construction. Therefore, the concept of distribution will be employed here to understand qualitative changes, and will here refer to the types of clausal environments in which a given construction is used. A further interesting question is what types of syntactic variants of the transitive construction are typical at each of the levels. This question has been approached from the point of view of Construction Grammar (e.g. Fillmore & Kay, 1996; Goldberg, 1995), according to which constructions constitute a continuum on the basis of their productivity and abstractness, including idioms with fixed lexical content, idioms that are partially filled, constructions with some filled material, and fully general linguistic patterns (e.g. Goldberg 2003, pp. 219–220).

The findings of a tentative qualitative analysis show that the prototypical construction is the most typical variant at level A. However, it can to some extent be applied to different clausal contexts: for example, it is used in certain subordinate positions, and combined with simple infinitive constructions. The

ditransitive variant (e.g. *give somebody something*) is also used early on, which can be expected due to its semantically concrete nature and syntactically regular form in Finnish.

It is somewhat surprising that even though the use of the transitive construction increases between A2 and B1, the quality does not change that much. The prototypical and ditransitive variants are still frequent, but at the B1 level the transitive constructions are mastered also in subordinate clauses. Infinitive constructions used simultaneously with the transitive construction are more varied, and the necessity construction, for example, becomes frequent:

5) *jos mökillä on vieraat,* ***minun täytyy laitaa ruokaa*** *paljon ja tiskata paljon.* (B1)
If cottage-ALL be-PRES-3SG guest-PL-NOM I-GEN must make-INF1 food-SG-PARTIT a lot and do-INF1 (the dishes) a lot.
'If there are guests at the cottage, **I must make a lot of food** and do the dishes a lot.'

One apparent qualitative change, however, that differentiates B1 from B2 is that at B2 the construction is used in the passive voice much more often.

Lexically more constrained and thus semantically more specific and coherent (see Barðdal, in press) variants of the transitive construction do not emerge until level C. At level B, these kinds of idiomatic transitive expressions are used only occasionally. These partially filled or lexically fully fixed variants can also be understood as distinct constructions (Goldberg, 1995; Leino, 2009), but at the same time, they are still instances of the versatile, general and more abstract [S[VO]] pattern. From the viewpoint of learner language, it seems more fruitful to examine these occurrences as representatives of the prototypical construction, since this perspective reveals how the uses – and constraints – of the construction begin to diverge from those of the general one. The examples below illustrate the use of these less open types:

6) *He eivät ole* ***saaneet siirrettyä*** *opetusmateriaaliaansa uuteen järjestelmään.* (C2)
They no-PRES-3PL be-STEM-NEG **get**-PART-PL transfer-**PASS-PAST PART-PARTIT** teaching material-SG-PARTIT-POSS new-SG-ILL system-SG-ILL
'They haven't been able to transfer their teaching material into the new system.'

Example 6) above expresses resultative action. In the *saada tehtyä* ('get done') variant the auxiliary is always *saada* 'to get', whereas the main verb can vary, provided it is a passive past participle and in the partitive case.

7) *Kansa* ***teki*** *poliitikoi****sta*** *pellejä.* (C2)
People-SG-NOM **make**-PST-3SG politician-PL-**ELAT** clown-PL-PARTIT
'People made politicians clowns.'

Example 7) is an instance of the transitive construction that expresses a change of state. The fixed elements are the verb *tehdä* ('to make') and the elative case *-sta*. In addition to these more constrained transitive expressions, other typical advanced uses at level C are non-finite clauses: they are hardly used at lower levels.

In sum, the clausal environments in which the transitive construction is used become more diverse as language proficiency develops. An increase in linguistic means makes it possible to express meanings more specifically (level C). Furthermore, these results also support the findings concerning the use of passive (Study 3) and local expressions (Study 1).

5. Study 3 - The Finnish passive

In L2 Finnish the use of generic expressions, such as the passive, can be a good indicator of proficiency. Genericity signifies actions or events that are described without specifying the agent, i.e., the person in action or the person observing. By means of generic expressions it is therefore possible for language users to reach a more abstract level in their communication, as opposed to using personal expressions. The aim of this study is to examine the development of the uses of passive in learner language.

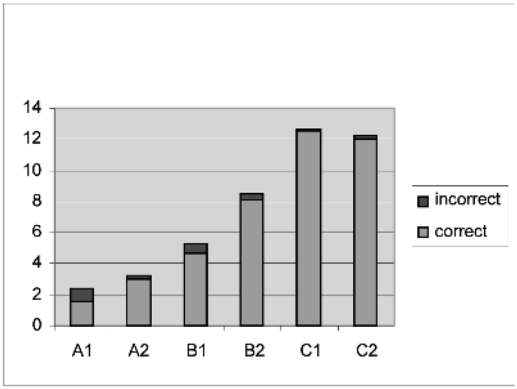
The use of all generic expressions is very common in Finnish, and it is a characteristic of both spoken and written language (see Hakulinen, Karlsson, & Vilkuna, 1980). Besides the passive, other generic forms include e.g. the zero person and the *sinä* 'you'-passive, all of which can be used to create open reference. Furthermore, passive forms are commonly used to replace 1st person plural forms. In this study, however, genericity in L2 Finnish is examined by focusing on only one of the generic expressions: the impersonal passive construction. It is investigated how this construction emerges, varies and develops in the L2 Finnish writing tasks.

The term *impersonal* is used here to convey that the agent has no specific referent. The use of the structure in question, with a passive form as the finite verb, has not been studied before in the context of L2 Finnish. The Finnish passive has the special characteristic that it always refers to a human agent: *Jos ovi avataan.* 'If the door is opened (by somebody)' (see e.g., Shore, 1986). Therefore the Finnish passive is currently often seen as a part of the personal system in Finnish (e.g., Helasvuo, 2006). The passive is a structure lacking an overt subject although the agent can be specified through the context (e.g., Laitinen, 2006). In a passive sentence the agent remains implicit and the passive describes events in less detail than an active sentence (ISK, 2004, pp. 1254–1256, 1284).

5.1. The use of passive in L2 Finnish

The passive construction as a generic expression is present at all proficiency levels from A1 to C2, and the use increases up until C1 level. A more distinguishable increase in the use of this impersonal expression occurs between levels B1 and B2, and levels B2 and C1.

Figure 6. Frequency and accuracy of the passive construction (/1000 words)

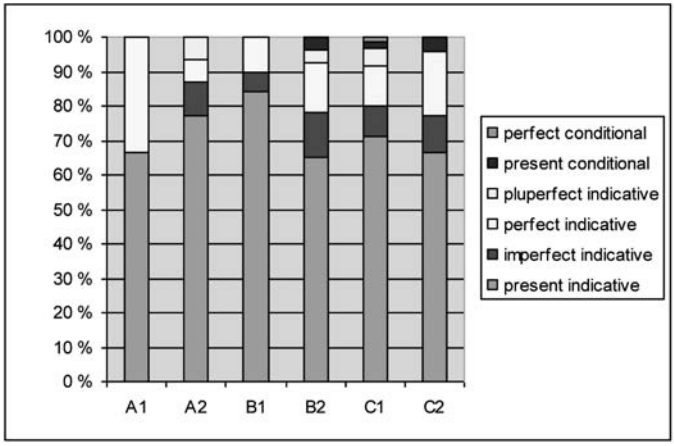


Non-target passive expressions can be found somewhat more frequently at levels A1 and B1, however, at level C the accuracy of the generic passives is almost 100 %: there are only three inaccurate passive forms in the texts. In terms of the DEMfad-model, mastery for accuracy is already reached at level A2 (over 90 % of the forms accurate). In L2 Finnish the passive occurs mostly in the present indicative tense (Figure 7). The morphological form of the passive present might be relatively easy to acquire for a learner of Finnish because in most verb types it is very similar to the dictionary form of the verb, the 1st infinitive (*tehdä* 'to do', *tehdään* 'is done'). The form is also familiar not only from its use as the 1st person plural imperative (*tehdään* 'let's do') but also from the spoken language, where, completed with a personal pronoun, it is widely used in the sense of the 1st person plural indicative (*me tehdään* 'we do').

The similarity between the 1st infinitive and the colloquial use of the passive form in active meaning is arguably also the source of non-target use of the passive, which is most common at levels A1 and B1 (Figure 6). There are examples of the inaccurate use of passives of this kind: the 1st infinitive form is used as a passive or vice versa, for example *myydä* 'to sell' and pro *myydään* 'to be sold'; *mahdollisuus puhutaan suomen kieliä* pro *puhua* 'possibility to speak Finnish'. Moreover, the auxiliary verb is used in the active and the main verb in

the passive in compound passive forms (*olin sanottu* - pro *minulle oli sanottu* 'I was told') and an active construction is blended with a passive one (*minä lähetetään* pro *minä lähetän* 'I will send'; *verkkosivuilla luetaan* pro *verkkosivuilla lukee/ verkkosivuilta voidaan lukea* 'one can read on the websites'.

Figure 7. Tense and mode variation within the passive construction occurrences



While the use of the passive in L2 Finnish increases gradually from one level to another, not only does the range of the verbs grow but also the variation in tenses and modes (Figure 7). However, the present tense is the most frequently used across all levels. At levels A1 – A2 the verbs used in the passive are basic types such as *puhua, mennä* 'speak, go'. As for tenses, at A1 only the present and perfect tenses occur. Figure 7 shows that also the passive imperfect and pluperfect tenses emerge in the texts soon, already at level A2. Conversely, the conditional form, for one, does not appear until level B2. All in all, at level B verb selection varies with synonyms (*sanotaan, mainitaan, on kehuttu* 'is said', 'is mentioned', 'has been praised'). The full repertory of passive forms, along with the past conditional, can be found at level C1 (Figure 7). Compared to B2, the expressions are more abstract and idioms are used at level C (*kulttuuri on valjastettu kaupallisten intressien vetäjäksi* (C2) 'the culture has been harnessed to lead commercial interests'; *on nostettu pöydälle aihe, josta kannattaa todella keskustella* (C1) 'a subject has been raised which is truly worthy of discussion'. The distribution and internal variance of the passive construction thus clearly develops across the levels.

6. Discussion

In addition to providing new information on the development of certain aspects of L2 Finnish writing, the purpose of drawing together results from somewhat disparate domains is theoretical and methodological. Below the implications of the results are discussed in relation to the CEFR and some constructs of the CAF triad.

The three domains of developing Finnish, the use of cases and transitive and passive constructions, studied in this chapter were each examined for frequency, accuracy, and distribution across the CEFR levels. The results of such a varied set of structures – and slightly different theoretical approaches – are obviously variable. As can be expected, the occurrence and accuracy of structures generally increases across the levels. The growth of the passive use in Study 3 is a good example of a fairly steady development, peaking at level C.

When the structure itself is extremely frequent and can hardly be avoided in any type of communicative task, as the local cases in Study 1, the total number of occurrences remains fairly stable while the proportions of the cognitively different uses of the structure have a converse relation: the concrete uses of local cases decrease and the metaphorical ones increase. A similar steady total frequency was found in Kynsijärvi's (2007) study of the verb *to be*, which was mainly used in present and past personal forms (i.e. single word forms) at level A. The auxiliary use of *to be* grew at level B, and other constructions, similarly restricted in meaning and form as the examples given in Study 2, were normally found only at level C.

The transitive construction in Study 1 is a good example of a structure whose frequency grows most between levels A2 and B1, while the accuracy "leap" is between B1 and B2. The development of noun phrases (Ukkola, 2009) was found to have a similar pattern but the frequency increased even more from level B to level C (both differences statistically very significant). For negation structures (Martin, 2008) the total frequency remained stable across the levels, as one might expect, while the accuracy growth from A2 to B1 was statistically significant when compared to the other steps of the CEFR scale. As the studies reported in this chapter are still at the stage of looking for the best indicators of development, no statistical testing has been done here. Nevertheless, it is interesting if leaps of development of any two domains should occur at the same place.

The rating scales used for proficiency were purely communicative, i.e. contained no reference to the language structures or vocabulary, and the raters were strictly instructed to pay attention to communicative proficiency only. They were also experienced and trained in this type of rating (see Alanen et al., this

volume). Yet it is possible – and unavoidable – that the linguistic features of the texts subconsciously influenced the rating and thus explain the results. Even if this is the case, however, there is no a priori reason to expect a similar pattern of development across any two domains, particularly as the domains studied in the project are quite different in nature.

Assuming that the leap from level A to level B, found in some studies in the frequency and/or accuracy of use of linguistic structures, is due to actual developmental factors, why is it where it is? Are the CEFR levels simply not equidistant in such a way that the communicative proficiency suddenly takes a bigger step between A2 and B1 than between other levels? This would explain some of our results: communicative and structural skills grow step-by-step. An alternate explanation is the one displayed in the title of this chapter: the concept of Independent User. It is possible that to become an independent user one needs a fairly large repertoire, and a degree of control, of target language structures, while the Basic User can function non-independently, with the help of interlocutors or, when writing, rely on the willingness of the readers to decipher the intended meaning from fragmentary expressions, without much grasp of the grammatical features of the language.

As to the threesome of complexity, accuracy, and fluency, the latter two seem to respond to the quantitative measures in this study, as in numerous studies before this one. Yet the theoretical question of construct definition remains: to what extent do these measures cover the construct of accuracy, let alone fluency? Finding an answer to this theoretical question remains elusive.

The third parameter, the complexity of the CAF triad, has not been subjected to any type of quantification here as the authors do not find any of the existing measures sufficiently refined. A potential measure which has been considered for future work is IPSyn (Scarborough, 1990). It has been applied to Finnish L1 development in Nieminen (2007). The results, however, were not entirely conclusive, because in the structural development of Finnish (whether L1 or L2) the morphological and syntactic issues are intricately entangled, which constitutes a challenge to all theories, models, and methods of the study of syntactic development. To clarify these issues Nieminen (2007) offers another approach, Utterance Analysis. It offers many insights into the co-development of morphology and syntax, but is too detailed and work-intensive to offer a solution to a large-scale study. In the future we plan to look at new ways of solving complexity issues, such as the application of Dynamic Systems Theory (see e.g. de Bot, Lowie, & Verspoor, 2007). In any case, the results of the Cefling project so far make it clear that the way linguistic structures are used changes across the CEFR levels. Whether this is called complexity or distribution, better qualitative and quantitative ways of accounting for this growth are required.

The construction-based view of language brings up two interesting issues. One is the question of emergence. All structures we have studied to date are present at level A1. Obviously all structures are not there in a single piece of writing but at the group level there is no doubt that structures emerge earlier than is often thought to be the case. There are examples of verb chains (Paavola, 2008), transitive constructions (Study 2 here), and passive (Study 3). Subordination is very common (Martin, 2009). In curricula and elementary textbooks these issues are considered difficult and presented late, if at all. In research it has been customary to write off these occurrences as unanalyzed chunks (see e.g. Pienemann, 1998), at best serving as input for later grammatical learning. In our approach the chunk is a construction which has been learnt, with the potential of expansion and variation provided by future encounters with similar occurrences.

Another question is raised by the construction-and-distribution approach of the Cefling studies: In addition to the three well-known CAF dimensions, is there another one, something which might be called abstractness? Unlike the rule-based approaches, the construction-based framework seems to bring out something about the growth of not only complexity of structures but also the growth from concrete to abstract uses of the structures. In each of the three domains discussed in this chapter there are signs of some "fourth dimension" which also seems to develop across the levels. The abstractness (metaphorical use) is clearly present in the use of local cases (Study 1), as it is built in the theoretical framework of the study, but also transitive and passive constructions seem to extend not only in the number of verbs but also in the quality and variety of verbs. The contexts where the construction is used become more abstract. Transitive constructions are used with abstract subjects, verbs, and objects. The many types of generic expressions of Finnish, including the passive, differentiate and indicate more refined details and implications. In all domains of this chapter (local case use, transitive and passive constructions) expressions become more idiomatic as constructions with limited possibilities of variation and non-literal meaning are added.

The very tentative notion abstractness, however, requires careful definition to avoid circularity. After all, some reference to more abstract uses of language is often made in the assessment criteria of communicative development. What is required is to separate the abstract uses of constructions from the abstraction level of the topic and ideas. Like distribution and variability, abstractness could also be included as one face of complexity in the future search for its more multidimensional definition.

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249–306.

Barðdal, J. (in press). Predicting the productivity of argument structure constructions. *Berkeley Linguistics Society* 32.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

De Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition, 10* (1), 7–21.

Eskildsen, S. W. (2008). *Constructing a second language inventory – the accumulation of linguistic resources in L2 English* (Doctoral dissertation). University of Southern Denmark.

Fillmore, C. J., & Kay, P. (1996). *Construction grammar.* [Lecture notes]. Center for the Study of Language and Information, Stanford, CA.

Franceschina, F., Alanen, R., Huhta, A., & Martin, M. (2006, December). *A progress report on the Cefling project.* Paper presented at SLATE Workshop, Amsterdam.

Goldberg, A. E. (1995). *Constructions. A construction grammar approach to argument structure. Cognitive theory of language and culture*. Chicago, IL: The University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Science*, *7*, 219–224.

Hakulinen, A., Karlsson, F., & Vilkuna, M. (1980). *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Publications/Department of General Linguistics. No 6. University of Helsinki.

Helasvuo, M.-L. (2006). Passive personal or impersonal? A Finnish perspective. In M.-L. Helasvuo & L. Campbell (Eds.), *Grammar from the human perspective: Case, space, and person in Finnish* (pp. 233–255). Amsterdam: Benjamins.

Housen, A., & Kuiken, F. (Eds.). (2009). Complexity, accuracy and fluency (CAF) in second language acquisition research [Special issue]. *Applied Linguistics*, *30*(4).

Huumo, T., & Ojutkangas, K. (2006). An introduction to Finnish spatial relations: Local cases and adpositions. In M.-L. Helasvuo & L. Campbell (Eds.), *Grammar from the human perspective: Case, space and person in Finnish* (pp. 11–20). Amsterdam: Benjamins.

ISK = Hakulinen, A., Vilkuna, M., Korhonen R., Koivisto, V., Heinonen, T., & Alho, I. (2004). *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL: The University of Chicago Press.

Kynsijärvi, T. (2007). *Se johtuu siitä, että minulla oli muistinmenetys. Olla-verbirakenteiden kehkeytyminen oppijankielessä* (Unpublished master's thesis). Department of Languages, University of Jyväskylä.

Laitinen, L. (2006). Zero person in Finnish. A grammatical resource for construing human reference. In M.-L. Helasvuo & L. Campbell (Eds.), *Grammar from the human perspective: case, space, and person in Finnish* (pp. 209–231). Amsterdam: Benjamins.

Lakoff, G. (1987). *Women, fire and dangerous things*. What categories reveal about the mind. Chicago, IL: Chicago University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: The University of Chicago Press.

Langacker, R.W. (1991). *Foundations of cognitive grammar: Descriptive application*. Stanford, CA: Stanford University Press.

Lauranto, Y. (1997). *Ensi askeleita paikallissijojen käyttöön. Espanjankielisten suomen oppijoiden sisä- ja ulkopaikallissijat konseptuaalisen semantiikan näkökulmasta*. Department of Finnish, University of Helsinki.

Leino, J. (2003). *Antaa sen muuttua. Suomen kielen permissiivirakenne ja sen kehitys* (Doctoral dissertation). Helsinki: Suomalaisen Kirjallisuuden Seura.

Leino, J. (Ed.). (2008). *Constructional approaches to language* (Vol. 5). Amsterdam: Benjamins.

Leino, J. (2009). *Results, cases, and constructions. Argument structure constructions in English and Finnish*. Manuscript in preparation.

Leino, P., Herlin, I., Honkanen, S., Kotilainen, L., Leino, J., & Vilkkumaa, M. (2001). *Roolit ja rakenteet. Antaminen, allatiivi ja suomen datiivin arvoitus*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Martin, M. (2008, September). *Negation as a marker of proficiency level in L2 Finnish*. Poster presented at the 18th EUROSLA Conference, Aix-en-Provence, France.

Martin, M. (2009, November). *Lauseet ja sanat kielitaidon kehityksen osoittimina: onko määrä laatua?* Paper presented at the symposium of AFinLA, the Finnish association of applied Linguistics in Tampere, Finland.

Nieminen, L. (2007). *A complex case: a morphosyntactic approach to complexity in early child language* (Doctoral dissertation). Jyväskylä Studies in Humanities 72. University of Jyväskylä.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578.

Paavola, V. (2008). *Haluatko menna muunkansa kalastaman? Verbiketjujen kehkeytyminen suomi toisena kielenä -oppijoiden kielessä* (Unpublished master's thesis). Department of Languages, University of Jyväskylä.

Pallotti, G. (2007). An operational definition of the emergence criterion. *Applied Linguistics*, *28*, 361–382.

Pallotti, G. (2009). CAF: defining, refining and differentiating constructs. *Applied Linguistics, 30*, 590–601.

Pienemann, M. 1998. *Language processing and second language development. Processability theory.* Studies in bilingualism 15. Amsterdam/Philadelphia: Benjamins.

Scarborough, H.S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*(1), 1–22.

Shore, S. (1986). *Onko suomessa passiivia?* Suomi 133. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tomasello, M. (2003). *Constructing a language. A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Ukkola, A. (2009). *Taas yksi nollatutkimus – substantiivilausekkeiden määritteet S2-oppijoiden kielessä* (Unpublished master's thesis). Department of Languages, University of Jyväskylä.

Visapää, L. (2008). *Infinitiivi ja sen infiniittisyys. Tutkimus suomen kielen itsenäisistä A-infinitiivikonstruktioista* (Doctoral dissertation). Helsinki: Suomalaisen Kirjallisuuden Seura.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-E. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity* (Technical report No. 17). Second Language Teaching & Curriculum Center: University of Hawai'i, Mānoa.

Communicative Adequacy and Linguistic Complexity in L2 writing

Folkert Kuiken, Ineke Vedder and Roger Gilabert
University of Amsterdam / University of Amsterdam /
University of Barcelona

The chapter investigates the relationship between communicative adequacy and linguistic complexity (syntactic complexity, lexical diversity, accuracy) of the written output of L2 writers of Dutch, Italian and Spanish. The main goal of the *CALC* study ('Communicative Adequacy and Linguistic Complexity') discussed in the chapter is to investigate the relationship between the communicative aspects of L2 writing, as defined in the descriptor scales of the Common European Framework of References (CEFR, Council of Europe 2001), and the linguistic complexity of L2 performance. It is argued that the interpretation of syntactic complexity, lexical diversity and accuracy is not possible without also taking into account the communicative dimension of L2 production.[1]

1. Introduction

The notion of language proficiency presented in the Common European Framework (CEFR) rests on two pillars, as has been pointed out in several studies (Hulstijn, 2007). Language proficiency is defined both functionally ('can-do statements'), describing the number of domains, functions and roles language users can deal with in the L2 (*what*), and in terms of the quality of language proficiency, e.g. the degree to which language use is effective, precise and efficient (*how* well; Hulstijn, this volume). Whereas the majority of research conducted so far has been concerned with the can-do-statements and the functional scales of the CEFR (Little, 2007), fewer studies have focused on the linguistic dimension, particularly regarding the question of whether it is possible for L2 learners to be situated at different linguistic scales and levels (for instance the

1 We would like to thank the raters for their evaluation of the data of Dutch, Italian and Spanish for both the L2 writers and the L1 writers. We also thank Luuk Nijman for his invaluable help with the statistical analysis of the data. Finally we thank the two anonymous reviewers for their useful comments on an earlier draft of this chapter.

B1 level for vocabulary range, and the A2 level for grammatical accuracy), or the specific ways in which L2 proficiency develops in different European languages. Moreover, the CEFR doesn't indicate, for a given target language, which particular developmental features can be identified as being characteristic for a given scale level (Alderson, 2007). The relationship between language proficiency and language acquisition and the overall development of L2 proficiency (in terms of syntactic complexity, lexical diversity, fluency and accuracy) and the way in which they interact, is thus still unclear (Hulstijn, 2007, this volume).

The relationship between the functional descriptor scales of the CEFR on the one hand and the linguistic scales on the other hand has not been addressed much in the literature either. One of the few studies which have investigated the relationship between the functional and the linguistic dimension of L2 performance is the so called *WISP* study ('What Is Speaking Proficiency'; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2007, in press). In the *WISP* study the oral performance of 208 L2 speakers and 59 native speakers of Dutch was examined both in terms of communicative success and in linguistic terms, concerning the mastery of a number of linguistic skills, such as fluency (i.e. breakdown fluency, speed fluency and repair fluency), syntactic complexity, and vocabulary control. The main question in this type of research is to what extent it may be expected that L2 learners who are situated at the B2 level of the functional descriptor scales of the CEFR have also attained the B2 level with regard to their linguistic performance. In other words, the issue at stake is if and how the communicative adequacy of L2 performance ('getting the message through') is related to the syntactic complexity, lexical variation, fluency, and accuracy of the output.

Whereas in several studies on L2 speaking and writing general measures for assessing the complexity, accuracy and fluency (CAF) of L2 performance are employed, few studies in SLA research report on the communicative success and adequacy of the L2 output. This, however, is in clear contrast with language teaching practice and testing, where often both the communicative dimension and the linguistic complexity and accuracy of the L2 production are independently assessed (Pallotti, 2009). A possible reason for the paucity of studies which explore the relationship between the communicative adequacy of the L2 production and the linguistic forms by means of which the message is conveyed may be due to the absence in the literature of a coherent and clear-cut definition of communicative adequacy as a construct. While communicative adequacy is often interpreted as socio-pragmatic appropriateness (McNamara & Roever, 2007), in other cases it is mainly considered in terms of communicative effectiveness (i.e. success of information transfer; Upshur & Turner, 1995) or successful task completion (i.e. relevance and effectiveness of content according to task instruction; De Jong et al., 2007, in press; Pallotti, 2009). In the present chapter communicative adequacy is interpreted as a task-related, dynamic

and interpersonal construct, focusing both on the specific communicative task which has to be carried out by the speaker or writer (e.g. writing an email to a friend to suggest a restaurant for dinner), and the way the message is received by the interlocutor (the listener or reader).

There is no unanimity in the literature either as to how communicative adequacy could best be assessed. Contrary to CAF measures, general and quantitative measures to rate communicative adequacy are lacking. Moreover, it is not clear by which textual and linguistic features communicative adequacy, in the eyes of raters, is mainly determined (see however the study by Iwashita, Brown, McNamara, & O'Hagan, 2008, for an investigation of the relationship between certain features of the oral L2 production of test-takers and the holistic scores awarded by raters to these performances). In order to be able to assess communicative adequacy, it is thus necessary to resort to proficiency scales, like the ones of the CEFR, containing a set of descriptors to evaluate the features of L2 performance relevant for a particular level of proficiency. However, as has been pointed out in a number of studies, one of the problems of the use of proficiency scales is that they are generally not calibrated or empirically validated and that they do not refer to any theoretical paradigm (see also the chapters in this volume by Alanen, Huhta, & Tarnanen; Hulstijn, Alderson, & Schoonen; Pallotti).

In order to explore the role of communicative adequacy in L2 writing and to establish whether, at a given level of L2 proficiency, communicative adequacy and linguistic complexity develop at the same pace or at the expense of each other, the *CALC* study ('Communicative Adequacy and Linguistic Complexity') was set up. The basic assumption, underlying *CALC*, is that syntactic complexity, lexical diversity, and accuracy cannot satisfactorily be interpreted without taking into account the communicative adequacy of the L2 text. The *CALC* study examines the extent to which the communicative adequacy of the written L2 production is related to the linguistic complexity and accuracy of the text. The corpus on which the analyses have been conducted consists of 206 short written essays. Participants in the study are 34 L2 learners of Dutch, 42 L2 learners of Italian, and 27 L2 learners of Spanish. To create a baseline comparison, the writing tasks have also been administered to a group of 18 native speakers of Dutch, 22 native speakers of Italian, and 10 native speakers of Spanish. In this chapter the data of the L2 learners of Dutch, Italian and Spanish are discussed.

The main goal of CALC is to provide evidence of learner performance, both in communicative and in linguistic terms (i.e. grammar, lexis, accuracy), at a particular scale level of the CEFR. More specifically, the study investigates the relationship between the communicative adequacy and the linguistic complexity, operationalized as syntactic complexity, lexical diversity and accuracy, of

learner output elicited by two writing tasks at the B1 level of overall written production of the CEFR, e.g. a short essay on a topic of interest for a particular functional purpose, in which an opinion has to be reported about factual information (Council of Europe, 2001).

A second aim of the CALC project is to contribute to the description of interlanguage and the role of proficiency in L2 writing by analysing the use of particular linguistic features and structures that typically characterize L2 performance at a given proficiency level, such as the elaboration of the noun phrase and the use of subordinate clauses. Finally, the study investigates the learning dimension of L2, in relation to the CEFR levels. The outcomes of the study are thus relevant for assessment and syllabus design. In this chapter we therefore focus on the first goal of the CALC study.

2. CALC: Design of the study

2.1. Research goals

The main goal of the present study, as pointed out above, is to investigate the relationship between communicative adequacy and linguistic complexity in L2 writing. In more general terms and related to this main goal, this study also aims at contributing new data to the description of interlanguage by using 'diagnostic' linguistic measures which may shed light on the role of L2 proficiency in writing. In order to achieve such goals, students with three different target languages were asked to perform two tasks in writing, and their productions were both rated holistically and measured by means of standardized measures of L2 writing performance. The level of proficiency of the students ranged from A2 to C1, although a large majority of them fell within the range A2-B1.

2.2. Research questions

The following questions were formulated in relation to the goals of this study:

1) What is the relationship in L2 between communicative adequacy as assessed by individual raters, and linguistic complexity (i.e., syntactic complexity, lexical diversity, and accuracy), as assessed also by the same individual raters?
2) What is the relationship in L2 between communicative adequacy, as assessed by individual raters, and linguistic complexity, as assessed by general measures of linguistic complexity?
3) What is the relationship in L2 between linguistic complexity, as assessed by individual raters, and linguistic complexity, as assessed by general measures of linguistic complexity?

The first research question aims at exploring whether a correlation exists between L2 learners' communicative adequacy when performing the task and their linguistic performance. In order to answer this question, holistic measures based on the CEFR descriptors are used and written productions are assessed by individual raters. The second research question tackles the issue of whether communicative adequacy as holistically rated by experienced[2] raters correlates with linguistic complexity, which is calculated this time by means of general linguistic measures (i.e., measures of structural complexity, lexical diversity, and accuracy, which are further described below). The third question deals with the potential correlation between linguistic complexity as perceived by individual raters using holistic measures and linguistic complexity as analysed and calculated by means of general measures of linguistic complexity.

Given the paucity of studies in this area[3] to motivate any directional hypothesis, no specific hypotheses are advanced. We have no sufficient grounds to hypothesize whether communicative adequacy will develop at the same pace as or separately from linguistic complexity. Our study is thus what Seliger & Shohamy (1989, p. 29) have labelled as heuristic or hypothesis generating kind of research.

2.3. Participants

Three groups of university students participated in the study. One group consisted of 34 international students learning Dutch as a second language. Their average age was 26,1 years. There was a wide variety of L1s in this group. Another group consisted of 42 Dutch students who had Italian as a foreign language. Their average age was 21,5 years. The third group consisted of 27 Dutch students taking Spanish as a foreign language. Their average age was 24.9 years. All of the students were enrolled in the modern language section of the University of Amsterdam.

2.4. Materials

Two communicative tasks were used in this study (see Appendix 1 for an example of task 1). Both tasks were similar in terms of type and structure. In both tasks learners were required to make a decision about which of three non-gov-

2 By 'experienced' we mean language teachers who have experience testing students orally, and so may be able to subjectively assess whether a learner is doing a good job communicating a message or not.

3 Although not specifically dealing with the issue of 'communicative adequacy', see Iwashita et al. (2008) for a large scale study on the issue).

ernmental organizations to choose as a candidate for receiving a grant in task 1, and which of three topics presented to the learners they would like to see published in their favourite newspapers in task 2. Both tasks were open in the sense that learners could choose from a number of possibilities. The communicative goal of the tasks was to provide arguments to convince a university board in task 1 and a board of journalists in task 2 to choose their recommended options. Learners were given four instructions in each case: to specify which organization and topic they would support; to describe the aims of the organization and the importance of the topic; to indicate the beneficiaries of the organization's work and the readers that would potentially be interested in the article of their choice; and to provide at least three reasons to convince their addressees. The two tasks were designed with CEFR descriptors which are associated with B1 level, and therefore accessible for Dutch L2, Italian L2 and Spanish L2 participants in our study. Students were told they had 35 minutes for each task and they were told to write at least 150 words (i.e., roughly 15 lines).

2.5. Procedures

Data collection took place during a two-week period. Learners were contacted in class and they were briefly told about the research project. All students participated on a volunteer basis. They were informed that the projects would help researchers understand what students are able to do at each CEFR level. Then learners took a C-test[4], which is described below, and they were also asked to fill out a personal data questionnaire which was either administered during this session or at the teacher's discretion. After having completed the C-test half of the learners were presented with task 1, while the other half started with task 2 and vice versa. (see Appendix 1, example of task 1). Immediately after task performance the students were asked to fill out a perception questionnaire which asked them about the difficulty in performing the task, their own evaluation of their performance, and the interest of the task. Students performed the second task under the same circumstances as the first task. Again, after finishing the second task, they had to fill out a perception questionnaire.

---

4 DIALANG was used as a backup proficiency test but is not reported here. The DIALANG is the test associated with the CEFR, and it provides an indication of the current level of the test-taker at a given point. In this test, which is a subset of the whole DIALANG test, learners are asked to look at lists of verbs in the target language and decide whether they are verbs that exist in the target language. The test provides the learner with a score and places him or her within one of the CEFR levels. In this study only C-test proficiency results are reported.

2.6. Instruments and measures

A number of instruments and measures were used to calculate the learners' proficiency, their productions in holistic terms, and the linguistic dimensions of their written productions. Regarding proficiency, a C-test was used in which learners are asked to complete 100 words in five short texts in which half the letters of every other word have been replaced by blanks. Learners are asked to reconstruct the words by considering contextual clues. The C-test has been shown to correlate with other general proficiency tests (see Babaii & Ansary, 2001; Jafarpur, 1999; Klein-Braley, 1997). Beyond their already assessed discriminatory power and standardized use in the literature, the criterion used to select this test was the fact that it could be completed in just 15 to 25 minutes. Given that data were collected in a classroom context it was important for the tests not to take too long and not to disrupt the class too much.

For the holistic rating of learners' productions, the researchers drew on two main sources: the general descriptors provided by the CEFR on the one hand and the measures developed for the calculation of speaking proficiency by the *WISP* group at the University of Amsterdam, which were also inspired by the CEFR, on the other hand. Based on these two sources, the criteria used for holistic rating (See Appendices 2 and 3) were adapted to the specific tasks learners were presented with. Meetings with the raters for each target language were held in which holistic assessment was presented, discussed, and piloted on a small number of productions. When sufficient agreement was reached, raters were given written productions which were assessed by means of the holistic criteria of communicative adequacy and linguistic complexity. They were asked to do this on their own time and were instructed to rate the productions separately for communicative adequacy and linguistic complexity. For Dutch four raters were asked to judge each text, for Italian and Spanish there were three raters.

As for the general linguistic measures, standardized measures in both oral and written task performance literature were used.

Syntactic complexity:	Clauses per T-unit Subclause ratio
Lexical diversity:	Guiraud Index of Lexical Richness
Accuracy:	Total number of errors per 100 words Total number of errors per T-unit

Clauses per T-Unit and subclause ratio are two well-established measures of structural complexity (Wolfe-Quintero, Inagaki, & Kim, 1998). For lexical diversity the Guiraud Index of Lexical Richness (see Vermeer, 2000, for an eval-

Table 1. Interrater reliability scores as assessed by Cronbach's alpha

L2	Task 1		Task 2	
	Comm. Adeq.	Ling. Compl.	Comm. Adeq.	Ling. Compl.
Dutch L2	0.761	0.889	0.777	0.882
Italian L2	0.702	0.868	0.752	0.735
Spanish L2	0.700	0.756	0.717	0.793

Table 2. Means and standard deviations of measures used in the experiment

	Means and *Standard Deviations*															
	Clauses/ T-unit		Dep.C/ Clause		Guiraud		Tot.Err/ T-unit		Tot.Err/ 100 w.		C-test		Comm. Adeq.		Ling. Compl.	
Dutch L2																
Task 1	1.94	*.83*	.46	*.12*	7.00	*.62*	2.44	*.22*	17.95	*8.06*	78.31	*8.13*	2.99	*.84*	2.82	*01*
Task 2	1.84	*.30*	.45	*.09*	6.88	*2.73*	2.51	*1.19*	18.98	*8.37*	77.29	*7.70*	3.11	*.85*	2.59	*.97*
Italian L2																
Task 1	1.87	*.35*	.45	*.10*	6.47	*1.11*	1.88	*.90*	15.61	*6.83*	69.49	*12.81*	2.97	*.86*	2.19	*.93*
Task 2	1.94	*.41*	.47	*.11*	6.57	*.94*	1.85	*1.30*	15.58	*9.46*	69.43	*12.41*	2.98	*.84*	2.39	*.80*
Spanish L2																
Task 1	1.89	*.27*	.46	*.07*	6.87	*.74*	1.06	*.60*	6.18	*3.45*	83.54	*.55*	2.39	*.71*	2.50	*.81*
Task 2	1.80	*.22*	.44	*.22*	6.87	*.81*	1.20	*.73*	7.20	*4.08*	83.08	*8.54*	2.35	*.80*	2.60	*.89*

uation of the measure) has shown to discriminate among learners at different levels. One of its advantages is that it corrects for differences in text length. In order to discriminate among the different levels of accuracy, two standardized measures in the psycholinguistic and the task-based performance literature were used. The total number of errors per 100 words and the total number of errors per T-units also compensate for differences in text length.

2.7. Statistical instruments

Both descriptive statistics and correlations are used in this study. Descriptive statistics are used to specify means and standard deviations, and Pearson correlations are applied to capture the potential relationship between communicative adequacy and linguistic complexity as measured by raters, and by these two holistic measures and the general measures of performance employed in the study. As will be seen below, correlations were calculated separately for task 1 and task 2. Cronbach's alpha was used for the calculation of interrater reliability.

3. Results

First, interrater reliability for Dutch L2, Italian L2 and Spanish L2 was assessed by means of Cronbach's Alpha, both for tasks 1 and 2 and for communicative adequacy and linguistic complexity (for results see Table 1). The interrater reliability coefficients can be considered sufficient to good, as they varied from 0.700 (Spanish L2, task 1, communicative adequacy) to 0.882 (Dutch L2, task 2, linguistic complexity). In general interrater reliability scores tend to be higher on linguistic complexity than on communicative adequacy.

The descriptives (i.e. means and standard deviations) of the measures that have been used in order to answer our research questions are presented in Table 2. At first sight these numbers look rather stable over the two tasks and the different measures used for the three groups of L2 learners. Perhaps the only salient finding is that the Spanish L2 learners make fewer mistakes than the Dutch L2 and Italian L2 learners. They also obtain higher scores on the C-test, so it might be the case that their general level of language proficiency is higher. However, this is not reflected in the scores of the raters on communicative adequacy and linguistic complexity.

We also considered a possible interdependency of the measures used: for instance, if T-units are longer, then there is more room for errors to be made. Using Pearson correlations we found a positive correlation between the number of clauses per T-unit and the number of errors per T-unit for Dutch L2-learners on task 1 ($r = .366$, $p < .05$), and for Spanish L2 learners on task 1 ($r = .505$, $p < .01$); for Spanish L2 learners on task 1 the correlation between the number of

dependent clauses per clause and the total number of errors per T-unit was significant as well ($r = .504$; $p < .01$). On the other hand, for Italian L2 learners on task 1, the syntactic measures correlated significantly with the Guiraud index (number of clauses per T-unit: $r = .364$, $p < .05$); number of dependent clauses per clause: r. = .365, p < . 05). For task 1 we also noted a (negative) correlation between Guiraud and the number of errors per 100 words ($r = -.340$; $p < .05$).

In order to answer the three research questions regarding the relationship between communicative adequacy and linguistic complexity Pearson correlation coefficients were calculated. We looked at the correlation between these two variables in two ways: bivariately and controlling for the participant's proficiency, as measured by their score on the C-test.

Our first research question concerns the relationship between communicative adequacy and linguistic complexity, both assessed by the raters on a six point Likert scale (see Table 3). Pearson correlation coefficients varied bivariately from 0.604 (Spanish, task 1) to 0.827 (Italian, task 1) and can be considered moderate to good. Taking into account the proficiency level of the participants, the correlation coefficients decreased (from 0.479 for Spanish on task 2 to 0.653 for Italian on task 2). Nevertheless, all correlations but one remained significant at $p < 0.01$.

Table 3. Pearson correlations between communicative adequacy and linguistic complexity, both based on ratings on a six point Likert scale

L2	Bivariate		Controlling for proficiency	
	Task 1	Task 2	Task 1	Task 2
Dutch L2	0.820**	0.768**	0.650**	0.559**
Italian L2	0.827**	0.777**	0.639**	0.653**
Spanish L2	0.604**	0.636**	0.534**	0.479*

* $p < 0.05$; ** $p < 0.01$

Our second research question regards the correlation between communicative adequacy as assessed by individual raters on a six point Likert scale and linguistic complexity as assessed by general measures of syntactic complexity, lexical diversity and accuracy. As mentioned in section 2, the measures we used were the number of clauses per T-unit and the number of dependent clauses per clause for syntactic complexity, the Guiraud index for lexical diversity, and the number of total errors per T-unit as well as the number of total errors per 100 words for accuracy (for results see Table 4). If we first consider the bivariate correlations in Table 4, we notice that no significant correlations can be established for the

measures of syntactic complexity (clauses per T-unit, dependent clauses per clause), whereas almost all correlations for lexical diversity (Guiraud) are significant (except for Spanish on task 1); the same holds for accuracy (total errors per T-unit, total errors per 100 words), since all correlations are significant except for Italian on task 1. However, if we calculate these correlations by factoring in the proficiency level of the participants, the correlation coefficients radically drop and the significant correlations decrease in number. All in all, this seems to indicate that raters, when making communicative adequacy judgments, rely more on lexical diversity and accuracy than on syntactic complexity.

Table 4. Pearson correlations between communicative adequacy as assessed by raters on a six point Likert scale and linguistic complexity as assessed by general measures

CORRELATIONS	Bivariate					Controlling for proficiency				
	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w
Dutch L2										
Task 1	-,015	,079	,582**	-,541**	-,676**	-,353	-,241	,305	-,531**	-,495**
Task 2	,090	,100	,376*	-,394*	-,531**	,010	,010	,278	-,322	-,279
Italian L2										
Task 1	,288	,251	,671**	-,199	-,473**	,105	,088	,318*	,080	-,064
Task 2	-,066	-,016	,568**	-,453**	-,578**	-,226	-,195	,320	-,348*	-,318
Spanish L2										
Task 1	-,279	-,254	,262	-,732**	-,713**	-,361	-,345	-,034	-,717**	-,670
Task 2	-,059	-,310	,636**	-,638**	-,580**	-,189	-,159	,543	-,504*	-,515

* $p < 0.05$; ** $p < 0.01$

A similar picture emerges if we turn to the third research question, concerning the relationship between linguistic complexity as assessed by the raters on a six point Likert scale and linguistic complexity as assessed by the general measures mentioned above (for results see Table 5). Again, concentrating first on the bivariate correlations, Table 4 shows that there are no significant correlations for the measures in terms of syntactic complexity, whereas almost all correlations for lexical diversity are significant (except for Dutch on task 2 and Spanish on task 1), while all correlations are significant with respect to accuracy. Also, taking into account here the proficiency level of the participants, the correlation coefficients tend to drop and the number of significant corre-

lations decreases, although less drastically than in Table 4: most of the correlations concerning accuracy stay significant (except for Italian on task 1), while the correlations on lexical diversity also remain significant for Italian. As with respect to raters' judgements on communicative adequacy, raters also seem to rely more on lexical diversity and accuracy when making linguistic complexity judgments.

Table 5. Pearson correlations between linguistic complexity as assessed by raters on a six point Likert scale and linguistic complexity as assessed by general measures

CORRELATIONS	Bivariate					Controlling for proficiency				
	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w
Dutch L2										
Task 1	,049	,075	,433*	-,757**	-,873**	-,197	-, 194	,081	-,824**	-,833**
Task 2	-,200	-,212	,197	-,726**	-,816**	-,352	-,292	,044	-,756**	-,738**
Italian L2										
Task 1	,213	,188	,673**	-,352*	-,566**	-,015	-,012	,313*	-,157	-,229
Task 2	-,038	,025	,634**	-,471**	-,584**	-,224	-,177	,378*	-,366*	-,334*
Spanish L2										
Task 1	-,249	-,226	,173	-,725**	-,777**	-,349	-,317	-,150	-,660**	-,689**
Task 2	,188	,194	,398*	-,641**	-,761**	,048	,050	,216	-,491*	-,623**

* $p < 0.05$; ** $p < 0.01$

Because the participant's proficiency level as measured by the C-test seems to play a substantial role if it is being controlled for, it was decided to look into this effect more thoroughly. Therefore the groups were split up into two subgroups depending on their score on the C-test. Participants with a C-score in the lowest 40th percentile were assigned to the 'low level' subgroup, and students with a C-score ranking into the highest 40th percentile were placed in the 'high level' group. The intermediate category was excluded from this part of the analysis. Next, the correlations between communicative adequacy and linguistic complexity, both assessed by the raters on a six point Likert scale, were established for each subgroup separately (for results see Table 6). As can be seen from Table 6 these correlations turned out to be significant for Dutch L2 and Italian L2, but not for Spanish L2. It also appears that, in general, the correlations tend to be higher for the high level group in comparison to the low level group.

Table 6. Proficiency level (based on C-test) versus correlation between communicative adequacy and linguistic complexity on a six point Likert scale

L2	Task 1	Task 2
Dutch L2		
LowCtest	,758**	,678**
HighCtest	,854**	,807**
Italian L2		
LowCtest	,575*	,677**
HighCtest	,759**	,772**
Spanish L2		
LowCtest	,340	,615
HighCtest	,534	,225

* $p < 0.05$; ** $p < 0.01$

4. Conclusion and discussion

First of all, the reliability among the raters as measured by Cronbach's alpha was sufficient to good, but the reliability scores for linguistic complexity tended to be higher than for communicative adequacy. All correlations were significant, and they remained significant when we took into account the proficiency level of the participants. In our view, then, raters reached a reasonable level of agreement in the interpretation of both scales. Raters, however, seemed to agree more clearly on their interpretations of the linguistic complexity criteria. It may be the case that experienced raters may have more often dealt with linguistic criteria than with functional ones, which may explain some differences in the interpretation of the communicative adequacy criteria. It may also be the case that the way the various levels of language proficiency were defined in terms of linguistic complexity was more clear to the raters than in terms of communicative adequacy.

If the participants were split up according to their proficiency scores based on a C-test, it appeared that generally the correlations between communicative adequacy and linguistic complexity tended to be higher for the high level group than for the low level group. Our findings also suggest that communicative adequacy and linguistic complexity seem to be more balanced in the case of advanced learners. This is less the case for lower level learners who may concentrate either on communicative adequacy or on linguistic complexity, and for whom it is probably more difficult to focus on communicative adequacy while they are still struggling with form. Another explanation for the higher correla-

tions of the more advanced learners might be that they tend to use longer sentences, which might encourages raters to give them a higher score on communicative adequacy.

With regard to the results of the first question, there are at least two possible explanations as to why high correlations between communicative adequacy and linguistic complexity were obtained, one which refers to the raters and one to the learners themselves. The first one is that the raters either may have perceived linguistically complex compositions as also communicatively adequate or vice versa. Such high correlations suggest that the development of communicative adequacy and linguistic complexity may go hand in hand. As pointed out by Alanen et al. (this volume) the accuracy and complexity of grammar and vocabulary will always have some influence on the communicative adequacy of the L2 production and the extent to which learners are able to complete the task they are rated on. Therefore the linguistic features will influence the ratings of the communicative adequacy. The second explanation is that more proficient learners, who obtained a higher score for linguistic complexity, may also have had more attentional and memory resources to deal with communicative adequacy, while lower level learners need to devote their cognitive resources to working out language problems, at the expense of the communicative and functional aspects of task performance.

The second research question concerned the correlation between communicative adequacy as assessed by individual raters on a six point Likert scale and linguistic complexity as assessed by general measures of syntactic complexity, lexical diversity, and accuracy. We found significant correlations for lexical variation and accuracy, but not for syntactic complexity. This may be explained in the following way: whether learners use simple or complex syntactic structures may simply not have an impact on the perception by raters that learners are being more or less communicatively adequate. On the contrary, the range of vocabulary employed by learners as well as the accuracy of the productions may be associated with the perception that they are also communicatively adequate. This is especially the case when proficiency is factored in, since results for accuracy show a moderately strong correlation with communicative adequacy. This finding also suggests that it may be worthwhile to take a closer look at the results of each individual learner.

A similar picture emerged with respect to the third research question, concerning the relationship between linguistic complexity as assessed by the raters on a six point Likert scale and linguistic complexity as assessed by general measures, that is, there were significant correlations for lexical diversity and accuracy but not for syntactic complexity. Results of structural complexity did not trigger any significant correlations and they also suggest that learners, in general, did not use highly complex structures. It is an issue whether more fine-

grained measures would capture instead any differences in structural complexity (like the elaboration of the noun phrase and the use of subordinate clauses, which were mentioned in the introduction). The results seem to imply that the decisions by raters to grade the students' general linguistic complexity may have been influenced more by the range of vocabulary they used and the accuracy of their productions than by the linguistic complexity of the text, as shown by the moderately strong correlations that were obtained for lexical diversity and accuracy. Accuracy in particular seems to determine the teachers ratings as the correlations tend to decrease when we take into account the proficiency level of the students.

The answers to the three research questions have broadened our view with respect to the nature of the relationship between communicative adequacy and linguistic complexity in L2 writing. The results have given us some insight into the role of communicative adequacy in relation to linguistic complexity and into the assessment of language proficiency by means of raters versus the use of general measures known from SLA research literature. We should, of course, take account of the limitations of our study. One such limitation is that although the participants were submitted to two tasks we only used one task type. But the problem that is troubling us most is the question of what makes raters decide whether a text is considered to be communicatively adequate or not. Interviews with raters and think aloud protocols while raters are judging texts might give us more insight into the motives raters are using to determine the communicative adequacy of a text.

Many other issues remain to be investigated. These include for instance the comparison between the written production of the L2 learners compared to that of the control group of native speakers. Another interesting comparison concerns the differences regarding the relationship between communicative adequacy and linguistic complexity in the three target languages: Dutch L2, Italian L2 and Spanish L2. It would perhaps also be worthwhile to consider carrying out in-depth analyses of specific features of L2 writing, such as the construction of the noun phrase and the verbal phrase or the use of subordinated clauses. But what may be by far the most tempting endeavour is to further explore the role of communicative adequacy in relation to complexity, accuracy and fluency. Another challenge is to investigate whether there are other, more 'objective' ways of measuring communicative adequacy. However, attempts to grasp this notion are still in their infancy.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*(4), 659–663.

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*, 209–219.

Council of Europe. (2001). *Common European framework of references for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

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

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (in press). The effect of task complexity on native and non-native speakers' functional adequacy, aspects of fluency, and lexical diversity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Investigating complerxity, accuracy and fluency in SLA*. Amsterdam: John Benjamins Publishing Company.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantatative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*, 663–667.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System, 27*(1), 79–89.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14*(1), 47–84.

Little, D. (2007). The Common European Framework of reference for language perspectives on the making if supranational language education policy. *The Modern European Language Journal*, *91*, 644–652.

McNamara, T. F., & Roever, C. (2007). *Testing: The social dimension*. Malden, MA/Oxford UK: Blackwell Publishing.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. Special issue, complexity, accuracy and fluency (CAF) in second language acquisition. *Applied Linguistics*, *30*(4), 590–601.

Seliger, H., & Shohamy, E. (1989). *Second language research methods*. Oxford University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, Hawai'i: University of Hawai'i Press.

APPENDIX 1. Task

Every month your favourite newspaper invites its readers to have a say in what will be the leading article for the monthly supplement. This time the Editorial Board has come up with three suggestions: 1) global warming, 2) physical education 3) animal experiments.
Out of these three suggestions one has to be selected. The selection is made by a Readers' Committee. Every member of the committee has to write a report to the editors in which she/he states which article should be selected and why. On the basis of the arguments given by the committee members the Editorial Board will decide which article will be published on the front page. This month you have been invited to be a member of the Readers' Committee. Read the brief descriptions of the suggestions for articles below. Determine which article should be on the front page and why. Write a report in which you give at least three arguments for your choice. Try to be as clear as possible and include the following points in your report:

- which article should be selected;
- what the importance of the article is;
- which readers will be interested in the article;
- why the editorial board should place this article on the front page of the Special Magazine (give three arguments),

You have 35 minutes available to write your text and you need to write at least 150 words (about 15 lines). The use of a dictionary is not allowed.

Suggestions for articles:

1. *Global warming*: there is an ongoing political and public debate worldwide regarding what, if any, action should be taken to reduce global warming.
2. *Physical education*: the government is launching a campaign in order to prevent people from becoming obese and to encourage them to move more.
3. *Animal experiments*: it is estimated that 50 to 100 million animals worldwide are used annually and killed during or after experiments.

APPENDIX 2. Levels of communicative adequacy

0	1	2	3	4	5	6
The participant does not communicate any relevant information	The information which the participant communicates scarcely describes the chosen organization but does not describe its objectives or the people who will benefit from it. The arguments are extremely simple (i.e., single statements not supported by other arguments). They are also not connected to the organization's goals and beneficiaries. And/or: The text lacks coherence, and it is very confusing. > an unsuccessful contribution	The information which the participant communicates briefly describes the chosen organization, and scarcely describes any of its objectives, or the people who will benefit from it. The arguments are simple (i.e., single statements not supported by other arguments) and poorly connected to the organization's goals and beneficiaries. And/or: The text lacks coherence, and it is confusing. > a weak contribution	The information which the participant communicates briefly describes the chosen organization, its objectives, or the people who will benefit from it. The arguments are simple (i.e., single statements not supported by other arguments) but connected to the organization's goals and beneficiaries. And/or: The text is somewhat coherent but a little effort needs to be made. The text meets the minimum requirements. > a moderately successful contribution	The information which the participant communicates clearly describes the chosen organization, its objectives, or the people who will benefit from it. The arguments are complex (e.g. single or multiple statements supported by other arguments) and connected to the organization's goals and beneficiaries. And: The text is coherent. > a successful contribution	The participant contributes complete information about the chosen organization, its objectives, or the people who will benefit from it. The arguments are complex and elaborate (e.g. single or multiple statements supported by other arguments) and clearly connected to the organization's goals and beneficiaries. And: The text flows smoothly, it is coherent, and is convincing. > a very successful contribution	The participant contributes very complete and precise information about the organization, its objectives, or the people who will benefit from it. The arguments are well developed and well structured, and also highly convincing. This information is very well connected with the organization's goals and beneficiaries. And: The text flows smoothly. It is well constructed, coherent and convincing. > a highly successful contribution

APPENDIX 3. Levels of linguistic complexity

0	1	2	3	4	5	6
The participant has a very limited and basic range of simple expressions and vocabulary related to factual information. Word choice is often wrong. Sentences are extremely simple. The participant shows only a limited control of a few simple grammatical structures and sentence patterns. And/or The mistakes in the text make comprehension nearly impossible. > a very poor text in terms of vocabulary, grammar, and orthography.	The participant uses a limited range of highly frequent words and expressions. Some instances of wrong word choice. Sentences are quite simple. The participant uses some simple grammatical structures correctly, but still makes a considerable number of mistakes, including both grammatical (e.g. agreement, verb tenses, prepositions) and orthographic. And/or The mistakes in the text make comprehension difficult. > a rather poor text in terms of vocabulary, grammar, and orthography.	The participant uses a limited range of highly frequent words and expressions with a few instances of wrong word choice. Sentences are simple. The participant uses simple grammatical structures with some mistakes, including both grammatical (e.g. agreement, verb tenses, prepositions) and orthographic. And/or The many mistakes in the text require a little effort from the reader to comprehend the text. > a poor text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of highly frequent words and expressions but there are no instances of wrong word choice. Sentences are simple but they display some complex structures (e.g. relative clauses). The participant uses simple grammatical structures with the odd mistake (e.g. wrong preposition). And There are a few basic/ elementary mistakes that do not prevent comprehension. > an acceptable text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of both high and low frequency words and expressions and there are no instances of wrong choice and words and expressions are appropriate in general. Sentences are moderately complex (e.g. relative clauses, less frequent structures) with no mistakes. And There are minor mistakes or mistakes which are the consequence of trying to use more complex language. > a well written and accurate text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of low frequency and specific words and expressions and there are no instances of wrong word choice which is appropriate to the context. Sentences are quite complex (e.g. relative clauses, infrequent structures) with no mistakes. And There are almost no mistakes in the text and they may be the consequence of trying to use more complex language > a very well written and accurate text in terms of vocabulary, grammar, and orthography.	The participant uses complex (infrequent and rich) and specific vocabulary and expressions that are especially appropriate to the context. Sentences are highly complex, combining many different grammatical structures. And The text contains only the odd mistake. > an extremely well written and accurate text in terms of vocabulary, grammar and orthography.

Exemplifying the CEFR: criterial features of written learner English from the English Profile Programme

Angeliki Salamoura & Nick Saville
University of Cambridge ESOL Examinations

English Profile (EP) is a collaborative programme of interdisciplinary research, whose goal is to provide a set of Reference Level Descriptions (RLDs) for English for all six levels of the Common European Framework of Reference (CEFR). This chapter summarises work and outcomes to date from one of the EP research strands which focuses on corpus linguistics, second language acquisition, psycholinguistics and computational linguistics. The findings discussed are based on the Cambridge Learner Corpus, a database of over 39 million words of written English produced by English learners from around the world, taking Cambridge ESOL examinations. The chapter illustrates how hypotheses formulated from models of second language acquisition (SLA) and psycholinguistics, and a corpus-informed approach are used to investigate second language learner data in order to develop the RLDs for English. By adopting such an approach, English Profile aims to produce an exemplification of the CEFR informed by SLA theory and at the same time, shed light to questions and issues raised by SLA and psycholinguistic theory based on empirical (learner) data. A main focus within EP is the identification of 'criterial features', i.e. features from all aspects of language which can distinguish CEFR levels from one another and thus serve as a basis for the estimation of a learner's proficiency level.

1. Introduction: The English Profile Programme

The English Profile Programme (henceforth EP) is a collaborative programme of interdisciplinary research, whose goal is to provide a set of Reference Level Descriptions (RLDs) for English for all six levels of the Common European Framework of Reference (CEFR) from A1 to C2 (Council of Europe, 2001; see Little, 2007, pp. 167-190 for an extended discussion of the CEFR). (For an overview of the EP research programme see Kurteš and Saville, 2008; Salamoura, 2008.) One strand of the EP research programme focuses on *corpus linguistics, second language acquisition, psycholinguistics* and *computational linguistics*. This chapter will summarise work and outcomes to date from this

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 101-132

research strand.[1] In particular, it will illustrate how hypotheses formulated about prevalent issues in second language acquisition (SLA) and psycholinguistics, and a corpus-informed approach, are used to investigate second language learner data in order to develop RLDs. By adopting such an approach, EP aims to produce an exemplification of the CEFR informed by SLA theory and at the same time shed light on questions and issues raised by SLA and psycholinguistic theory based on empirical (learner) data. The EP aims and approach reflect strongly the objectives and concerns of the SLATE network as detailed in the introductory chapter of this volume (Hulstijn, Alderson, & Schoonen, this volume). This close relationship will become apparent in the remainder of this chapter, where references will be made to SLATE's research questions and goals as presented in the introductory chapter of this volume (see sections "SLATE's overarching research question" and "SLATE's more specific research questions and research goals").

A main focus within the programme is the identification of 'criterial features' for each CEFR level, or in other words, how each level differs from adjacent levels (cf. Hendriks, 2008). This focus closely matches, in fact, SLATE's central research question as formulated in the Introductory chapter, particulary research question 4 (Hulstijn et al., this volume). Of course, EP is concerned with the identification of 'criterial features' for L2 English. 'Criterial features' are linguistic properties from all aspects of language (phonology, morphology, syntax, semantics, discourse, etc.) which can distinguish the different proficiency (here CEFR) levels from one another and thus can serve as a basis for the estimation of a learner's proficiency level. In fact, EP researchers have drawn an analogy between criterial features and the defining characteristics for recognising faces in a police identikit. In an identikit, one does not need to see all the features of a person's face in order to distinguish that person from others; the important defining characteristics that capture essential qualities are typically enough for such a distinction. Criterial features operate in a similar way – they capture essential distinguishing properties of the CEFR proficiency levels (Hawkins, MacCarthy, & Saville, personal communication, April 12, 2010).

What makes a feature 'criterial' is an open question which the EP researchers have been addressing as part of their collaborative research agenda. The programme has adopted an iterative approach to formulating and testing research questions and hypotheses: as empirical evidence is accumulated and

1 For work and outcomes in other more pedagogically and assessment oriented EP research strands, see e.g. Green (2008) and Capel (2009).

shared, more criterial features will be identified. The more the criterial features are understood in relation to the empirical data, so the research questions will be refined over time.

A more comprehensive definition of criterial features is discussed in a following section (*Towards a definition of criterial features*). We will define four types of feature whose use or non-use, accuracy of use or frequency of use may be criterial for distinguishing one CEFR level from the others: (i) acquired/learned language features, (ii) developing language features, (iii) acquired/native-like usage distribution of a correct feature, (iv) developing/non native-like usage distribution of a correct feature. We will also provide examples of linguistic features that have been identified as criterial from the hypotheses formulated and tested thus far.

Although informed by SLA, EP research does not aim to put to the test existing SLA theories or compare and contrast competing SLA models. As [the authors of the introductory chapter in this volume] demonstrate, there are inherent problems in trying to find direct links between proposed acquisitional stages in SLA theory and any of the levels defined in the functional and formal CEFR scales, as these two may not necessarily coincide and, in fact, they were not designed to coincide – the CEFR authors clearly say that the Framework is deliberately atheoretical (Council of Europe, 2001). Instead, Hulstijn et al., (this volume) argue that a potentially fruitful and meaningful research approach for the exemplification of the CEFR should be directed at trying to characterize successful performance at a given functional CEF level, in terms of *how learners express meaning with linguistic form*. Once linguistic forms and their equivalent functions per CEFR level have been identified, the next step would be to find an *explanation* for these findings. The EP research approach follows a similar pathway.

The source of the findings reported in this chapter is the *Cambridge Learner Corpus* (CLC), a database which currently comprises approximately 39 million words of written production from over 160,000 learners of English from a wide range of L1 backgrounds and, critically, is linked to the CEFR (see the next section for a detailed description of the CLC). Starting from such an extensive empirical database, EP research aims at identifying systematic patterns in the learner data either inductively or deductively, which current theories or models of SLA do not necessarily predict. Informed by current issues in SLA and related disciplines (e.g. frequency of input, L1 transfer, etc.), we then formulate explanatory principles that can account for these emerging data patterns, and in turn, can inform current SLA issues. This approach is explained in Hawkins and Filipović (2010):

> In order to find the criterial features of a level we use a mix of inductive and deductive techniques. The CLC reveals many patterns that are not theoretically predictable and that emerge in response to inductive search queries. We also proceed deductively by searching selectively in the corpus for grammatical and lexical patterns that we believe will be distinctive for the different levels, after consulting a broad range of linguistic and psycholinguistic theories that help us make informed decisions about what is likely to be criterial. These theories come from studies of first and second language acquisition, language processing, grammatical complexity, the lexicon and lexical semantics, and language typology. A set of hypotheses was formulated at the outset of the EPP for emerging patterns in second language acquisition, derived from these theories, which were then gradually tested and refined as the project developed.

As emphasised above, EP is a collaborative, interdisciplinary programme of research involving a number of researchers working on different but ultimately interrelated aspects of the EP research agenda. The EP research work referred to in this chapter has been carried out by or under the supervision of Prof. John Hawkins of the Research Centre for English and Applied Linguistics at the University of Cambridge, and in collaboration with the authors who have been critically involved in the EP research agenda in a variety of roles, including commissioning research, reviewing and interpreting research findings and planning future research directions. This chapter aims to collate the main EP research findings as discussed and presented in a number of interim publications that have, thus far, been circulated as internal reports and papers within the EP research circle or presented at internal EP seminars or meetings.

As mentioned above, the research findings discussed in this chapter are based on analyses of learner data from the Cambridge Learner Corpus, a detailed description of which is provided in the next section.

2. Cambridge Learner Corpus

The Cambridge Learner Corpus (CLC) is a unique collection of learner written English, developed since the early 1990s by Cambridge University Press and the University of Cambridge Local Examinations Syndicate (now Cambridge ESOL). At the time of writing the CLC consists of approximately 39 million words of learner written English produced by candidates taking Cambridge ESOL examinations. Approximately 20 million of these data are error-coded. (For the latest updates on the size and scope of the CLC, please consult the official website at http://www.cambridge.org/fi/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=fi_FI)

Alongside the candidates' written responses to the Writing examination papers (extended writing tasks which require candidates to produce an extended piece of text as opposed to short answer questions or cloze tasks) the corpus contains information about the candidates, such as their gender, age, first language, reason for taking the exam etc., and the candidate's overall mark or grade and marks on the other components (typically Reading, Listening and Speaking). The tasks to which the candidates responded are also available, both as images and as a searchable sub-corpus.

At present, EP is using data from a subset of the CLC, amounting to some 26 million words (half of which are error-coded). The examinations in this corpus subset are Main Suite tests (a general purpose suite) consisting of: Certificate of Proficiency in English (CPE), Certificate in Advanced English (CAE), First Certificate in English (FCE), Preliminary English Test (PET), and Key English Test (KET).

The examinations are aligned with the CEFR (Council of Europe, 2001) as shown in Table 1 (adapted from Taylor, 2004, p. 3; see also Jones, 2000, 2001, 2002; Taylor & Jones, 2006, for empirical validation of the alignment). Thus, the EP findings reported in this paper derive from data ranging from A2 to C2 (CF. Table 1) and A1 level may be not mentioned in analyses.

Table 1: Alignment of Cambridge ESOL Examinations with the CEFR scale.

CEFR level	Descriptive title	Main Suite
A1	Breakthrough	
A2	Waystage	KET
B1	Threshold	PET
B2	Vantage	FCE
C1	Effective Operational Proficiency	CAE
C2	Mastery	CPE

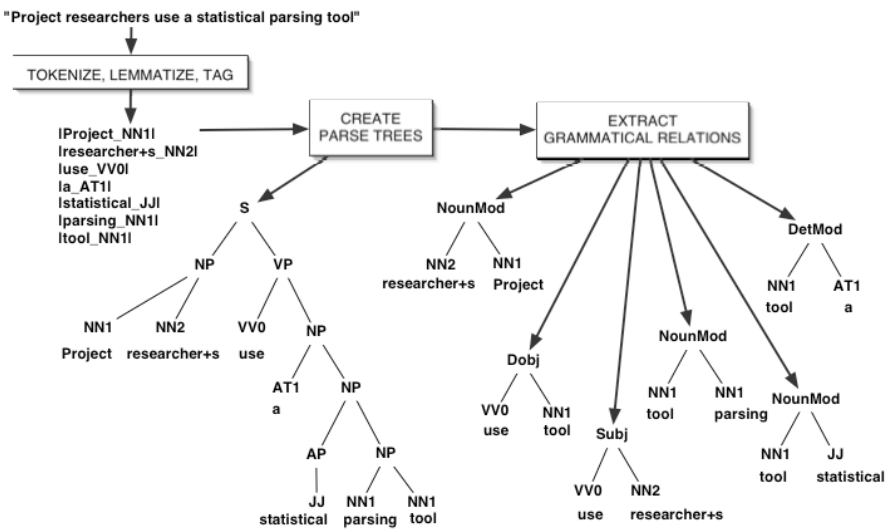
The CLC's system of error codes consists of over 70 codes, each containing two parts: the type of error and the part of speech it applies to. Some examples are provided in Table 2 (after Hawkins & Buttery, 2009a; see Nicholls, 2003, for a description of the error-coding system).

Table 2: Sample error codes in the Cambridge Learner Corpus.

Error Code	Explanation	Exemplification
RN	Replace noun	*Have a good travel* (journey)
RV	Replace verb	*I existed last weekend in London* (spent)
MD	Missing determiner	*I spoke to President* (the) *I have car* (a)
AGN	Noun agreement error	*One of my friend* (friends)
AGV	Verb agreement error	*The three birds is singing* (are)

The existence of these error codes together with the meta-data about the candidate and the exam enabled the calculation of frequency statistics for each exam level, language group, age, etc. (see Tables 4 and 5 on determiner errors in the criterial features section below). To enable searches beyond the individual word level and over a wide range of lexical and grammatical features, the CLC was subsequently tagged for parts-of-speech and parsed using the Robust Accurate Statistical Parser (RASP) (Briscoe, Carroll, & Watson, 2006). RASP is an automatic parsing system

Figure 1: Parsing of 'Project researchers use a statistical parsing tool' using RASP (from Hawkins & Buttery, 2009a, p. 162).



incorporating both grammatical information and statistical patterns, and its operation is summarised by Hawkins and Buttery (2009a, p. 161).

> When the RASP system is run on raw text, such as the written sentences of the CLC, it first marks sentence boundaries and performs a basic 'tokenisation'. Part-of-speech tags are assigned in a probabilistic basis. The text is then 'lemmatized', based on the tags assigned to word tokens. For each sentence a parse forest representation is generated containing all possible parse trees and subanalyses, with their associated probabilities. And a weighted set of grammatical relations is extracted associated with each parse tree. These operations are shown in Figure 1 (*see next page*) for the sample sentence 'Project researchers use a statistical parsing tool', using just one illustrative parse tree and its associated grammatical relations.

For more details on the annotation of the CLC with part-of-speech tags, word lemmas, grammatical relations and complexity metrics, as well as illustrations see Hawkins and Buttery (2009a, pp. 169–172).

3. CLC and SLA

CLC offers a rare opportunity to explicate the CEFR levels for the English language and at the same time explore a number of SLA issues. Previous research in SLA has been constrained by three main sets of limitations. The first set pertains to the reliable identification of the proficiency level of the study participants. Not many SLA studies provide a systematic or assessment-based classification of learners according to proficiency, or if they do, the measurement tools used vary from study to study.[2] Moreover, the varied use of terminology for level description (which is not always adequately defined), e.g. "advanced learners vs. beginners", "high proficiency vs. low proficiency learners", does not help either. As a consequence, the degree of generalisabilty and comparability of findings across different SLA studies has always been a major caveat in these studies. Furthermore, the variability exhibited in the

2 There are, of course, notable exceptions to this statement (e.g. Bialystok, 2001; Hulstijn, 2006; Kessler, 2007; Pienemann & Kessler, 2007; Skehan, 1998, etc.) who have addressed the issue of defining and measuring second language profiles in a systematic way. Pienemann's (1992) COALA programme also represents an early attempt to systematically and reliably classify stages of SLA from empirical data.

learner interlanguage (IL) makes the systematic classification of learners into levels an all important issue for SLA research. It has been observed, for instance, that learners show more variability than native speakers in terms of target language forms, e.g. learners may alternate between forms (e.g. *no* and *not*) to express the same language function (negation) in a non-systematic way and only with increased proficiency do they establish a one-to-one from/function relationship (Gass & Selinker, 2008, p. 259). Or, learners may alternate between use or nonuse of a form, e.g. plural marking, even on the same lexical item (Young, 1991). Given the existing variability in second language learning, second language data can only be a useful tool for research and pedagogic purposes if their proficiency level has been reliably identified, e.g. via valid alignment to a widely accepted proficiency framework, such as the CEFR. CLC's learner data fulfil this requirement.

In the current strand of EP research, learner proficiency (linguistic ability) at each CEFR level will be described in terms of the following properties (Hawkins & Buttery, 2009a, p. 160):

- meaningful units or morphemes;
- lexical items (e.g. nouns and verbs);
- basic grammatical constructions;
- productive syntactic and morpho-syntactic rules;
- exceptions to some of these, i.e. lexical idiosyncrasies.

We do not argue that this is a comprehensive definition of second language proficiency or ability. It is one that covers the lexico-semantic, morpho-syntactic and syntactic aspects of language learning which is the focus of the EP research described in this paper.

The second set of limitations involves the features studied. Most SLA research to date has focused on the investigation of single language features or phenomena across levels (the vertical or developmental dimension of SLA), thus offering little information about the co-occurrence of different developing SLA features within the same proficiency level or, most importantly, about their interaction. The CLC provides a large empirical database of L2 data and an advanced search capability, as outlined above and detailed in Hawkins and Buttery (2009a). These two features allow EP researchers to search for correlations among a large set of diverse lexical and grammatical features and thus draw conclusions about the interrelation of developing SLA features, i.e. the horizontal principles of SLA mentioned above.

The third set of limitations in SLA research concerns the combination of L1s commonly used in L1 transfer research. The majority of SLA studies reach preliminary conclusions based on the study of a handful of L1s. In contrast, CLC comprises data from learners from 130 L1s, thus permitting an extensive study of L1 transfer effects across all major language families. Cross-linguistic differences per CEFR level is one of the main premises under investigation in EP, reflecting SLATE's objective of examining the extent of the source language (L1) involvement in determining a learner's linguistic profile (research question 2 in Introductory chapter, this volume).

4. Towards a definition of 'criterial features'

As discussed in the Introduction, a main focus of EP is the identification of 'criterial features' of English for each CEFR level, which will differentiate one level from adjacent levels. By criterial features we mean features that characterise and distinguish the six CEFR levels. By definition then the criterial features form a subset of all possible features that may appear at a given level and by definition these criterial features can be used, by virtue of their distinctiveness, for diagnostic purposes in language learning, teaching and assessment. For example, the occurrence or not of criterial features in a learner's output can diagnose their CEFR level and/or distinguish them from other learners whose use of the same criterial features differs (significantly) from that of the first learners.

In the *Corpus and Computational Linguistics* strand, Hawkins and Buttery (2009b) have identified four types of feature that may be criterial for distinguishing one CEFR level from the others. Although couched primarily in grammatical terms (i.e. lexical semantic, morpho-syntactic and syntactic features), this classification may also be extended to encompass other types of language features. The four categories, illustrated with examples, are as follows:

1. Acquired/Learned language features

These are language features a learner masters at a given level and uses accurately and consistently at the higher levels. In this category fall the *'positive grammatical properties'* that Hawkins and Buttery (2009b) describe as:

> …correct properties ... that are acquired at a certain L2 level and that generally persist at all higher levels. E.g. property P acquired at B2 may differentiate [B2, C1 and C2] from [A1, A2 and B1] and will be criterial for the former. Criteriality characterises a set of adjacent levels in this case. Alternatively some property might be attained only at C2 and be unique to this highest level.

For instance, verb co-occurrence frames appearing for the first time at B1 level, e.g. NP-V-NP-NP structures (*She asked him his name*), are criterial for [B1, B2, C1, C2], whereas new verb co-occurrence frames appearing at B2, e.g. NP-V-NP-AdjP (Obj Control) (*He painted the car red*), are criterial for [B2, C1, C2] (see Tables 3b-c).

2. Developing language features

These are features that appear at a certain level but they are unstable, i.e. they are not used correctly in a consistent way. This category includes what Hawkins and Buttery (2009b) call 'negative grammatical properties of an L2 level, i.e.:

> …incorrect properties or errors that occur at a certain level or levels, and with a characteristic frequency. Both the presence versus absence of the errors, and the characteristic frequency of the error (the 'error bandwidth') can be criterial for the given level or levels. E.g. error property P with a characteristic frequency F may be criterial for [B1 and B2]; error property P′ with frequency F′ may be criterial for [C1 and C2].

Hawkins and Buttery (2009b) define criteriality for "*negative grammatical properties"*, i.e. errors, as follows:

An error distribution is criterial for a level L if the frequency of errors at L differs significantly from their frequency at the next higher and lower levels, if any. Significance amounts to a difference of at least 29% from level to level, which guarantees at least one standard deviation from the mean. Two or more levels can be grouped together for criteriality if each is not significantly differentiated from any immediately higher and lower levels (i.e. by less than 29%).

For instance, preposition errors (e.g. *When I arrived at London*) do not show significant differences in frequency of occurrence between B1 and B2 or between B2 and C1, but their frequency of occurrence drops significantly from C1 to C2 level. Therefore, the relevant "error bandwidth", as defined above, is criterial for B1, B2, C1 versus C2. As explained above, the three levels (B1, B2, C1) are grouped together in this case since they are not significantly different from each other with respect to their error frequency/bandwidth.

Given the evolving nature of second language acquisition/learning, one would predict that several language features would pass through a developing stage before they are acquired/learned. So, one feature that is still developing at one proficiency level may be acquired at the next level up, or a feature may be developing across more than one level.

Although we analyse incorrect language properties, we do not conduct a mere error analysis in its traditional form (e.g. Corder, 1967). We are not looking at individual, random errors; we are looking at error patterns; we are not only considering errors, i.e. what learners cannot do (cf. (2) and (4) below), but also what learners can do at the same CEFR level (cf. (1) above and (3) below). It is the combination of correct and incorrect language features in a learner's IL that will provide a comprehensive insight into learners' second language performance and overcome shortcomings inherent in traditional error analysis (see e.g. Gass & Salinker, 2008, p. 102ff). As Gass and Selinker (2008) note, errors "provide windows onto a system – that is, evidence of the state of a learner's knowledge of the L2" (p. 102). It is in this sense that we take into account incorrect properties when defining and identifying criterial features.

3. Acquired/Native-like usage distributions of a correct feature

> Positive usage distributions for a correct property of L2 that match the distribution of native speaking (i.e. L1) users of the L2. The positive usage distribution may be acquired at a certain level and will generally persist at all higher levels and be criterial for the relevant levels, e.g. [C1 and C2] (Hawkins & Buttery, 2009b).

For example, the relative distribution of indirect object/oblique relative clauses (*the professor that I gave the book to*) at C1 (4.63%) and C2 levels (4.29%) in relation to relatives on other positions *(*subjects, direct objects and genitives) in CLC learner data approximates the distribution observed in the British National Corpus (BNC; at least 4.33% - see Table 6 for the usage of different types of relative clauses as percentage of total within each CEFR level in the CLC learner data). This distribution is, therefore, a strong candidate for a criterial feature (acquired usage distribution) for C1 & C2 (Hawkins & Buttery, 2009b).

4. Developing/Non native-like usage distributions of a correct feature

> Negative usage distributions for a correct property of L2 that do not match the distribution of native speaking (i.e. L1) users of the L2. The negative usage distribution may occur at a certain level or levels with a characteristic frequency F and be criterial for the relevant level(s), e.g. [B2] (Hawkins & Buttery, 2009b).

The same distribution described above, i.e. the relative distribution of indirect object/oblique relative clauses (*the professor that I gave the book to*) to relatives in

other positions (subjects, direct objects and genitives), at CEFR levels A2 (1.61%), B1 (1.62%) and B2 (2.80%) in the CLC departs significantly from the typical usage distribution of native speakers (at least 4.33% in BNC - see Table 6 for percentages of relative clause usage in the CLC learner data). This distribution is thus an example of a developing usage distribution at A2-B2 levels and is criterial for A2, B1 and B2 versus C levels (Hawkins & Buttery, 2009b).

Of course, not all criterial features will have diagnostic power at an individual learner output/script level as this was described at the beginning of this section. A number of both acquired and developing properties, defined as criterial for a certain level based on frequency of occurrence across all scripts (or a wide range of learner data), may be absent from an individual script not necessarily because the learner hasn't mastered them but simply because the small size or specific focus of the individual script may not allow or encourage the use of these properties. Or in other words, there is also the question of how dense the learner data is in terms of criterial features – the absence of a sufficient amount of such features may also be due to too little evidence of structures in the individual learner data rather than the unsuitability of criterial features for diagnostic purposes.[3] Hawkins and Filipoviç (2010) are currently investigating this issue further in an attempt to capture the diagnostic relevance of criterial features when applied to individual scripts. Identifying which criterial features can also serve as diagnostics for language learning, teaching and assessment purposes is one of the main research leads pursued within EP and is in accordance with SLATE's concerns about "which linguistic features, emerging from [CEFR] profiling research, can serve as successful tools in the diagnosis of learners' proficiency levels and of weaknesses that require additional attention and training" (research question 4 in Introductory chapter, this volume).

Moreover, the aforementioned four types of criterial features describe not only what learners *can do* (types 1 and 3) but also what learners *cannot do* (type 2) and what learners *cannot do as well as or to the same extent as* native speakers (type 4). All these three aspects of language performance form fundamental parts of second language learning which, according to the SLATE group, should be investigated to identify the limits of learners' performance at each CEFR level (research question 3 in Introductory chapter, this volume).

3 We are grateful to an anonymous reviewer for this remark.

5. Criterial features and SLA

The search for the above four types of criterial features within the CLC is primarily driven by current issues and questions in SLA theory (e.g. frequency of occurrence of L2 structures, L1 transfer; see Doughty & Long, 2005, for an overview of such issues), as well as by psycholinguistic principles of processing efficiency and complexity (Hawkins, 2004). Based on Hawkins' (2004) theory of *Efficiency and Complexity in Grammars*, a number of general patterns and principles of developing SLA stages of English have been identified using the CLC. These principles and patterns, in turn, enable us to define a number of criterial features across the CEFR levels. This section will illustrate three of these principles drawing on Hawkins and Buttery (2009a), and Hawkins and Filipoviç (2010).

Maximize Frequently Occurring Properties (MaF).

> Properties of the L2 are learned in proportion to their frequency of occurrence (as measured, for example, in the British National Corpus): more frequent exposure of a property to the learner facilitates its learning and reduces learning effort.

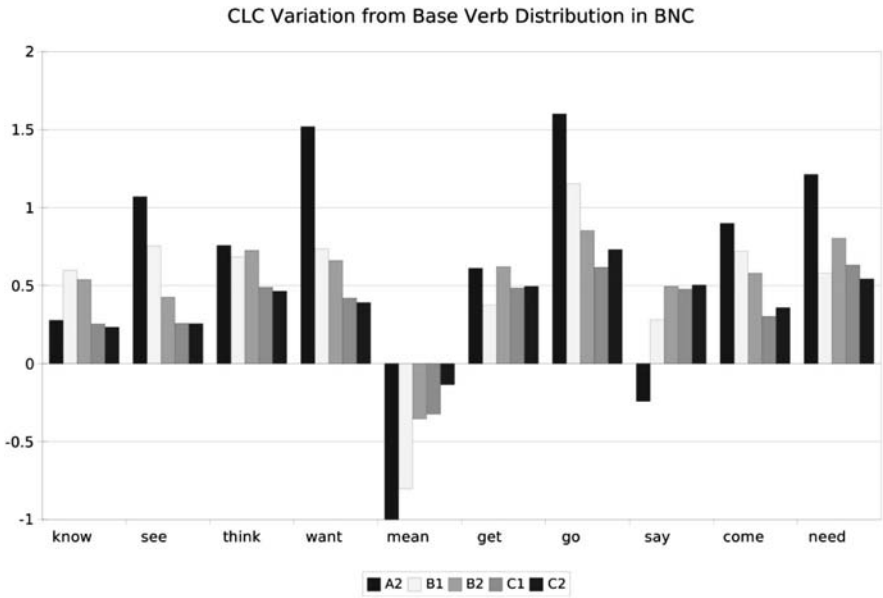
I.e. more frequent properties result in earlier L2 acquisition, more of the relevant properties learned, and fewer errors, in general. Infrequency makes learning more effortful.

In the L1 literature, the importance of frequency of occurrence of a property for its acquisition is attested in the work of Tomasello (2001, 2003) and Diessel (2004) among others. Tomasello (2001, 2003) claims that children's early production consists of specific linguistic expressions that imitate the language they hear around them. It is only later that children creatively combine these expressions to reach adult competence. Tomasello and Brooks (1998), for instance, showed that children younger than 3 years of age use novel verbs only in sentence frames in which they have heard these verbs occurring. In their study, children heard a novel verb (*tam*) for the first time in an intransitive sentence frame, such as *The sock is tamming* (with a meaning similar to *roll* or *spin*). When they were subsequently shown a picture where someone was "tamming" something and were asked *What is X doing?*, most children preferred to re-use the novel verb with the intransitive frame rather than with a transitive one (despite the fact that the question was prompting them to produce a transitive frame). A number of other studies using different constructions corroborate these findings (e.g. Dodson & Tomasello, 1998; Brooks & Tomasello, 1999; Akhtar, 1999). Diessel (2004) also discusses frequency of occurrence in the ambient language as one of the major factors

motivating the order of acquisition of complex sentences in the first language, such as infinitival and participial complement constructions, relative clauses, adverbial and co-ordinate clauses.

In SLA, the work of MacWhinney and colleagues on the Competition Model (Bates & MacWhinney, 1987; MacWhinney, 1987a, 1992, 1997) provides examples of the frequency principle. In the framework of the Competition Model (MacWhinney, 1987b) and more recently of the Unified Model (MacWhinney, 2008), forms (e.g. lexical items) provide cues to functional interpretations for sentence comprehension and conversely, underlying functions provide cues to retrieving forms for sentence production. "Cue availability", i.e. *how often* a particular form occurs as a cue for a certain underlying function, is one major factor in determining "cue validity"[4],

Figure 2: CLC variation from base verb distribution in BNC, including present tense, uninflected forms only from both corpora (reproduced from Hawkins & Buttery, 2009a, p. 164).[5]



4 The other factor is "cue reliability", that is, how reliably a form marks a function.

5 A1 is not included in this figure and the subsequent data tables as the EP findings reported in this paper derive from data ranging from A2 to C2 (cf. Table 1).

which, in turn, determines ease of learning of form/function correspondences. In other words, second language learners are figuring out the form/function mappings of the target language through repeated exposure to these mappings and the frequency of occurrence of these mappings in the input is a decisive factor. Connectionism approaches to SLA also emphasise the role of frequency. In the framework of a connectionist model, second language learning is seen as the extraction of regular patterns from the input. Using a connectionist model, Ellis and Schmidt (1997), for instance, demonstrated frequency effects for the acquisition of L2 morphology, supporting earlier findings in this area by Larsen-Freeman (1976), who argued that frequency of occurrence is a major determiner of the order of acquisition of morphemes in L2.

Turning now to the CLC learner data, one illustration of the MaF principle is the use of the ten most common (frequent) verbs in English (*know, see, think, want, mean, get, go, say, come, need*) by L2 learners in CLC. Hawkins and Buttery (2009a) found that nine out of ten of these verbs are overrepresented in the earlier stages of L2 learning, moving gradually to more native-like L1 English use. Figure 2 shows the ratio of relative frequency of the base form of these ten lexical verbs in the CLC compared with their occurrence in the BNC, taking into account present tense, uninflected forms only in both corpora. Bars above the zero line indicate an over-use in the CLC in comparison to the BNC; bars below the zero line indicate under-use.

Figure 2 shows that overall these common verbs are indeed overrepresented in the CLC relative to the native speaker's usage in the BNC – apart from the verb *mean*. This overrepresentation is a feature of all levels but it declines at the higher level in accordance with the MaF principle above.

Another illustration of MaF in the CLC data comes from verb co-occurrence frames. Williams (2007) analysed verb co-occurrence frames using the Briscoe-Korhonen subcategorisation frame system (cf. Briscoe, 2000; Briscoe & Carroll, 1997; Korhonen, Krymolowski, & Briscoe, 2006; Preiss, Briscoe, & Korhonen, 2007) and recorded the appearance of new frames from A2 to B2 as shown in Tables 3a-3c. There were no new verb co-occurrence frames at C levels suggesting that these basic constructions of English are, more or less, learned by B2. As Hawkins and Buttery (2009a) remark, C2 levels require a different and more subtle kind of analysis in order to capture progress, and projects are planned to explore this issue.

Table 3a: New verb co-occurrence frames at A2 level (Williams, 2007)

Frame	Example
• NP-V	*He went*
• NP-V (reciprocal Subj)	*They met*
• NP-V-PP	*They apologized* [*to him*]
• NP-V-NP	*He loved her*
• NP-V-Part-NP	*She looked up* [*the number*]
• NP-V-NP-Part	*She looked* [*the number*] *up*
• NP-V-NP-PP	*She added* [*the flowers*] [*to the bouquet*]
• NP-V-NP-PP (P = *for*)	*She bought* [*a book*] [*for him*]
• NP-V-V(+ *ing*)	*His hair needs combing*
• NP-V-VPinfinitival (Subj Control)	*I wanted to play*
• NP-V-S	*They thought* [*that he was always late*]

Table 3b: New verb co-occurrence frames at B1 level (Williams, 2007)

Frame	Example
• NP-V-NP-NP	*She asked him* [*his name*]
• NP-V-Part	*She gave up*
• NP-V-VPinfin (WH-move)	He explained [*how to do it*]
• NP-V-NP-V(+ *ing*) (Obj Control)	*I caught him stealing*
• NP-V-NP-PP (P = *to*) (Subtype: Dative Movement)	*He gave* [*a big kiss*] [*to his mother*]
• NP-V-NP-(*to be*)-NP (Subj to Obj Raising)	*I found him (to be) a good doctor*
• NP-V-NP-Vpastparti (V = passive) (Obj Control)	*He wanted* [*the children*] *found*
• NP-V-P-V*ing*-NP (V = + *ing*) (Subj Control)	*They failed in attempting the climb*
• NP-V-Part-NP-PP	*I separated out* [*the three boys*] [*from the crowd*]
• NP-V-NP-Part-PP	*I separated* [*the three boys*] *out* [*from the crowd*]
• NP-V-S (Wh-move)	*He asked* [*how she did it*]
• NP-V-PP-S	*They admitted* [*to the authorities*] [*that they had entered illegally*]
• NP-V-S (*whether* = Wh-move)	*He asked* [*whether he should come*]
• NP-V-P-S (*whether* = Wh-move)	*He thought about* [*whether he wanted to go*]

Table 3c: New verb co-occurrence frames at B2 level (Williams, 2007)

Frame	Example
• NP-V-NP-AdjP (Obj Control)	*He painted* [*the car*] *red*
• NP-V-NP-*as*-NP (Obj Control)	*I sent him as* [*a messenger*]
• NP-V-NP-S	*He told* [*the audience*] [*that he was leaving*]
• NP-V-P-NP-V(+*ing*) (Obj Control)	*They worried about him drinking*
• NP-V-VPinfin (Wh-move)(Subj Control)	*He thought about* [*what to do*]
• NP-V-S (Wh-move)	*He asked* [*what he should do*]
• NP-V-Part-VPinfin (Subj Control)	*He set out to win*

Critically, Williams found that the progression from A2 to B2 correlates with the frequency of these frames in native speaker corpora – the most frequent frames appear at A2 moving progressively to less frequent frames at B1 and B2. Table 4 provides the average token frequencies in BNC for the subcategorisation frames identified as occurring for the first time at A2-B2 in Tables 3a-c.

Table 4: Average token frequencies in native English corpora (BNC) for the new verb co-occurrence frames appearing from A2-B2 in the CLC (cf. Tables 3a-c).

CEFR Level	A2	B1	B2
Average token frequency	1,041,634	38,174	27,615

Maximise Structurally and Semantically Simple Properties (MaS)

> Properties of the L2 are learned in proportion to their structural and semantic simplicity: simplicity means that there are fewer properties to be learned and less learning effort is required. Simpler properties result in earlier L2 acquisition, more of the relevant properties learned, and fewer errors. Complexity makes learning more effortful, in general, since there are more properties to be learned.

In addition:

> Properties of the L2 are used in proportion to their structural and semantic simplicity: simplicity means that there are fewer properties to be processed in on-line language use and less processing effort is required.

That is, simplicity and complexity affect both learning and processing. This principle then predicts that simpler constructions will be acquired earlier than more complex ones. But what constitutes a simple or a complex structure? There is not as yet a satisfactory account of complexity in SLA research (see Rimmer, 2006, for a review of complexity in SLA) and this is why we turn to L1 and typology research. Hawkins (2004) defines complexity as:

> Complexity increases with the number of linguistic forms and the number of conventionally associated (syntactic and semantic) properties that are assigned to them when constructing syntactic and semantic representations for sentences. That is, it increases with more forms, and with more conventionally associated properties. It also increases with larger formal domains for the assignment of these properties. (p. 9; see also Hawkins, 2009)

The different types of relative clauses as listed in the Keenan-Comrie Accessibility Hierarchy Hypothesis (AH; 1977) provide a good example of increasing complexity across structures. Table 5 below lists types of relative clauses in order of complexity following the Keenan-Comrie AH and Hawkins (2004). Hawkins (1994, pp. 37-42, 2004, pp. 177-8, 2009) has argued that the relative clauses down the AH involve increasing complexity of the processing domains for the different relativizable positions. According to Hawkins, the number of dominating and co-occurring syntactic nodes required for these relativizable positions correlates strongly with their AH ranking – the lower the ranking the higher the number of syntactic nodes required and the greater the complexity of the relative clause. (For a detailed account of this argument, the reader is referred to Hawkins, 1994, 2004, 2009).

Table 5: Relative clause types in order of complexity based on the Keenan-Comrie Accessibility Hierarchy (1977) and Hawkins (2004)

Relative Clause Types *(in order of complexity based on the Keenan-Comrie Accessibility Hierarchy, 1977, and Hawkins, 2004)*	
Subject Relatives	*The student **who/that** wrote the paper*
Direct Object Relatives	*The student **who(m)/that** I taught*
Indirect/Oblique Object Relatives	*The student **to whom** I gave the book* *The student **who/that** I gave the book **to***
Genitive Relatives (within a Subject)	*The student **whose** supervisor retired*
Genitive Relatives (within a Direct Object)	*The student **whose** supervisor I know*
Genitive Relatives (within an Indirect Object)	*The student **to whose** memory I devoted the book* *The student **whose** memory I devoted the book **to***

According to the MaS principle above then, simpler relative clause types, such as subject relatives, will be learned earlier than more complex ones, such as relatives in genitive positions (see Table 5). And this is what is actually attested in the CLC data (see Table 6 below). Hawkins and Buttery (2009b) found a clear progression from A2 to C2 in the appearance of new relative clause types. This progression correlates with the increasing complexity of the relative clauses involved (subject > object > indirect/oblique > genitive; cf. Keenan & Comrie, 1977; Hawkins, 1994). For example, fairly complex relative clauses from the lower positions of the Keenan-Comrie AH (1977), such as relatives in indirect/oblique object positions (*the student to whom I gave the book*), are rare

before B2, whereas relatives in genitive positions (*the student whose supervisor retired*) do not appear at all before C1.[6]

Table 6: Usage of different types of relative clauses as percentage of total within each CEFR level

	A2	B1	B2	C1	C2
Subject RCs	67.7%	61.1%	71.1%	70.3%	74.4%
Direct Object RCs	30.7%	37.3%	26.1%	25.0%	20.9%
Indirect Object RCs	1.6%	1.6%	2.8%	4.6%	4.3%
Genitive RCs	0.0%	0.0%	0.0%	0.1%	0.2%

These findings are in accordance with a substantial body of SLA research on the acquisition of relative clauses by second language learners from various L1 backgrounds which show that the order of acquisition appears to follow the ranking of the AH (e.g. Doughty, 1991; Eckman, Bell, & Nelson, 1988; Gass 1979, 1980; Gass & Ard, 1984; O'Grady, Lee, & Choo, 2003). Gass (1979, 1980), for example, found that the production of different types of relative clauses by learners of English with a wide range of L1s could be predicted based on their rank order in the AH – that is, the percentage correct of subject relatives was higher than the percentage correct of direct object relatives, and so on. These findings were based on empirical data from a variety of experimental tasks, including free compositions, sentence combining and grammaticality judgements. (But see also the more recent studies of Jeon & Kim, 2007; Ozeki & Shirai, 2007, for some exceptions.)

Lexical semantic properties appear to follow a similar acquisitional path in L2: simpler semantic senses are learned earlier than more complex, figurative senses (Hawkins & Filipoviç, 2010). Table 7 illustrates this principle by looking at the acquisition of *break*.[7]

---

6 In all levels, the learners were responding to open-ended, essay type questions (with minimal prompts) that allowed considerable freedom of expression in terms of content and form. It is thus unlikely that the distribution in Table 6 is due to what the writing tasks in the different CEFR levels allowed the learners actually to use rather than progressive learning across the levels. However, this remains an empirical question that will be further investigated as additional learner data from non-exam settings are collected and analysed within EP (see the Conclusions section).

7 These findings deriving from learner corpora appear to be in accordance with earlier work on prototypicality and language transfer by Kellerman (1977, 1978, 1979).

Table 7: Occurrence of break in the CLC

CEFR level	Example of occurrence	Type of use
A2	break	basic physical sense
B1	break the routine	additional sense of INTERRUPT
B2	break an argument / promise	additional sense of NOT OBEY
C1	break the bank	Idiomatic
C2	break the wall that surrounds him	original figurative

Maximise Positive Transfer (MaPT)

> Properties of the L1 which are also present in the L2 are learned more easily and with less learning effort, and are readily transferred, on account of pre-existing knowledge in L1.

Similar or identical L1/L2 properties will result in earlier L2 acquisition, more of the relevant properties learned, and fewer errors, in general, unless these shared properties are impacted by other factors, such as high complexity (cf. MaS) or low frequency of occurrence (cf. MaF). Dissimilar L1/L2 properties will be harder to learn by virtue of the additional learning that is required, in general, and this learning may be more or less effortful depending on other factors (cf. Odlin, 2005, for a summary of relevant research literature in SLA). With respect to errors, dissimilar L1/L2 properties will result either in more errors or in structural avoidance (and hence possibly in fewer errors, e.g. Schachter, 1974). The more obligatory or unavoidable the lexical/grammatical property in question, the more we will see errors rather than avoidance.

A test for the above principle is the use of definite and indefinite articles in L2 English by learners whose first language has an article system and learners whose first language does not use articles. MaPT predicts that the acquisition of

---

In a series of studies, Kellerman investigated learners' intuitions about which of the different meanings of the Dutch verb *breken* can be 'transferred' in English, i.e. translated into its English cognate *break*. He found that the dimension of 'prototypicality' largely determined Dutch learners' judgements in that more core senses of *breken* (e.g. "He broke his leg", "The cup broke") presented higher percentages of transferability than less core, figurative senses (e.g. "The underground resistance was broken", "A game would break up the afternoon a bit"). This order of 'transferability' closely resembles the order of acquisition of the senses of *break* in English as L2 identified in the present study.

the article system of English will be easier for the former group of learners. And this is what is attested in the CLC data (Hawkins & Buttery, 2009a). Table 8 displays error rates for *the* (definite) and *a* (indefinite) articles at CEFR levels A2-C2 by French, German and Spanish learners of English, whose first language has a similar article system to that of English. The numbers are percentages of errors compared to the total number of correct uses. As Table 8 shows, error rates are generally low for these learners without significant differences between the CEFR levels (Hawkins & Buttery, 2009a).

Table 8: Missing Determiner Error Rates for L1s with Articles (Hawkins & Buttery, 2009a, p.168)

Missing 'the'					
	A2	**B1**	**B2**	**C1**	**C2**
French	4.76	4.67	5.01	3.11	2.13
German	0.00	2.56	4.11	3.11	1.60
Spanish	3.37	3.62	4.76	3.22	2.21
Missing 'a'					
	A2	**B1**	**B2**	**C1**	**C2**
French	6.60	4.79	6.56	4.76	3.41
German	0.89	2.90	3.83	3.62	2.02
Spanish	4.52	4.28	7.91	5.16	3.58

Now compare the data in Table 8 above with those displayed in Table 9 below which shows error rates for the same articles by Turkish, Japanese, Korean, Russian and Chinese learners of English. These languages do not have an article system. Error rates are significantly higher across all CEFR levels than error rates at the equivalent levels by learners with first languages which have articles. However, unlike learners whose first language has articles, learners whose first language has no articles show, in general, a linear improvement, i.e. a decline in error rates as they progress through the CEFR levels.[8] (A more detailed study

8 The only exception are the Chinese learners who present an inverted U-shaped pattern of error rates, particularly in the use of a, with significant improvement only at C2.

(Alexopoulou, 2010) on definiteness and indefiniteness and the various ways in which articles (including zero article) are used across the different CEFR levels is currently underway in EP).

Table 9: Missing Determiner Error Rates for L1s without Articles (Hawkins & Buttery, 2009a, p. 169)

Missing 'the'					
	A2	**B1**	**B2**	**C1**	**C2**
Turkish	22.06	20.75	21.32	14.44	7.56
Japanese	27.66	25.91	18.72	13.80	9.32
Korean	22.58	23.83	18.13	17.48	10.38
Russian	14.63	22.73	18.45	14.62	9.57
Chinese	12.41	9.15	9.62	12.91	4.78
Missing 'a'					
	A2	**B1**	**B2**	**C1**	**C2**
Turkish	24.29	27.63	32.48	23.89	11.86
Japanese	35.09	34.80	24.26	27.41	15.56
Korean	35.29	42.33	30.65	32.56	22.23
Russian	21.71	30.17	26.37	20.82	12.69
Chinese	4.09	9.20	20.69	26.78	9.79

These findings are in line with an abundance of earlier studies that report L1 transfer effects in SLA – not only for articles (e.g. Dušková, 1983) but for all aspects of language learning (for a comprehensive review of studies on L1 transfer and cross-linguistic influence see e.g. Odlin, 2005; Gass & Selinker, 1992, 2008). There are also numerous transfer theories and models, ranging from generativist ones (e.g. Full Transfer/Full Access Model, Schwartz & Sprouse, 1996) to processing approaches (e.g. Developmentally Moderated Transfer Hypothesis, Pienemann, Di Biase, Kawaguchi, & Håkansson, 2005; L1 and L2 cue competition, MacWhinney, 1992, 2008) and connectionist approaches (Ellis, 2008). The main questions investigated by these theories include the extent of availability of L1 features during the initial and later stages of SLA, and the interaction of or competition between L1 and L2 features. It is beyond the scope of this chapter and EP to provide a full review of existing transfer the-

ories or test their predictions, as explained in the Introductory section. Odlin (2005) remarks that "[t]he highly diverse evidence for transfer has impeded attempts to develop truly comprehensive theories of cross-linguistic influence. In the more credible attempts at theory-building, researchers have focused on what is admittedly only part of an overall model" (p. 437). The EP goal is to account for L1 transfer as one of the factors within a multi-factor model of SLA using a complex adaptive system and computer simulation (see the concluding section for more details).

With respect to the identification of criterial features, the aim of EP research is two-fold. First, it aims at identifying criterial features that are not L1-specific as one of the main aims of the CEFR is to compare L2 learners across their L1s. Examples of such features are provided under the MaF and MaS principles above. However, as the cross-linguistic data in the last section indicate, cross-linguistic variation does exist among L2 learners across the CEFR levels, and in some cases, e.g. the omission of articles, this variation persists up to C2 level. EP research, therefore, also aims at investigating to what extent criterial features and thus linguistic profiles per CEFR level may differ depending on the L1 of the learner. Is it the case that different groups of learners make use of different criterial features and as a result have a substantially different linguistic profile per CEFR level according to their L1? Is the description of *one single* general linguistic profile per CEFR level viable or should extensive reference be made to subgroups of L1s? These are issues currently under investigation in EP in line with SLATE's objectives (cf. research question 2 in Introductory chapter, this volume).

Moreover, the gradual transition in recent years from traditional to more diverse learning settings around the globe provides an additional impetus for the investigation of cross-linguistic differences across the CEFR levels. In a traditional learning setting (e.g. classroom), the teacher would typically address the needs of a homogeneous group; in a modern learning setting (e.g. an e-learning environment where learners from around the world meet virtually in chat rooms), teachers would cater to learners from a wide range of linguistic backgrounds with diverse learning needs. This gives rise to a new need for learners: personalised learning. Linguistic research on the effect of the L1 can inform teaching and address requirements for personalised learning, of increased importance in settings of globalised e-learning environments (Alexopoulou, Yannakoudakis, & Briscoe, 2010).

In summary, this section illustrated three principles that appear to drive SLA development in the CLC data. This is work in progress and current EP research is formulating further principles to account as comprehensively as possible for the emerging learning patterns identified in CLC (for more details see

Hawkins & Filipović, 2010). It needs to stressed, however, that we do not claim that any of the above principles alone (or any single learning principle for that matter) can fully account for SLA development and performance independently of one another and/or other factors. In fact, further EP research now focuses on identifying the interactions between these principles and their predictive power within a multi-factor SLA model (see *Conclusions and the way forward* below for specific proposals of how this investigation will take shape).

6. 'Criterial features' of English across the CEFR levels

The last two sections outlined some language features evident across the CEFR levels as illustrations for the L2 principles and patterns identified thus far. For a more comprehensive inventory of the criterial features identified so far (on the basis of the principles described above), the reader is referred to Salamoura and Saville (2009). A more complete list is currently developed by Hawkins and Filipović (2010). The list of criterial features has also been informed by earlier research on the properties of learner English at different learning stages, namely the T-series (*Breakthrough*, Trim, 2001; *Waystage*, *Threshold* and *Vantage*, van Ek & Trim, 1998a, 1998b, 2001). Researchers within the EP team are currently revisiting these publications in search of features that are novel at each level and that could thus qualify for the status of criterial features. Some preliminary results from this project are again provided in Salamoura and Saville (2009).

It should be stressed that these preliminary findings are a 'snapshot' of EP research as it stands at the time this chapter goes to press. It is expected that these findings will be refined, revised and complemented as more data become available and as more research is carried out.

7. Conclusion and the way forward

In summary, in this chapter we discussed EP's approach to profiling the CEFR levels – a search for criterial features of learner performance that distinguish the CEFR levels, informed by SLA theory and empirically derived from an extensive L2 English corpus, the CLC, using psycholinguistic and computational principles and metrics. The findings discussed in this chapter are mostly preliminary but they already reveal a promising picture of a learner's profile. A number of criterial features have emerged across the CEFR levels which show, from A2 to C2, an increasing progression in terms of frequency, syntactic and semantic complexity, and at the same time, a decreasing tendency for errors, non-native like usage and L1 influence. Critically, this progression can be systematically quantified and

measured against a large corpus of learner data. The emerging performance patterns per CEFR level are potentially highly informative for our understanding of the development of SLA, as they can inform us about the order of acquisition of linguistic features and elucidate the role and interaction of factors such as frequency, complexity and L1 transfer. It thus appears that the EP approach, which reflects closely SLATE's goals and objectives, is fulfilling the original EP aim, which is to produce an exemplification of the CEFR following SLA theory and at the same time shed light on questions and issues raised by SLA and psycholinguistic theory based on empirical (learner) data.

Although quite informative, these initial findings raise a number of further questions which EP researchers are currently addressing. These include:

- How do the different SLA patterns and principles that emerge from the study of CLC interrelate? This is a particularly important question as some of the emerging principles in CLC may be in competition with each other. Which competing principle takes precedence over the other? Hawkins and Filipoviç (2010) argue that the principles of frequency, complexity and L1 transfer can be incorporated within a multi-factor model of SLA and used to define possible versus impossible, and likely versus unlikely, IL stages. They propose to investigate the relative strength and interaction between principles by setting up a computer simulation that defines these possible/impossible and likely/unlikely IL stages in the manner of a complex adaptive system (Gell-Mann, 1992), and in the manner of Kirby's (1999) computer simulation of the emergence of possible/impossible and likely/unlikely word order variants, using the processing principles of Hawkins (1994).
- How do the different kinds of criterial features (lexical semantic, morpho-syntactic, syntactic, discourse, notional, functional, etc.) interrelate? In particular, which linguistic features realise which language functions across the CEFR levels? Answering these questions is fundamental in bringing together the different strands of EP research.
- To what extent does the criteriality of features vary depending on the L1 of the learner?
- Which criterial features can be used as diagnostics at the individual learner level?
- What is the effect of task type on learner production and criterial features? (Parodi, 2008)
- How does the type of context in which some linguistics properties (e.g., spatial verbs) occur help explain the emerging patterns in CLC (Hendriks, 2008)?

The immediate future of the EP will involve extending the current analyses to broader samples from the CLC and collecting other kinds of non-exam written data (e.g. from classroom settings) from learners of English worldwide. A major data collection exercise is currently being undertaken worldwide to this effect. Another major challenge being addressed is how to include *spoken language* in the analysis (McCarthy & Saville, 2009) in order to be able to describe a learner's linguistic profile at each CEFR level for speaking too. Such a profile will complement the current linguistic profile being investigated for writing (as defined in terms of lexico-semantic, morphosyntactic and syntactic features listed on pp. 6–7 and exemplified through the SLA principles identified thus far in the CLC data: MaF, MaS and MaPT). This would bring the CEFR profiling a step closer to what SLATE envisages as a complete linguistic description of the CEFR for all four skills (see research question 1 in Introductory chapter, this volume). Finally, another future aim is to collect data that will make it possible to foster a closer relationship between the EP outcomes and teachers/learners of English in their different contexts world-wide (Alexopoulou, 2008).

The method described in this chapter for profiling the CEFR ('criterial features', SLA theory- and corpus-informed empirical approach) was illustrated for the English language. However, once fully developed and established, it has the potential for application to any second language (Hendriks, 2008). Such a prospect would without doubt facilitate the comparison of CEFR profiles across different L2s as recommended by the SLATE network (see research question 2 in Introductory chapter, this volume).

It is envisaged that the description of English across the CEFR levels in terms of criterial features will result in a valuable data source for researchers and a useful tool for practitioners in the fields of English language learning, teaching and assessment. Moreover, as an outcome of the EP, it is hoped that the CEFR itself can be operationalised more effectively for English and that it will become a more useful tool for its intended purposes. The search for criterial features will lead to better *linguistic descriptions*, and this in turn will lead to better *functional descriptors*, thus addressing a current weakness (see Milanovic, 2009). Already the focus on empirical research at the bottom and top ends of the scale (A1, and C1/2) is providing more precise information about the nature of proficiency in English at these levels.

Acknowledgements

We are grateful to Prof. John Hawkins, Dr. Luna Filipović, Dr. Teresa Parodi, Dr. Henriette Hendriks and Prof. Roger Hawkey for their input in an earlier version of this chapter.

References

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26*, 261–278.

Alexopoulou, T. (2008). Building new corpora for English Profile. *Research Notes, 33*, 15–19.

Alexopoulou, T. (2010). *"She's great teacher": Article omission in predicative sentences or the non-linearity of L1 effects on English L2 predicative nominals*. Manuscript in preparation.

Alexopoulou, T., Yannakoudakis, H. & Briscoe, E. J. (2010). *L1 effects and personalised learning in globalised learning settings*. Manuscript in preparation.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–194). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy and cognition.* Cambridge: Cambridge University Press.

Briscoe, E. J. (2000). *Dictionary and system subcategorisation code mappings*. MS, Computer Laboratoty, University of Cambridge.

Briscoe, E. J., & Carroll, J. (1997). Automatic extraction of subcategorisation from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing* (pp. 356–363). Washington DC. Retrieved from http://delivery.acm.org/10.1145/980000/974609/p356-briscoe.pdf?key1=974609&key2=8461832721&coll=GUIDE&dl=GUIDE&CFID=88116625&CFTOKEN=35472625

Briscoe, E. J., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (pp. 77–80). Sydney, Australia. Retrieved from http://acl.ldc.upenn.edu/P/P06/P06-4020.pdf

Brooks, P., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology, 35*, 29–44.

Capel, A. (2009, February). *A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project*. Paper presented at the English Profile Seminar, Cambridge.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*, 161–170.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Diessel, H. (2004). *The acquisition of complex sentences.* Cambridge: Cambridge University Press.

Dodson, K., & Tomasello, M. (1998). Acquiring the transitive construction in English: The role of animacy and pronouns. *Journal of Child Language, 25*, 555–574.

Doughty, C. (1991). Second language instruction does make a difference. *Studies in Second Language Acquisition, 13*, 431–469.

Doughty, C. J., & Long, M. H. (Eds.). (2005). *The handbook of second language acquisition*. Malden, MA: Blackwell.

Dušková, L. (1983). On sources of errors in foreign language learning. In B. Robinett & J. Schachter (Eds.), *Second language learning: Contrastive analysis, error analysis, and related aspects* (pp. 215–233). Ann Arbor: University of Michigan Press.

Eckman, F., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics, 9*, 1–13.

Ellis, N. C. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention and the limited L2 endstate. In P. Robinson & N. C. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp. 372–405). New York: Routledge.

Ellis, N. C., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition, 19*, 145–171.

Gass, S. M. (1979). Language transfer and universal grammatical relations. *Language Learning, 29*, 327–344.

Gass, S. M. (1980). An investigation of syntactic transfer in adult second language learners. In R. Scarcella & S. Krashen (Eds.), *Research in second language acquisition* (pp. 132–141). Rowley, MA: Newbury House.

Gass, S. M., & Ard, J. (1984). Second language acquisition and the ontology of language universals. In W. E. Rutherford (Ed.), *Language universals and second language acquisition* (pp. 33–67). Amsterdam: John Benjamins.

Gass, S. M., & Selinker, L. (Eds.). (1992). *Language transfer in language learning.* Amsterdam: John Benjamins.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). New York / London: Routledge.

Gell-Mann, M. (1992). Complexity and complex adaptive systems. In J.A. Hawkins & M. Gell-Mann (Eds.), *The evolution of human languages* (pp. 3–18). Redwood City, CA: Addison Wesley.

Green, T. (2008). English Profile: Functional progression in materials for ELT. *Research Notes, 33*, 19–25.

Hawkins, J. A. (1994). *A performance theory of order and constituency.* Cambridge: Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars.* Oxford: Oxford University Press.

Hawkins, J. A. (2009). An efficiency theory of complexity and related phenomena. In D. Gill, G. Sampson & P. Trudgill (Eds.), *Complexity as an evolving variable* (pp. 252–268). Oxford: Oxford University Press.

Hawkins, J. A., & Buttery, P. (2009a). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In L. Taylor & C. J. Weir (Eds.), *Language Testing Matters: Investigating the wider social and educational impact of assessment* (pp. 158–175). Cambridge: Cambridge University Press.

Hawkins, J.A., & Buttery, P. (2009b). *Criterial features in learner corpora: Theory and illustrations.* Manuscript submitted for publication.

Hawkins, J. A., & Filipović, L. (2010). *Criterial features in the learning of English: Specifying the reference levels of the Common European Framework.* Manuscript in preparation.

Hendriks, H. (2008). Presenting the English Profile Programme: In search of criterial features. *Research Notes, 33*, 7–10.

Hulstijn, J. H. (2006, June). *Defining and measuring the construct of second language proficiency*. Plenary paper presented at the joint conference of AAAL and ACLA/CAAL, Montreal.

Jeon, K. S., & Kim, H-Y. (2007). Development of relativization in Korean as a foreign language: Noun phrase accessibility hierarch in head-internal and head-external relative clauses. *Studies in Second Language Acquisition, 29*, 253–276.

Jones, N. (2000). Background to the validation of the ALTE Can Do Project and the revised Common European Framework. *Research Notes, 2*, 11–13.

Jones, N. (2001). The ALTE Can Do Project and the role of measurement in constructing a proficiency framework. *Research Notes, 5*, 5–8.

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Case studies* (pp. 167–83). Strasbourg: Council of Europe. Retrieved from www.coe.int/T/DG4/Portfolio /documents/case_studies_CEF.doc

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry, 8*, 63–99.

Kellerman, E. (1977). Towards a characterization of the strategy of transfer in second language learning. *Interlanguage Studies Bulletin, 2*, 58–145.

Kellerman, E. (1978). Giving learners a break: Native language intuitions as a source of predictions about transferability. *Working Papers on Bilingualism, 15*, 59–92.

Kellerman, E. (1979). Transfer and non-transfer: Where are we now? *Studies in Second Language Acquisition, 2*, 37–57.

Kessler J-U. (2007). Assessing EFL-Development online: A feasibility study of Rapid Profile. In F. Mansouri (Ed.), *Second language acquisition research: Theory-construction and testing* (pp. 119–144). Newcastle: Cambridge Scholars.

Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals.* Oxford: Oxford University Press.

Korhonen, A., Krymolowski, Y., & Briscoe, E. J. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC06), Genova, Italy. Retrieved from http://www.cl.cam.ac.uk/~alk23/lrec06-lexicon.pdf

Kurteš, S., & Saville, N. (2008). The English Profile Programme: An overview. *Research Notes, 33*, 2–4.

Larsen-Freeman, D. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning, 26*, 125–134.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*, 645–55.

MacWhinney, B. (1987a). Applying the competition model to bilingualism. *Applied Psycholinguistics, 8*, 315–327.

MacWhinney, B. (1987b). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249–308). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1992). Competition and transfer in second language learning. In R. J. Harris (Ed.), *Cognitive processing in bilinguals* (pp. 371–390). Amsterdam: Elsevier Science.

MacWhinney, B. (1997). Second language acquisition and the Competition Model. In A. de Groot & J. Kroll (Eds.), *Tutorials in bilingualism* (pp. 113–142). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008). A Unified Model. In N. Ellis & P. Robinson (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). New York: Routledge.

McCarthy, M., & Saville, N. (2009, March). *Profiling English in the real world: What learners and teachers can tell us about what they know*. Paper presented at the American Association for Applied Linguistics Conference, Denver, Colorado.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes, 37*, 2–5.

Nicholls, D. (2003). *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. Retrieved from http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition, 25*, 433–448.

Odlin, T. (2005). Cross-linguistic influence. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 436–489). Malden, MA: Blackwell.

Ozeki, H., & Shirai, Y. (2007). The consequences of variation in the acquisition of relative clauses: An analysis of longitudinal production data from five Japanese children. In Y. Matsumoto, D. Oshima, O. Robinson, & P. Sells (Eds.), *Diversity in language: Perspectives and implications.* Stanford, CA: CSLI Publications.

Parodi, T. (2008, December). *L2 morpho-syntax and learner strategies*. Paper presented at the Cambridge Institute for Language Research Seminar, Cambridge, UK.

Pienemann, M. (1992). COALA – A computational system for interlanguage analysis. *Second language Research, 8*, 59–92.

Pienemann, M., Di Biase, B., Kawaguchi, S., & Håkansson, G. (2005). Processing constraints on L1 transfer. In J. F. Kroll & A. M. B. DeGroot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 128–153). New York: Oxford University Press.

Pienemann, M., & Keßler J-U. (2007). Measuring bilingualism. In P. Auer & L. Wei, (Eds.), *Handbook of multilingualism and multilingual communication* (pp. 245–275). Berlin: Mouton de Gruyter.

Preiss, J., Briscoe, E. J., & Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (pp. 912–919), June 2007, Prague, Czech Republic. Retrieved from http://www.cl.cam.ac.uk/~alk23/acl-07.pdf

Salamoura, A. (2008). Aligning English Profile research data to the CEFR. *Research Notes, 33*, 5–7.

Salamoura, A., & Saville, N. (2009). Criterial features across the CEFR levels: Evidence from the English Profile Programme. *Research Notes, 37*, 34–40.

Schachter, J. (1974). An error in error analysis. *Language Learning, 24*, 205–214.

Schwartz, B., & Sprouse, R. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research, 12*, 40–72.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Taylor, L. (2004). Issues of test comparability. *Research Notes, 15*, 2–5.

Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). *Research Notes, 24*, 2–5.

Tomasello, M. (2001). The item-based nature of children's early syntactic development. In M. Tomasello & E. Bates (Eds.), Language development: The essential readings (pp. 169–202). Oxford: Blackwell Publishing.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.

Tomasello, M., & Brooks, P. (1998). Young children's earliest transitive and intransitive constructions. *Cognitive linguistics, 9*, 379–395.

Trim, J. L. M. (2001). *Breakthrough.* Unpublished manuscript. Retrieved from www.englishprofile.org

Van Ek, J., & Trim, J. L. M. (1998a). *Threshold 1990.* Cambridge: Cambridge University Press. (Original work published 1991).

Van Ek, J., & Trim, J. L. M. (1998b). *Waystage 1990.* Cambridge: Cambridge University Press. (Original work published 1991).

Van Ek, J., & Trim, J. L. M. (2001). *Vantage.* Cambridge: Cambridge University Press.

Williams, C. (2007). *A preliminary study into the verbal subcategorisation frame: Usage in the CLC.* Unpublished manuscript, RCEAL, Cambridge University, UK.

Young, R. (1991). *Variation in interlanguage morphology.* New York: Peter Lang.

Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French

Fanny Forsberg & Inge Bartning
Stockholm University

The study presents results from an ongoing project which tries to match communicative abilities as proposed in the CEFR- scale with the development of linguistic proficiency features. The main purpose is to look for linguistic features of each CEFR-level (A1-C1), in terms of morpho-syntax, discourse organisation and use of formulaic sequences. The selected linguistic phenomena have already been shown to be 'criterial' for acquisitional orders in oral L2 French. To this end, written data have been collected from 42 Swedish university students of L2 French. The method implied included placement of the students on CEFR.scales by DIALANG, production by the students of written argumentative texts and summaries according to CEFR-criiteria criteria, raters' judgements and, finally, narrow linguistic analysis of the same productions.
The first results show that

- Measures of morphosyntactic deviances yield significant differences between the CEFR-levels up to B2
- Links can be observed between already established late acquisitional features, like *gérondif*, *dont* and *plus-que-parfait*.

Use of lexical formulaic sequences increases at higher CEFR-levels, but significant differences were found only between A2/B2/C2.

1. Introduction

This chapter presents results from an ongoing project which investigates whether it is possible to find correlations between the development of certain linguistic interlanguage features and the proposed language proficiency levels of the CEFR scale (Council of Europe, 2009). The linguistic phenomena investigated in this study are morpho-syntax (NP and VP morphology), discourse (discourse markers and subjunctions) and the use of formulaic language.

This linguistic analysis has been motivated by the model of six developmental stages of morphosyntax and discourse in oral French as presented in Bartning and Schlyter (2004, p. 282) which was elaborated as an empirically

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 133-158

based bottom-up construct (see Appendix). This study presented results from work using the InterFra corpus, Stockholm University (Forsberg, 2008; Hancock, 2000; Kirchmeyer, 2002) (http//:www.fraita.su.se.interfra) and the Lund corpus (see Granfeldt, 2003; Schlyter, 2003). The theoretical underpinnings of the Bartning and Schlyter model (2004, p. 281) include work by Klein and Perdue (1997, the ESF programme with 5 target languages including French L2), studies of grammaticalisation processes (form/function relations, see Bybee & Hopper, 2001) and processability theory (Pienemann, 1998).

The theoretical perspective in this chapter is guided by work on developmental stages (Bardovi-Harlig, 2006; Bartning & Schlyter, 2004; Sharwood-Smith & Truscott, 2005) and formulaic language (Erman & Warren, 2000; Wray, 2008). For earlier studies concerning acquisition and learning routes, esp. in Europe, see the ESF project (Perdue, 1993) and the Pavia project for Italian L2 (Giacalone Ramat, 1992). For studies on accuracy in the tradition of CAF (complexity, accuracy and fluency), see e.g. Van Daele, Housen, Kuiken, Pierrard, and Vedder (2007).

The choice of the three linguistic domains, viz. morphosyntax, discourse phenomena and formulaic language, is motivated by the fact that they have been shown also in recent studies on the InterFra corpus (Bartning, Forsberg, & Hancock, 2009; Bartning & Hancock, in press) to be discriminators in the development of L2 French. The morpho-syntactic areas concern verbal and nominal morphology (see e.g. overviews Ågren, 2008; Granfeldt & Nugues, 2007; Herschensohn, 2006; Véronique, 2009).

Some researchers consider developmental stages as one of the main findings in SLA (see Ellis, 2008, p. 72; Long, 2009), others have recently started to question them (see Hulstijn, this volume; Larsen-Freeman, 2006). This scepticism is thus expressed by Larsen-Freeman (2006) in her article about an emergentist perspective in SLA. She proposes individual profiles permitting great variation with many paths to development suggesting that development of learner language is not discrete and stage-like but more like 'the waxing and waning of patterns', (p. 590). The perspective of interlanguage as a dynamic process is, however, caught by the well-found terms by Bardovi-Harlig (2006, p. 69), 'main routes' and /or 'individual paths' according to different sources of influences.

One of the main aims of SLATE (see Introductory chapter) is to relate communicative development, as expressed by the CEFR-scale, to linguistic development, where the different chapters show examples of various linguistic domains and structures. However, it is important to stress already at this stage that the present study does not make a clear-cut distinction between communicative and linguistic development, since our CEFR-raters use both the communicative criteria stated in general proficiency scales (Finnish National Certificates of Language Proficiency, based on the CEFR general proficiency

scales) and the more language-oriented criteria presented in the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe, 2009). These latter criteria are quite general and vague, however, and do not make reference to language-specific structures, in this case French. One example goes as follows: "Maintains consistent *and highly accurate* grammatical control of *even the most complex language forms. Errors are rare and concern rarely used forms* (Council of Europe, 2009, p. 187). Accordingly, our study presents linguistic profiles of written productions that have been rated as belonging to the same general CEFR-levels, based on communicative and/or language-oriented criteria (the two different scales mentioned above). Our aim is thus mainly to map what language-specific features characterize the CEFR-levels in written L2 French.

To our knowledge there has been little earlier work on French L2 concerning the relation between the pragmatic functions described in the CEFR levels and corresponding developmental linguistic features. However, for other L2s such as English, Finnish, Italian, and Dutch there are now several studies as illustrated by this volume (cf. Kuiken, Vedder, & Gilabert, this volume; Martin, Mustonen, Reiman, & Seilonen, this volume; Pallotti, this volume).

Following the reasoning above, our two research questions are:

1. Is it possible to establish linguistic developmental features to match the general communicative CEFR levels? And, if so,
2. How do the proposed linguistic domains of morphosyntax (Bartning & Schlyter, 2004, VP and NP morphology), discourse markers/subjunctions (Hancock, 2000, 2007) and formulaic sequences (Forsberg, 2008) develop along the CEFR-levels?

In this study we focus on written data since there has not been as much research carried out on written French L2 as on spoken French L2, although some studies have worked with written corpora (Ågren, 2008; Bolly, 2008; Granfeldt & Nugues, 2007; Granger, Hung, & Petch-Tyson, 2002) but not in relation to the CEFR scale. In addition, written data is less time-consuming to collect than oral data and it was therefore agreed upon in the SLATE-group that it would be convenient to start off with written production. We stress the point that in order to relate the communicatively defined levels of CEFR to linguistic development, it is appropriate to start with testing students' communicative abilities according to the CEFR levels (cf. Hulstijn, 2006; Martin et al., this volume) and then analyse their productions in terms of linguistic categories.

The general plan of the chapter is as follows: methodological issues are discussed in section 2, the main results from the three domains under investigation, viz. morpho-syntax, discourse markers and formulaic sequences, are shown in sec-

tion 3 followed by the conclusion, section 4. Section 3.1 investigates the non-target-like forms of morpho-syntactic categories drawn from the stages of Bartning and Schlyter and adjusted for written French. The second part of this investigation (section 3.2), which is considered as a qualitative complement to the first part, concentrates on the emergence of a limited number of morpho-syntactic and discursive target features. The third part of the analysis (3.3), focusses on formulaic language and is a quantitative study of the use of a specific category of formulaic language, viz. lexical formulaic sequences, which has been shown to be successful at discriminating between developmental levels (cf. Forsberg, 2008).

2. Methodological issues

2.1 Participants, data collection and tasks

The participants were recruited among students of French at Stockholm University, from various levels during 2007-2008. Most participants were 1st or 2nd term students of French, which explains the fact that many of the participants are placed at the B1 level and that fewer participants are to be found at the highest CEFR levels according to the DIALANG test (see below).

The students (N= 42) were gathered in the computer room of the language laboratory at Stockholm University, where they were asked to perform three tasks during the course of 2-2.5 hours. Time constraints were thus not entirely rigid, but no one was allowed to spend more than 2.5 hours on all of the tasks. The participants were not allowed to use any aids, such as dictionaries or grammar books, and the spell and grammar check had been deactivated on the computers used for the tasks.

The tasks

1) Participants were placed at the CEFR level by the DIALANG test, but only using the sub-test devoted to written production. In order to receive a CEFR level from the DIALANG test, the test taker had to first of all pass the vocabulary placement test, then take the self-assessment test, before finally taking the diagnostic test measuring written production skill. The level of the diagnostic test reported by the student is the level taken into consideration when administering the tasks to the students.
2) Participants then performed two written tasks – one which was given to all levels, and one which was specific to their estimated CEFR level.

The written tasks were developed by the authors of this article and were modelled on the tasks in the Cefling project (see Alanen, Huhta, & Tarnanen, this

volume; Martin et al., this volume). They were also developed, as were the Cefling tasks, in order to test communicative abilities as stated in the CEFR descriptors, such as expressing personal views and summarizing. The tasks were thus not designed to trigger some specific linguistic features, but rather, to test communicative abilities. Text length was not indicated, but there was a time limit for the whole set of tasks as indicated above.

Task 1: Given to all levels
Subject: Write a summary of a film that you have seen or a book that you have read recently.

Task 2: Specific for each CEFR level
A2: Write an e-mail to a friend and tell him/her about what you did last weekend.
B1: Argumentative/personal task: Why is it important to learn French?
B2: Argumentative/topic-based task: Can the individual do anything to counter the climate threat?
C1: Genre-specific task: Write a letter of complaint to the "Préfecture de police" regarding permission to stay in France.
C2: Genre-specific task: Write an application letter to a university/school/publisher in France.
(The six CEFR levels are A1, A2, B1, B2, C1 and C2.)

2.2 The rating procedure

The rating procedure selected for this particular study needs to be briefly discussed. Because there is risk of circularity, Hulstijn (this volume) claims that it is not appropriate to use the linguistic scales of the CEFR in the rating procedure if the aim is to relate the *communicative* CEFR levels to *linguistic* features of development. His position is theoretically laudable, but in our view, it entails practical limitations. Which professional language proficiency rater – be he/she trained in the CEFR or not – will not take linguistic form into account at some level, especially in the written modality? Can a person judge a written text, without noticing, for instance, grammar and orthography (cf. Alanen et al., this volume)? For a study aiming at teasing apart communicative adequacy (according to the CEFR) and linguistic complexity (both according to raters' perception and according to general CAF measures), see Kuiken et al. (this volume).

We have come to the conclusion that it is possible to use raters, such as ours, who take both communicative function and linguistic form into account when rating a task (see criteria below), especially since the trained raters themselves propose this procedure. However, we do not know *which* linguistic features (morphosyntax, discourse and formulaic sequences) seem to discriminate between the

CEFR-levels, determining whether a learner is placed at a B1 level and not a C1 level and this will hopefully constitute the main contribution of the present study.

After the data were collected, the tasks were rated by professional CEFR raters, in order to verify whether the students had, in fact, performed at the CEFR-level for which they had been tested. One main rater of French at the University of Jyväskylä rated all 83 productions. (Unfortunately experienced CEFR raters are rare in Sweden and the Jyväskylä rater was recommended by colleagues.) For some productions, in case of difficult rating decisions, a second rater (also from Finland) was involved to ascertain the main rater's decision (this procedure was suggested by the main rater, as it corresponded to their practices). The raters were asked to use their regular rating practices, which involved making use of the following criteria:

1. *Finnish National Certificates of Language Proficiency* (based on the CEFR levels). In Finland, extensive work has been done at e.g. the University of Jyväskylä to align the national language tests with the CEFR (cf. Alanen et al., this volume). As a result, an evaluation scale, based on the six CEFR levels, is used when testing both Finnish as a second language and modern languages in school and higher education. Given their closeness to the CEFR scale, the raters could thus make use of the criteria stated for each level.

The level is described holistically, taking into account most language skills such as comprehension, writing and speaking. A few sentences are devoted specifically to written production, but they are quite general e.g. "is able to write both private and semi-official texts and to express thoughts as a coherent whole." (http://www.jyu.fi/hum/laitokset/solki/yki/english/about/skill_level/).

2. The other criteria are taken from *"Relating Language Examinations to the CEFR – Written assessment criteria grid* (http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp). Besides the criteria from the Finnish National Certificate, the raters also referred to the manual provided by the Council of Europe (2009), which contains more detailed linguistic criteria, developed in order to facilitate the raters' work.

However, it is precisely these criteria that are regarded as problematic by the SLATE-group, since they are not empirically validated through second language acquisition research. Below follows an example from the written assessment criteria grid, B1 level, overall rating (Council of Europe, 2009, p. 187):

> "Can write straightforward connected texts on a range of familiar subjects within his field of interest [1], by linking a series of shorter discrete elements into a linear sequence [2]. The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading."

As indicated by this citation raters thus use both holistic and general criteria [1] as well as criteria that contain more linguistic specifications [2].

However, the more specific criteria do not include any particular linguistic structures to which the raters should pay attention, let alone any structures in French, but they do direct the rater's attention, at least to some extent, to the linguistic form of expression and not only the communicative value of the text.

Following the rating procedure described and discussed above, the 83 productions by the 42 writers (one writer produced only one task), were grouped as shown in Table 1 below. As stated above, most productions were found to be at the B1 level. For statistical comparisons to be made between the levels, only levels A2, B1, B2 and C2 had enough values for the statistical analysis to be performed, hence their marking in bold.

Table 1. The CEFR levels of the productions after the rating

CEFR level (N: participants)	N: productions	N: words	Mean number of words/scriptor
A1 (1)	2	76	38
A2 (6)	12	1831	152
B1 (22)	43	11890	276
B2 (8)	16	5632	352
C1 (2)	4	1399	349
C2 (3)	6	2620	437
Total (42)	83	234 448	

3. Analysis: Linguistic criteria in the written productions

As stated in the introduction, potentially discriminatory features have been selected based on the results of many years of research on French as a second language, e.g. the work of the InterFra-project in Stockholm and the Corpus Lund at Lund University as shown in Bartning and Schlyter (2004). Besides the work on French morphosyntax, the InterFra-project has also been investigating the development of certain discursive features such as connectors (Hancock, 2007) and also the development of lexical competence as manifested in formulaic language (Forsberg, 2008). These different linguistic phenomena, which have to date been shown to be fruitful measures for the description of oral L2 French development, have now been investigated in the present *written* corpus.

Other studies which have independently confirmed the proposed six stages by Bartning and Schlyter (2004) are Housen, Kemps, and Pierrard (2008, oral data), Labeau (2009, written and oral data) and Bolly (2008, written data) for the advanced levels (stages 4-6), and Granfeldt and Nugues (2007) and Ågren (2008) for stages 1-4 in written production. Véronique (2009) also refers to the stages of Bartning and Schlyter (2004) when describing the development of French IL grammar in written and oral data.

It may seem methodologically problematic to apply criteria used in an oral corpus to a written one, but as the studies cited above on written French L2, e.g. Granfeldt and Nugues (2007) and Ågren (2008), have already shown, the criteria proposed in Bartning and Schlyter (2004) work surprisingly well even for written French.

The main criteria for establishing the developmental stages were the following: utterance structure, finiteness, verb morphology, subject – verb agreement, tense, mode and aspect (TMA), negation, noun phrase morphology, gender marking and discourse phenomena. (See Appendix, for a presentation of criteria and stages; and for the InterFra corpus, see http://www.fraita.su.se/interfra).

It is important to take into account that the morpho-syntactic stages proposed in Bartning and Schlyter (2004) were not originally meant to be related to the CEFR-levels. It may be tempting, however, to simply map the two scales onto each other, since they both contain six levels or stages, but this cannot be done automatically. The CEFR-scale is developed for all language skills, is supposed to be language independent and is based on teachers' experiences, ranging from the earliest levels to the most advanced level, the levels all being hypothetical. The Bartning and Schlyter stages, on the other hand, are also hypothetical as stages, but at the same time they are empirically based clusters of developmental features / sequences *(itinéraires acquisitionnels*) of two oral corpora of French interlanguage, which obviously offers advantages and disadvantages compared to the CEFR-scale. The obvious advantage is the argument of objectivity; they are based on how actual L2 production develops, without involving personal experiences. The most obvious disadvantage is that they are limited by the corpora and the levels attained by the learners in these corpora. As a consequence, the most advanced level in the Bartning and Schlyter continuum corresponds to the most advanced learners in the corpora at hand, but not necessarily to the most advanced levels of the CEFR-scale (C1-C2).

As already stated, section 3.1 below concentrates on the non-target-like forms of the categories taken from the stages of Bartning and Schlyter (2004) and adjusted for written French. They are here called morpho-syntactic deviances (MSDs). This first analysis of the 83 written productions is a quantitative morpho-syntactic one that takes into account a number of features. The second part of the study (section 3.2) has isolated a limited number of morpho-

syntactic and discursive target-like features, based on the InterFra corpus, and tries to trace the emergence of the features and their use at the different CEFR levels.

The third part (3.3) focusses on formulaic language and offers a quantitative study of the use of a specific category of formulaic language, viz. lexical formulaic sequences, which have been shown to be successful at discriminating between levels, in a number of studies such as Forsberg (2008) and Bartning et al. (2009). These well-documented features discriminating among linguistic levels are now related to the CEFR-based written productions of the study and thus to SLATE work.

3.1 Morpho-syntactic deviances

3.1.1 Presentation of the morpho-syntactic categories

In order to combine the CEFR levels with grammatical development in French L2, we propose the areas of developmental features of NP and VP morphology. In our study we have made a first screening of these features in written L2 French. Space limits do not permit us to problematize the method of identification of these MSDs (morphophonological rules, audible/non audible oppositions, orthography etc.). Nevertheless, we have indicated in the classification below non-target-like/target-like oppositions.

Presentation of the morpho-syntactic categories of deviances (MSDs):

VP morphology

1. Subject-verb agreement, opposition plural/singular in person and number: *ils *sort* (TL (=Target Language): ils sortent), *j'*a* (TL: j'ai); *ils *a* (TL: ils ont),
2. Subject-verb agreement, opposition in person, number: *ils *parle* (TL: ils parlent), *je *peut* (TL: je peux), *il *peux* (TL: il peut), *c'est nous qui *pouvont* (TL: pouvons)
3. Tense, Mode and Aspect (TMA) simplification: the present tense form instead of the *passé composé* (PC), etc. PC instead of *plus-que-parfait* (PQP), non-finite forms instead of finite: *je *donnE*, or the opposite: a finite form in stead of a non finite form: *je peux *parle* (TL: parler); *pour *s'occupaient* (TL: s'occuper)
4. TMA subjunctive: *il faut que *j'ai* (TL: j'aie), *que tu *as* (TL: tu aies)
5. TMA auxiliary: *j'*ai tombé* (TL: je suis tombé)
6. Clitic object: *je *lui aide* (TL: je l'aide) (In this study classified under VP since the choice of the object depends on the verb construction, *aider qn*, e.g.)

NP morphology

7. Gender and number on subject personal pronoun: *ils, elles, il/elle* (and their NTL variants)
8. Gender on definite article: **la père* (TL: le)
9. Gender on indefintie article: **un fille* (TL: une)
10. Number on nouns: *les *fille* (TL: les filles)
11. Gender, number on audible attributive adjectives: *une femme *fort* (TL: forte)
12. Gender, number on non-audible attributive adjectives: *une *joli fille* (TL: jolie), *des *joli* (TL: jolies) *filles*
13. Gender, number on audible predicative adjectives: *la fille est *fort* (TL: forte)
14. Gender, number on non-audible predicative adjectives: *la fille est *joli* (TL: jolie)
15. Naked nouns (without obligatory determiners): **liberté* (TL: la liberté)

3.1.2 Morpho-syntactic deviances (MSD) at the different CEFR levels

Table 2 below shows the raw figures of MSDs in the 83 written productions rated at the different CEFR-levels. A distinction has also been made between deviances belonging to the VP and the NP respectively. It becomes quite clear that most of the MSDs are found within the NP even in this written learner corpus, esp. A1-A2, B1 and B2 levels (as in the oral very advanced sub-corpora of InterFra, see Bartning et al., 2009). This tendency of many MSDs in NP morphology in written French also confirms the results of Ågren (2008).

Table 2. Results of morpho-syntactic deviances at six CEFR levels

CEFR-level (N=productions)	VP Total	NP Total	Total MSD	Total No of words
A1-A2 (14)	41 (33%)	83 (67%)	124	1.907
B1 (43)	74 (27%)	205 (73%)	279	11.890
B2 (16)	21 (30%)	48 (70%)	69	5.632
C1 (4)	14 (45%)	17 (55%)	31	1.399
C2 (6)	7 (41%)	10 (59%)	17	2.620
Total (83)				

Whereas table 2 above shows the raw figures of MSDs, table 3 below shows the mean values of MSDs / 100 words at four levels. The figures clearly indicate that the MSDs decrease with higher levels of CEFR, indicating an increase in

students' accuracy. The four selected levels in table 3 (A2, B1, B2, C2) contained sufficiently many productions to be submitted to statistical tests, while productions at the A1 and C1 levels did not[1].

Table 3. Mean number of morpho-syntactic deviances/ 100 words

CEFR-level	Mean value
A2	8.6
B1	3.2
B2	1.2
C2	0.4

Table 4 below shows statistically significant differences between five of the CEFR levels concerning morpho-syntcatic deviances.

Table 4. Statistical results for the comparison between different CEFR-levels (One way ANOVA, Tukey-Kramer post-hoc test)

CEFR-level	P-Values
A2>B1	0.001
A2>B2	0.001
A2>C2	0.001
B1>B2	0.01
B1>C2	0.01
B2>C2	Not significant

The results are interesting, since statistical differences are found up to the B2 level, but not between B2 and C2 (Table 4). However, there is an important difference in the mean values between B2 and C2, namely that the number of MSDs diminishes (see Table 3), although the difference was not significant.

Furthermore, one interpretation of the results might be that morpho-syntactic development, as manifested in quantity of MSDs, reliably discriminates

1 The software used for the statistical analysis, GraphPad InStat, did indicate that there were too few values in these given groups for the analysis to be performed.

Table 5. Quantitative results concerning the developmental features in the 83 productions at the six CEFR levels (raw and relative frequencies)

CEFR level	Total words	*dont*	*ce qui*	*ce que*	Gerund	Pluperf	Subj *soit, soient*	*donc*	*mais*	*parce que*	*pourtant*	*puisque*	*en effet*
A1	76	0	0	0	0	0	0	0	1	0	0	0	0
Token/total									1.32%	0.00%			
A2	1831	1	0	0	0	0	0	1	20	17	0	0	0
		0.05%						0.05%	1.09%	0.93%			
B1	11890	3	11	10	13	1	3	17	106	22	8	2	1
		0.03%	0.09%	0.08%	0.11%	0.01%	0.03%	0.14%	0.89%	0.19%	0.07%	0.02%	0.01%
B2	5632	3	8	5	12	0	4	6	40	7	4	1	1
		0.05%	0.14%	0.09%	0.21%	0.00%	0.07%	0.11%	0.71%	0.12%	0.07%	0.02%	0.02%
C1	1399	1	2	1	5	0	0	2	7	1	0	0	4
		0.07%	0.14%	0.07%	0.36%	0.00%	0.00%	0.14%	0.50%	0.07%	0.00%	0.00%	0.29%
C2	2620	5	3	1	5	4	1	4	16	1	3	3	1
		0.19%	0.11%	0.04%	0.19%	0.15%	0.04%	0.15%	0.61%	0.04%	0.11%	0.11%	0.4%
Total words	**234 448**												

between CEFR-levels up to level B2. Other linguistic criteria and a larger corpus are probably necessary to characterize the differences between the highest levels (B2-C1-C2).

It is interesting to note that Martin et al. (this volume) also see important differences between A2 and B1, suggesting that the step to B1, The Independent User, is an important one for the learner to take.

With the cautionary statements above in mind, we think that the results are nonetheless thought-provoking because they suggest that morpho-syntax does not develop beyond a certain CEFR-level of written L2 French. One could perhaps propose that problems in morphosyntax are stabile and do not altogether disappear. This result seems to concur with recent results proposed by Bartning et al. (2009), even though the results from this last study of late oral French were not matched against the CEFR-scale. However, it did show that three different learner groups, all highly proficient, two being resident in the TL country and one being in a foreign language setting, did not differ significantly with respect to morpho-syntactic deviances. Surprisingly, the MSDs persisted through these high stages (according to the literature near-native speakers' grammar is native-like at very advanced stages, cf. von Stutterheim, 2003).

To sum up, morphosyntax seems to develop along the CEFR-scale up to the level B2, as measured by frequency of grammatical deviances. Our question now is, of course, what other specific grammatical and discursive criteria could be tested against the CEFR-levels. This question will be qualitatively explored in the following section.

3.2 Candidates for indications of tendencies of developmental features

As stated in the introduction, the selection of morpho-syntactic and discursive features (here: discourse markers (e.g. *donc*), and subjunctions (e.g. *puisque*)), is based on features typical of the development of French interlanguage in the SLA literature (for an overview, cf. Bartning, 2009; Herschensohn, 2006). Many of them are presented in the appendix. Among relative pronouns we find: *dont, ce que, ce qui* (Bartning & Schlyter, 2004; Flament-Boistrancourt, 1984; Hancock & Kirchmeyer, 2005); the TenseModeAspect category is represented by the pluperfect (Bartning, 2009; Howard, 2009) and the subjunctive (Bartning, in press; Howard, 2008): *qu'il soit, ils soient.* At higher levels gerund emerges (Kirchmeyer, 2002): *venant, en venant.* Finally, we find connectors and subjunctions, as both late and early features, such as *donc, pourtant, puisque, en effet, mais* and *parce que* (Hancock, 2000, 2007; Kirchmeyer, 2002). These features have been examined in 83 productions in a pilot study. The results in table 5 below show the raw figures and percentages of the frequency of the use of these different phenomena.

Let us now consider Table 5.

Concerning the relative pronoun *dont* (column 3), according to earlier research on French L2 in *oral* productions *dont* is acquired late. In Bartning and Schlyter (2004) it does not show up until stages 5-6, if at all (see Appendix). It is thus interesting that it appears more frequently from level B1 (even A2) and upwards here. *Dont* is an example of condensed syntax which is expected to turn up late in IL (cf. written L2 French: Flament-Boistrancourt, 1984; oral L2 French: Hancock & Kirchmeyer, 2005).

It is also interesting to see that the other relative pronouns as *ce qui, ce que* do not appear until B1, B2. These pronouns refer anaphorically to anterior clauses in utterances and thus presuppose complexity in utterance building, a feature that belongs to higher acquisitional levels (Kirchmeyer, 2002).

The gerund is also a late feature in oral production (Bartning & Schlyter, 2004, Table 3, p. 294). This construction is also a manifestation of complex and condensed syntax which belongs to high levels in oral proficiency. This has been shown by others, e.g. Kirchmeyer (2002). Interestingly, as the data presented in Table 5 show, we see that the gerund turns up in several productions from B1 level and upwards. It is not surprising that it appears here in written productions since it belongs to written genres more than to oral. In any case, it appears to discriminate between the A levels, which show no appearances, and the other levels, which show surprisingly many. In the Bartning and Schlyter (2004) oral corpora the gerund was extremely rare, with only occasional uses at stages 5-6. It would be interesting to investigate the use of the gerund in a larger corpus of spontaneous non-native and native productions, oral and written. A relevant factor which could be revealing is the explicit/implicit dichotomy, as written language invites the learner to reflect on his language and time permits metalinguistic control. This issue will be explored in future studies focusing on the differences between oral/written French interlanguage.

The less frequent feature of our illustrations is the use of the pluperfect and, if it is used at all, it occurs at the highest levels. This pattern follows earlier studies of French interlanguage, e.g. works by Howard (2009) and Bartning (2009).

The subjunctive verbal forms *soit, soient* appear also at B1. This is an acknowledged late feature in oral French according to Bartning and Schlyter (2004) and Howard (2008).

The selection of discourse markers and subjunctions has been made in order to be independent of text genres, such as narrative and argumentative texts. As table 5 (above) also shows, *mais* and *parce que* turn up already as connectors at A1-A2 as in early oral French L2 (cf. Hancock, 2000). At levels A2-B1 there is a remarkable increase of *mais* and then a decrease at B2-C2. This result also reflects tendencies already found in oral French IL: the well-known overuse of *mais* in the rather limited repertoire of connectors in early IL. The

counts for *parce que* reveal an increase at A2 with very few occurrences at C1-C2, the hypothesis being that *parce que* is taken over by the use of other causal connectors such as *puisque* etc. Table 5 also clearly shows that the repertoire of different connectors (*donc*, *pourtant*, *pusique* and *en effet*) grows at B1 and the writers have access to more than just two connectors *mais* and *parce que*.

Table 5 also informs us that the grammatical structures *dont*, *ce que*, *ce qui*, the gerund, the pluperfect and the subjunctive, as well as the remaining connectors *donc, pourtant*, *puisque* and *en effet,* appear at all levels from B1. None of these phenomena, grammatical or discursive, turns up at the earlier CEFR levels A1-A2 (with the exception of two occurrences of *dont* and *donc*). They all belong to more elaborated language.

We now turn to the third domain of linguistic features to be investigated in this chapter in our search for potential candidates of developmental measures along the CEFR levels, namely, the use of formulaic language. In the conclusion, results from the investigations of the MSDs, discursive phenomena and formulaic language will be viewed together and be related to the CEFR levels.

3.3 Formulaic language: Lexical formulaic sequences in relation to the CEFR-scale

Formulaic language, such as collocations, idiomatic expressions and social routines, are known to be a stumbling block for second language learners and users (Schmitt, Grandage, & Adolphs, 2004; Wray, 2008). However, 'formulaic language' or 'formulaic sequences' is sometimes also used in the SLA literature to refer to unanalyzed sequences that help beginner learners to communicate before they master creative rules. Sometimes these formulaic sequences correspond to target-like formulaic sequences, such as 'You're welcome', but they can also be sequences which are only unanalyzed in the learner's production such as *Monique *j'habite* (Monique *I lives), in Myles, Hooper, and Mitchell's (1998) study. In the present study, these latter sequences will not be treated since the study only takes into account target-like, idiomatic sequences. Furthermore, "Idiomatic expressions and colloquialisms" are also mentioned in the Written Assessment Criteria grid of the CEFR (Council of Europe, 2009, p. 186) as a feature which is not mastered before the C2 level (the most advanced CEFR-level). This encouraged us even more to test this criterion on written L2 development.

Erman and Warren (2000) presented a taxonomy of formulaic sequences, where they were divided into Lexical, Grammatical and Discursive prefabs, based on their main function in language use. Forsberg (2008) and Lewis (2008) applied this taxonomy to second language data, French and English respectively, and both found that the Lexical prefabs were the sequences causing difficulties for second language learners. They were thus shown to be an efficient yardstick of second language proficiency, especially when distinguishing

between advanced formal learners and very advanced second language users residing in the TL community. Consequently, we decided to see how this category is used at the different CEFR levels.

3.3.1 Identification and classification of FSs

The two most commonly used ways of identifying formulaic language in corpora are the statistical method and the phraseological method (cf. Granger & Pacquot, 2008). The first method identifies recurrent sequences based on different statistical measures such as *log likelihood* and *Mutual information score* (MI). Words that occur together more often than predicted by chance are labelled as collocations (a category of formulaic language). This method is automatic and involves no human judgement. Within the phraseological methodology, on the other hand, the researcher identifies potentially formulaic/conventional sequences based on linguistic criteria related to syntactic, semantic and pragmatic restrictions.

In view of the small size of our corpus and the fact that the statistical method requires a large corpus, it was decided that a phraseological method would be more appropriate for this study.

The present study makes use of Erman and Warren's (2000) original categorisation of prefabs (their term) which was slightly modified in Erman, Forsberg, and Fant (2008). Sequences can thus be categorized into Lexical, Grammatical and Discursive types.

Only the Lexical FSs category will be presented in more detail here. For an overall presentation of the categorisation, see Forsberg (2008) or Forsberg (2010).

Lexical FSs:	Clausal:
	je vous en prie ('you're welcome')
	c'est pas grave ('that's OK')
	métro, boulot, dodo (no equivalent)
	Phrasal:
	agréablement surpris ('positively suprised')
	faire du sport ('practice a sport')
	poser une question ('pose a question')

Lexical FSs incorporate at least one content word. They are used for extralinguistic reference (as opposed to grammatical and discursive FSs) and denote actions (such as *faire la fête* 'to party'), states (*avoir peur* 'to be scared'), objects (*pomme de terre* 'potato') and so on (Forsberg, 2008, p. 96). They are sub-clas-

sified into clausal and phrasal sequences. Clausal sequences are full clausal, propositional language-specific sequences, often with pragmatic connotations among which conversational routines are probably the best known whereas phrasal sequences are primarily used for their denotative meanings, and as a rule constitute phrases, sometimes with open slots, such as *X tenir X au courant de X* ('keep X posted on X').

When it comes to the practical identification of these sequences, Erman and Warren (2000) make use of the criterion *restricted exchangeability.* In order for a sequence to qualify as *conventional* (a prefab in their terminology), an exchange of one of the words for a synonymous word must always result in a change of meaning or a loss of idiomaticity (Erman & Warren, 2000, p. 32).

The first step in identification is to find the Lexical FSs that meet the *restricted exchangeability* criterion (Erman & Warren, 2000). This is then complemented by internet searches using Google.fr. The Google tests are carried out following a specific procedure. To test the extent to which restricted exchangeability applies to a sequence, an analogous sequence, which has been subject to one of the modifications listed below, is constructed. The modifications were established based on Erman and Warren (2000) and on empirical evidence, i.e. some of the modifications were found to be decisive through work with the data.

1. One of the words is exchanged for a synonymous word
2. One of the words is exchanged for an antonymous word (for example *ça marche mal* 'it works bad' instead of *ça marche bien* 'it works well')
3. Change of article (from definite to indefinite or absence of article)
4. Change of number (from plural to singular or vice versa)
5. Change in word order (for example *égalité femmes/hommes* 'equality women/men' instead of *égalité hommes/femmes* 'equality men/women')

For a sequence to be counted as formulaic, it has to appear at least twice as frequently on Google as any of the modified versions. To sum up, the methodology is based on linguistic criteria and the researcher's intuition, which is complemented by searches on Google.fr, in order to ascertain the researcher's intuitions.

3.3.2 Results Lexical FSs in relation to the CEFR levels

A quantitative study was carried out which calculated the number of Lexical FSs per 100 words in all of the groups. As observed earlier in this article, only groups A2, B1, B2 and C2 have enough productions to pass the statistical tests. The mean numbers of Lexical prefabs /100 words are shown in the table below.

Table 6. Mean value of Lexical CS at the CEFR levels

Level	Mean no Lexical FSs / 100 words
A2	1.01
B1	2.03
B2	3.07
C2	4.18
A2<B2	p<0.001
A2<C2	p<0.001
B2<C2	P<0.01

The numbers suggest that the higher the CEFR level, the higher the number of Lexical FSs. However, a statistical analysis using One-Way ANOVA with Tukey-Kramer post-hoc test showed that significant differences were only found between the following groups: A2 – B2 – C2. Consequently, table 6 above shows that there are differences between low levels and high levels of the CEFR-scale and also between intermediate and high levels, but it does not show significant differences between 'neighbouring' levels such as A2/B1 and B1/B2.

One possible reason for this lack of significance, which applies to at least B1/B2, is that we have very few participants at the B2 level. More participants would probably yield a statistical significance between the B1 and the B2 level. Longer texts would probably also yield more robust results, since lexis, due to its lower frequency in interlanguage, requires longer texts.

If the results obtained for Lexical FSs are compared to those regarding MSDs, two interesting aspects are found: MSDs are better at discriminating between each level of the CEFR up to B2 level, whereas Lexical FSs do not seem to render differences that are fine-grained enough to separate between e.g. A2 and B1. On the other hand, Lexical FSs do the job that MSDs do not succeed in doing, i.e. discriminate between the higher levels B2 and C2. Possibly, this is due to the fact that Lexical FSs develop modestly up to a certain level and that development continues even at the highest levels, possibly never ending, since we are dealing with the growth of lexis, which is constant even in the L1.

4. Summing up

This chapter thus investigated the linguistic development of grammatical, discursive and formulaic structures in productions made by informants placed at different CEFR levels.

Our research questions were:

1. Is it possible to establish linguistic developmental features to match the communicative characteristics of the CEFR levels? and, if so,
2. How do the proposed linguistic domains of morphosyntax (VP and NP morphology), discourse markers/subjunctions and formulaic sequences develop along the CEFR-levels?

As an answer to the first question on the basis of the investigation of 83 productions, the results above show that there is a decrease of MSDs across the levels with significant differences between the levels (see Tables 3 and 4). Thus it seems that linguistic features do discriminate between CEFR levels and more specifically the features chosen for this study. Measures of morpho-syntactic deviances thus yield significant differences between the CEFR levels up to B2. These results of late MSDs in highly proficient learners/users concur with findings in Bartning et al. (2009).

An answer to the second question, as shown in Table 5 above, is that there is development of a selection of grammatical and discursive features across the levels. These developmental features are represented by *dont*, *ce que*, *ce qui*, gerund, pluperfect and some connectors. Our study also presented results concerning the lexicon: it was shown that the use of lexical formulaic sequences increases at higher CEFR levels, but significant differences were only found between A2/B2/C2 (Table 6). This is yet another linguistic feature that can be used as a measure of progression in IL development and in the CEFR scale.

In the future, when the written CEFR corpus has been enlarged, the MSDs and their TL equivalents, as well as the developmental features in Table 5 above, will be examined in order to work out a developmental continuum of written French IL linked to the CEFR scale.

It thus seems, according to this study, that a relationship is to be found between linguistic development (the three different measures) and communicative development as expressed in the CEFR scales, at least as regards written production in L2 French. Possibly the three measures, i.e. morpho-syntactic deviances (accuracy), emergence and use of discourse/grammatical markers and, finally, the rate of lexical formulaic sequences could be proposed as constituting ingredients of a global index of interlanguage development. However, the measures and the rating procedures need to be further refined before drawing any further conclusions.

Acknowledgements
We thank Hugues Engel, Department of French, Italian and Classical languages, Stockholm University, for his valuable help with some data collection and relevant comments on earlier drafts of this paper and Victorine Hancock for her precious remarks. We also warmheartedly thank our two raters from the University of Jyväskylä, Finland. Finally, we thank the anonymous reviewers of this volume for their valuable comments.

References

Ågren, M. (2008). *À la recherche de la morphologie silencieuse. Sur le développement du pluriel en français L2 écrit* (Doctoral dissertation). Lund University, Sweden.

Bardovi-Harlig, K. (2006). Interlanguage development. Main routes and individual paths. *AILA Review, 19,* 69–82.

Bartning, I. (2009). The advanced learner variety: 10 years later. In E. Labeau & F. Myles (Eds.), *The advanced learner variety: The case of French* (pp. 11–40). Bern: Peter Lang.

Bartning, I. (in press). Late morpho-syntactic and discourse features in advanced and very advanced L2 French – a view towards the end state. In S. Haberzettl (Ed.), *The end state of L2 acquisition.* Berlin: Mouton de Gruyter.

Bartning, I., Forsberg, F., & Hancock, V. (2009). Resources and obstacles in very advanced L2 French. Formulaic language, information structure and morphosyntax. *EUROSLA Yearbook, 9,* 185–211. Amsterdam: Benjamins.

Bartning, I., & Hancock, V. (in press). Morphosyntax and discourse at high levels of second language acquisition. In K. Hyltenstam (Ed.), *High level proficiency in second language use.* Berlin: Mouton de Gruyter.

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies, 14*(3), 281–299.

Bolly, C. (2008). *Les unités phraséologiques: Un phénomène linguistique complexe?* (Unpublished doctoral dissertation). Université catholique de Louvain, Belgium.

Bybee, J., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure.* Amsterdam: Benjamins.

Council of Europe. (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages* [Online version]. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp

Ellis, R. (2008). *The study of second language acquisition.* Oxford: Oxford University Press.

Erman, B., Forsberg, F., & Fant, L. (2008, October). *Nativelike selection in high-level L2 use.* Paper presented at High-level proficiency in a second language, Stockholm, Sweden.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29–62.

Flament-Boistrancourt, D. (1984). *La pratique des relatifs français chez les néerlandophones* (Unpublished doctoral dissertation). University of Lille III, France.

Forsberg, F. (2008). *Le langage préfabriqué – formes, fonctions et fréquences en français parlé L2 et L1.* Bern: Peter Lang.

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching, 48,* 25–51.

Giacalone Ramat, A. (1992). Grammaticalisation processes in the area of temporal and modal relations. *Studies in Second Language Acquistion, 14,* 297–322.

Granfeldt, J. (2003). *L'acquisition des catégories fonctionnelles. Étude comparative du développement du DP français chez des enfants et des apprenants adultes* (Doctoral dissertation). Lund University, Sweden.

Granfeldt, J., & Nugues, P. (2007, June). *Évaluation des stades de développment en français langue étrangère.* Paper presented at TALN 2007, Toulouse, France.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Language Learning & Language Teaching: Vol. 6. Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam: John Benjamins.

Granger, S., & Pacquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins.

Hancock, V. (2000). *Quelques connecteurs et modalisateurs dans le français parlé d'apprenants avancés. Étude comparée entre suédophones et locuteurs natifs* (Doctoral dissertation). Cahiers de la Recherche 16. Stockholm University, Sweden.

Hancock, V. (2007). Quelques éléments modaux dissociés dans le paragraphe oral dans des interviews en français L2 et L1. *Journal of French Language Studies, 17,* 21–47.

Hancock, V., & Kirchmeyer, N. (2005). Discourse structuring in advanced L2 French: The relative clause. In J.-M. Dewaele (Ed.), *Focus on French as a foreign language* (pp. 17–35). Clevedon: Multilingual Matters.

Herschensohn, J. (2006). Français langue seconde: From functional categories to functionalist variation. *Second Language Research, 22,* 95–113.

Housen, A. Kemps, N., & Pierrard, M. (2008). The use of verb morphology of advanced L2 learners and native speakers of French. In E. Labeau & F. Myles (Eds.), *The advanced learner varieties: The case of French* (pp. 41–61). Bern: Peter Lang.

Howard, M. (2008). Morpho-syntactic development in the expression of modality: The subjunctive in French L2 acquisition. *Canadian Journal of Applied Linguistics, 11,* 171–192.

Howard, M. (2009). Short- versus long-term effects of naturalistic exposure on the advanced instructed learner's L2 development: A case study. In E. Labeau & F. Myles (Eds.), *The advanced learner varieties: The case of French* (pp. 93–123). Bern: Peter Lang.

Hulstijn, J. (2006, December). *Linking meaning (function) and form (lexis, grammar, prosody).* Paper presented at the SLATE Workshop, Amsterdam, Netherlands.

Kirchmeyer, N. (2002). *Étude de la compétence textuelle des lectes d'apprenants avancés. Aspects structurels, fonctionnels et informationnels* (Doctoral dissertation). Cahiers de la Recherche 6. Stockholm University, Sweden.

Klein, W., & Perdue, C. (1997). The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research, 13,* 301–347.

Labeau, E. (2009). An imperfect mastery: The acquisition of functions of imparfait by anglophone learners. In E. Labeau & F. Myles (Eds.), *The advanced learner variety: The case of French* (pp. 63–90). Bern: Peter Lang.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27,* 590–619.

Lewis, M. (2008). *The idiom principle in L2 English* (Doctoral dissertation). Stockholm University, Sweden.

Long, M. (2009, September). *Second language acquisition and language teaching.* Paper presented at Stockholm University, Sweden.

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in the foreign language classroom. *Language Learning, 48,* 323–364.

Perdue, C. (1993). *Adult language acquisition: Cross-linguistic perspectives* (Vol. 2). New York: Cambridge University Press.

Pienemann, M. (1998). *Language processing and second language development. Processability theory.* Amsterdam: Benjamins.

Schlyter, S. (2003). Development of verb morphology and finiteness in children and adults acquiring French. In C. Dimroth & M. Starren (Eds.), *Information structure, linguistic structure and dynamics of learner language* (pp. 15–44). Amsterdam: Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences* (pp. 127–151). Amsterdam: Benjamins.

Sharwood-Smith, M., & Truscott, J. (2005). Stages or continua in SLA: A MOGUL solution. *Applied Linguistics, 26,* 219–240.

Van Daele, S., Housen, A., Kuiken, F., Pierrard, M., & Vedder, I. (Eds.). (2007). *Complexity, accuracy and fluency in second language use, learning and teaching.* Brussels: KVAB.

Véronique, D. (Ed.). (2009). *L'acquisition de la grammaire du français, langue étrangère.* Paris: Didier.

Von Stutterheim, C. (2003). Linguistic structure and information organisation. The case of very advanced learners. *EUROSLA Yearbook, 3,* 183–206. Amsterdam: Benjamins.

Wray, A. (2008). *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

APPENDIX

Table 1. Overview of the six developmental stages proposed by Bartning and Schlyter (2004, p. 293)

STAGE	MORPHO-SYNTACTIC AND DISCURSIVE FEATURES
Stage 1 – Initial stage This stage brings to mind the prebasic and basic varieties (see Bartning & Schlyter, 2004, p. 295; Klein & Perdue, 1997).	**Utterance structure** - Nominal and non-finite utterance structure (*je // mes amis* 'I meet friends'); some bare nouns ; - formulaic expressions (*je ne sais pas* 'I don't know', *je voudrais* 'I would like') **Connectors** - Emergence of the connectors *et* ('and'), *mais* ('but') and *puis* ('then') **Negation** - Negation of the type Neg X (*non grand lit* 'no double bed') and preverbal negation **TMA** - Emergence and rare use of *passé composé* (very few contexts are marked for past tense) **Verb morphology** - Mostly non-finite verb forms but some finite verb forms; - no opposition between personal verb forms; **Gender** - Default value on the value of masculine/feminine on determiners **Pronouns** - Use of first person pronouns *je* without elision
Stage 2 – Post-initial stage	**Utterance structure** - Continued use of non-finite utterance structure but also some finite forms; - frequent use of the 'passe-partout' formulas *c'est* ('it is'), and some *il y a* ('there is'); - subordination (see below connectors) **Connectors** - Simple subordination with temporal, causal and relative clauses **Negation** - Use of the preverbal negation *ne* (without *pas*) along with the first uses of the TL form *ne … pas*; **TMA** - Emergence of modal auxiliaries *pouvoir* 'be able to', *vouloir* 'to want to'; - increase of *passé composé* forms; - use of *imparfait* with *être* ('be') and *avoir* ('have'); - first uses of periphrastic future; **Verb morphology** - Still frequent use of non-finite forms; - emergence of forms of irregular verbs in the present plural 3rd person as *ils *prend* 'they take'; - subject – verb agreement between 1st and 2nd person singular of non-thematic verbs (*être*, *avoir*); - alternative forms between the verb forms of the 1st person present plural *nous V-ons* and the short from **nous V (*nous parle* 'we speak') **Gender** - Continued default value on determiners - Some adjectival agreement **Pronouns** - Object pronouns in postposition of the verb

>>>

STAGE	MORPHO-SYNTACTIC AND DISCURSIVE FEATURES
Stage 3 – Intermediate stage A regular and systematic interlanguage with overgeneralisations and analogies that make the system sometimes not TL (see Bartning & Schlyter, 2004, p. 295).	**Utterance structure** - More finite utterance structures; - Subordination (see below connectors) **Connectors** - Use of causal, temporal, relative, interrogative clauses, and the subordinator *que* ('that'); overuse of *mais*, *parce que* **Negation** - TL use of the negation *ne …pas* on the finite forms of the verbs, - Emergence of *ne…rien* **TMA** - Use of *passé composé* with marked past contexts; - use of the periphrastic future; - use of the *imparfait* on lexical verbs - first use of isolated cases of *futur simple;* - first use of the subjunctive; **Verb morphology** - Still some non-finite verb forms; - the 1st person present plural is now mostly correct; - subject – verb agreement with opposition between the 3rd persons of singular and plural begins to be established with *avoir* and *être* (*a*/*ont*, *est*/*sont*); - alternation between *ils *prendre* and *ils *prend* (with some cases of the correct *ils prennent*); **Gender** - More TL use on the definite determiner than on the indefinite; - overuse of the masculine (determiners, adjectives); - problems with adjectival agreement in preposition and attributive sentences **Pronouns** - Object pronouns placed before the lexical verb in simple and complex tenses (often incorrectly after the auxiliaries *est*/*a*)
Stage 4 – Low advanced stage	**Utterance structure** - Multipropositional syntactic structures; discourse complexity that demands temporal and conditional expressions (see below TMA): the emerging forms of tenses, mode and aspect; these contexts are not always systematically marked by relevant forms; the form/function relations are not yet TL **Connectors** - Use of *alors, après, finalement, mais, parce que* and temporal *puis ;* significant overuse of the polyfunctional markers *mais* and *parce que* **Negation** - Complex negation with *ne… personne, jamais, rien* **TMA** - The *passé composé* and the *imparfait* are more and more TL; - Emergence of the typical French use of the conditional, the pluperfect, the subjunctive **Verb morphology** - The non-finite forms in finite contexts *je donnE* 'I give' disappear, except for verb forms in /-r/ like *je lire* ('I read'); - the forms *ils ont, sont, vont, font* dominate over *ils *a/est/va/fait* ; - continued use of *ils *prend* but in competition with *ils prennent* **Gender** - Overuse of the masculine on determiners and adjectives **Pronouns** - Object clitic pronouns before verbs; - subject clitic pronouns with elision

>>>

STAGE	MORPHO-SYNTACTIC AND DISCURSIVE FEATURES
Stage 5 – Intermediate advanced stage	**Utterance structure** - Multipropositional utterances, some infinitives and the gerund ; - inflectional morphology becomes functional **Connectors** - Appearance of *donc*; native-like uses of *parce que* **Negation** - Use of *personne … ne*, *rien … ne* ; - TL use except for the omission of *ne* **TMA** - Development of inflectional morphology (subjunctive, conditional, pluperfect) - The pluperfect is not used in all obligatory contexts, nor is the conditional; **Verb morphology** - Fragile zones of morphology in multipropositional utterances **Gender** - Still overuse of the masculine in determiners; - problems in agreement with preposed adjectives (feminine) **Pronouns** - More or less TL use, even the relative *dont*
Stage 6 – High advanced stage	**Utterance structure** - High degree of embedding and of integrated propositions ; - capacity of keeping several information levels simultaneously in the same utterance; - discourse structuring according to L1 (fewer constituents than NS in the pre-frontfield, 'préambule') **Connectors** - TL use of connectors; use of *donc, enfin;* macro-syntactic relatives **Negation** - TL use; variation on the omission of *ne* **TMA** - Use of the pluperfect (sometimes still replaced by the *passé composé*); - the conditional in most contexts **Verbal Morphology** - Stabilised inflection which becomes functional ; some rare *ils *prend* may turn up in complex syntax/discourse ; **Gender** - Same as stage 5. Adjectival agreement TL but some form of the type *la *petit ville* may turn up; problems with gender, the indefinite determiner before feminine nouns starting with vowels **Pronouns** - TL in form and position

Doing interlanguage analysis in school contexts

Gabriele Pallotti
University of Modena and Reggio Emilia

In second language acquisition (SLA) research it is well-known that interlanguages are autonomous and rule-governed linguistic systems whose grammar cannot be described simply in terms of errors and deviations from L2 norms. However, assessment practices at school, both formative and summative, heavily rely on counting errors and scoring them based on various types of 'gravity'. Carrying out a systematic interlanguage analysis as it is done in SLA research would be highly time consuming and very impractical in most teaching and testing contexts. The chapter will show some examples of interlanguage analysis in such contexts. Different stages of the research process will be focussed upon, including data collection, transcription, coding, scoring, and quantitative and qualitative analysis. If CEFR scales are to be related to acquisitional sequences in teaching and testing contexts, it is necessary to find ways in which the latter can be assessed in a reasonable amount of time and without specialized skills, while preserving the procedure's validity with the respect to current SLA theorising and methodology.

1. Introduction

One of the main aims of this volume is to find relationships between linguistic development, as described in Second Language Acquisition (henceforth, SLA) research, and the acquisition of communicative proficiency, as described in the Common European Framework of Reference (Council of Europe, 2001; henceforth CEFR). Most chapters report on studies in which various measures of linguistic performance calculated by SLA researchers are matched to assessments of communicative language proficiency made by raters experienced in language testing and assessment. The question remains whether these raters, or other practitioners, would be able to conduct the linguistic part of the analysis themselves, if this could be implemented in concrete (formative and summative) assessment practices, how it should be conducted and what type of training would be needed to allow teachers and language testers to implement interlanguage analysis in their professional domains.

The aim of this chapter is to discuss these issues, by reporting on a project aimed at bringing interlanguage analysis to school. The project involved sever-

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 159-190

al teachers at different levels, from kindergarten to middle school, as part of wider teacher-training and action-research schemes. Teachers were involved in collecting data and in their subsequent analysis based on the notion of interlanguage (henceforth, IL). These procedures were designed so that they could be implemented in everyday school settings. The main goal was formative assessment, in the more usual sense of being oriented to improving didactic strategies, but also in the sense of being part of teacher training.

The present chapter thus differs from others in the volume. No attempt will be made at relating the development of communicative language proficiency, as described in the CEFR, with the acquisition of linguistic structures. The main goal will rather be that of presenting a scheme for training teachers in analysing their pupils' interlanguage. This will lead to a more general discussion on how practitioners, like teachers and language assessors, can acquire the skills that are necessary for describing linguistic development in ways that are consistent with the notion of interlanguage and current SLA research. If SLATE's efforts are to have an impact on teaching and testing practices, it is important to address the issue of how linguistic development can be assessed in such contexts and how professionals should be trained accordingly.

The present contribution is far from offering an exhaustive and systematic answer to such crucial questions. The project reported here is in fact limited in several ways. Besides being still at a rather exploratory stage, it concerns a single language, Italian, and a very special learners' group, i.e. children aged between 5 and 12. However, it is hoped that the present discussion will stimulate a more general reflection on how interlanguage analysis can be integrated into formative and summative assessment.

2. Interlanguage description

2.1. The notion of interlanguage

The term interlanguage was first introduced by Selinker (1972), who defined it as "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [= Target Language] norm" (p. 214). This orientation derived in turn from the Error Analysis Approach (Corder, 1967), which emphasized that errors are an important window on the learner's processes and strategies and that their careful analysis is more productive, from a pedagogic and scientific point of view, than the mere counting, scoring and sanctioning of 'wrong' forms.

The relationship between errors and interlanguage development has remained intricate and several authors have felt the need, at various times, to stress that the two notions should be kept conceptually apart. Bley-Vroman

(1983) for example warned against the risk of committing a "comparative fallacy", arguing that "work on the linguistic description of learners' language can be seriously hindered or sidetracked by a concern with the target language" (p. 2). Similarly, Sorace (1996) notes that, if the aim is to reconstruct a learner's linguistic system, "the evaluation of the distance between native and non-native grammars becomes an irrelevant criterion" (p. 386).

The fact that different researchers, at different times and from different perspectives, have stressed the independence of the notion of interlanguage from those of error and conformity to L2 norms testifies that the two are often mingled. However, this is not always the case and several authors have provided descriptions of interlanguage development eschewing any reference to accuracy or errors. For example Pienemann (1998) proposes the notion of "factorization" as a way of disentangling various factors bundled together in the L2 which may lead to 'errors'. A learner may develop an interlanguage system in which just one of such factors governs a set of form-function associations, which should be described in their own right, regardless of the fact that they yield forms not allowed by L2 rules. For instance, in a fusional language like Swedish or German, adjectives may be inflected based on a variety of factors, such as gender, number, definiteness or case. A learner who associates one inflectional morpheme with just one of these factors, e.g. number, will produce many forms deviating from L2 norms, but would nonetheless follow a clear interlanguage rule.

Another approach that has coherently looked at interlanguage development in its own right is that of the Basic Variety (Klein & Perdue, 1992, 1997). These authors have described some organizing principles of utterance construction in the early stages of IL development. The principles predict the order in which constituents will appear in an utterance and they have to do with semantic notions such as 'the referent with highest control on the situation', or pragmatic-textual ones, such as 'topic' and 'focus'. Based on the extensive ESF project database (Perdue, 1993), the authors found these principles to hold independently of the L1 and the L2, thus pointing to a generalisable functional explanation of interlanguage development making no reference to errors and conformity to L2 norms.

The Basic Variety Approach, as the name itself suggests, was developed to investigate the initial stages of second language acquisition. In such cases it is more obvious that learners' productions should be analysed according to their internal logic rather than by looking at their conformity to L2 norms. Norris and Ortega (2003) suggest that this way of looking at interlanguage as an independent system is more relevant for initial stages, while other types of analysis are more appropriate for later stages. At intermediate levels quantitative measures of the spread and consistency of use of a structure may be employed, while accuracy-based measures would be more meaningful for characterizing the production of advanced learners.

2.2. Describing interlanguage development in terms of accuracy

Despite these recommendations, the description of learner language in terms of errors is still quite widespread. In SLA research many studies characterise learning over time or after an experimental treatment with accuracy measures, like the number of error-free T-Units or number of errors per 100 words (e.g. Wolfe-Quintero, Inagaki, & Kim, 1998). Teachers, too, often describe and evaluate their students' performance based on the number and type of errors made, which is also a common practice in language assessment. For example, the CEFR (Council of Europe, 2001, p. 114) provides descriptors for "grammatical accuracy", which constantly refer to the number and type of errors made, but none on "grammatical development". Examples of these descriptors are: "systematically makes basic mistakes (A2); errors occur, but it is clear what he/she is trying to express (B1); does not make mistakes which lead to misunderstanding (B2); errors are rare and difficult to spot (C1)". Clearly, since the CEFR is a language-neutral instrument, it would have been impossible to refer to the development of specific grammatical features, and a scale referring to a general notion of "accuracy" (necessarily related to errors' quantity and quality) is perhaps unavoidable. Users of the CEFR are indeed invited to consider "which grammatical elements, categories, classes, structures, processes and relations are learners, etc. equipped/required to handle" (p. 114), but nothing else is said about how interlanguage development should be conceptualised and reported. Even English Profile, a comprehensive project for relating CEFR descriptors to linguistic development in L2 English, heavily relies on counting and scoring errors (Hendriks, 2008; Williams, 2008).

There is nothing inherently wrong in reporting accuracy scores, but one should be clear about what is achieved with this type of approach. As Wolfe-Quintero et al. (1998) write: "the purpose of accuracy measures is precisely the comparison with target-like use. Whether that comparison reveals or obscures something about language development is another question" (p. 33). In other words, 'accuracy', i.e. the degree of conformity to L2 norms (expressed as the ratio of 'wrong' forms to overall production), should not be taken as a direct indicator of interlanguage development, and may actually somehow distort the picture.

The error-counting approach has some clear advantages. It is relatively easy to understand and can easily be applied by teachers, assessors and other practitioners, for whom noticing the presence of an error is more intuitive than understanding the internal logic of an interlanguage system. Judgements thus tend to be rather reliable, although various authors have reported a number of problems in the operationalization of what errors are and how they should be scored, a problem that becomes even more apparent in the case of qualitative judgements on the gravity of different types of errors (James, 1998). In any case,

when searching for correlations between descriptions of linguistic and communicative competence (as is done in most chapters in this volume), one should always bear in mind that, as regards the former, 'accuracy growth' and 'interlanguage development' do not represent the same construct.

2.3. Interlanguage analysis in research contexts

Analysing interlanguages in research contexts is often a complex and laborious process. Some authors, especially in the generative framework, have focussed exclusively on acceptability judgements or comprehension tasks, as these should better reflect the underlying linguistic competence without the many confounding factors related to on-line performance. However, in other approaches the predominant sources are production data and more specifically those coming from communicative oral tasks. The preference for oral data has been motivated by their being more spontaneous, unplanned, and thus based on implicit linguistic knowledge.

Oral data need transcribing, which is very time-consuming. For a medium-grained transcription, about 10 hours are required for one hour of data, and the figure can easily increase in the case of more accurate transcriptions, e.g. including phonological, prosodic or gestural information. Transcribed data are then usually subjected to some kind of coding in which all instances of the phenomenon under investigation are tagged, counted and classified. This leads to analysis proper, consisting in observing frequencies and relationships among categories in order to arrive at generalizable conclusions.

The aim of analysis is typically the discovery of developmental orders, for instance how a certain structure emerges and then gradually spreads out in interlanguage, or how this path is related to that of other linguistic structures. Studies may be purely descriptive or they may test specific theory-based hypotheses. In both cases, conclusions need to be reliable, hence the need for robust data sets containing large numbers of tokens of the phenomenon under investigation. If the focus is on the development of a specific feature, researchers must ensure that their data sets are sufficiently 'dense' (Pienemann, 1998), i.e. that they contain several contexts requiring the production of a given structure. Studies on developmental orders, making generalisable statements about the appearance of certain structures before others, need to set explicit acquisition criteria specifying the conditions that must be satisfied in order to conclude that a structure has been acquired (Pallotti, 2007).

Performing such analyses requires not only a substantial amount of time, but also a specific expertise, including the ability to identify variable linguistic rules and to define them in abstract terms, transcending the norms of a specific language but grounded in general principles such as those proposed by Universal Grammar or Functional-typological linguistics. All this is usually

beyond the reach of most practitioners in teaching and testing contexts. The project reported here had the aim of developing simplified and practically manageable ways of doing interlanguage analysis at school, with learners aged 5-12. The main goal was that of training teachers in new ways of looking at their pupils' linguistic productions, in order to improve their teaching and promote language development in multilingual classes. The insights gained in this research might also be generalized to other contexts, such as large-scale assessment, as will be discussed in the final section.

2.4 Assessing young learners

Assessing learners younger than 12 poses specific problems and requires methods and procedures that may differ significantly from those used with adolescents and adults. The relevant literature is relatively scant but steadily growing - see e.g. reviews by Rea-Dickins and Dixon (1997), Cameron (2001), Ioannou-Georgiou and Pavlou (2003), Hasselgreen (2005), McKay (2005, 2006), Bailey (2008), and a special issue of *Language Testing* (Rea-Dickins, 2000).

Carpenter, Fujii, and Kataoka (1995) is one of the few studies entirely devoted to the identification of suitable protocols for assessing young learners' interlanguage. The authors describe an oral interview procedure for assessing L2 learning in children aged 5-10. The procedure begins with a task requiring children to manipulate objects in order to respond to simple commands and answer simple questions requiring no or minimal verbal production. After a few minutes of conversation, an information gap task follows, requiring children to discover differences between a picture they have in front of them and one held by the interviewer. Children are then given four pictures and asked to select one that does not belong with the others, motivating their choice; this elicits the production of comparisons and academic language on similarities, differences and categorization patterns. The next task is a picture-story narrative based on *Goldilocks and the three bears* and the last task is a role play where children, using puppets, impersonate first themselves talking to another child, then a teacher talking to students, in order to elicit register variation.

I am not aware of other published sources describing systematic research procedures for a comprehensive assessment of young learners' interlanguage. Ideas and practical suggestions can be drawn from empirical studies that have used one or, occasionally, more than one elicitation tool for data collection in child SLA (see e.g. chapters in Philp, Oliver, & Mackey, 2008). Other useful resources are contributions on data elicitation techniques for SLA research in general (for a comprehensive review, Gass & Mackey, 2007), most of which can be employed, perhaps with adaptations, with young children. Alternatively, many methods for assessing monolingual children can be used for bilinguals (a useful review of methods for assessing syntactic competence can be found in

McDaniel, McKee, & Smith Cairns, 1996). Finally, a few exams now exist that are specifically designed for assessing young learners of a second or foreign language, such as the Cambridge Young Learners English Tests (see Taylor & Maycock, 2007).

For the assessment of young children in school contexts, most authors indicate the portfolio as the approach of choice (e.g. Ioannou-Geogiou & Pavlou, 2003). Valdez Pierce and O'Malley (1992) examine a number of options for conducting performance and portfolio assessment in school contexts, pointing out the strengths of these approaches and some possible problems, including the amount of time required and the difficulty to administer tasks to individual students. More recently, Hasselgreen (2003) reports on a large-scale international project for the construction of portfolio based on CEFR 'can do' statements especially geared towards a population of young adolescents aged about 13-15.

Young learners' assessment is most of the times formative, i.e. it aims at diagnosing pupils' strengths and weaknesses, their achievements, difficulties and developmental paths, in order to make teaching more effective and, possibly, make pupils (especially older ones) aware of their own learning, which is supposed to increase their motivation and autonomy. Given these premises, issues of validity do not primarily concern the representativeness of the sample or its adequacy for evaluating the skills possessed by an individual in a range of everyday situations, as in more standardized high-stakes testing, but rather have to do with the usefulness of assessment for promoting learning.

3. Interlanguage analysis in school contexts: a case study

3.1. The project

The main goal of the approach described here is to bring techniques and strategies of interlanguage analysis commonly used in SLA research into school contexts, in order to increase teachers' "diagnostic competence, i.e., the ability to interpret students' foreign language growth, to skilfully deal with assessment material and to provide students with appropriate help in response to this diagnosis" (Edelenbos & Kubanek-German, 2004, p. 259).

The main aim is thus formative assessment, giving teachers conceptual tools to understand how their students make progress in the second language in order to better assist them with well-designed and appropriately timed pedagogic activities. Most teachers in the contexts where the project was carried out lack the skills needed to interpret interlanguage development. They notice pupils' errors, interpreting them with vague and completely impressionistic opinions about their 'gravity'. At the primary and middle school levels they believe that their intervention should just consist in marking all the errors and counting

them, while in preschool the prevailing attitude is 'let nature take its course', i.e. not making any specific effort at focussing children's attention on grammatical forms. Little or no attempt is made at understanding how and why errors are produced and it is often the case that even very different error types, such as grammatical, lexical, phonological or orthographical, are bundled together and hardly ever meaningfully set apart. This confusion is aggravated when it comes to summative assessment. School teachers discuss whether students should pass or fail based on their errors, what errors can be considered to be acceptable after a certain number of months or years of exposure to the L2, or how grades should be assigned based on the number and type of errors (for a discussion on critical issues and good practices in classroom-based assessment, see Rea-Dickins, 2008).

The project presented here had the aim of training teachers in a different approach. They were asked to collect and transcribe samples of their pupils' oral productions, which departs from their usual practice of assessing and grading written texts only. Although time-consuming, the act of transcribing itself promotes closer attention to interlanguage dynamics and the realisation that children construct their own rules in creative and systematic ways, rather than just 'make mistakes'. This awareness was further stimulated in subsequent analyses of the transcribed materials, in which teachers noted down, classified and interpreted various types of linguistic behaviour, making an effort to use positive formulations of what structures are present and how they work rather than just listing errors and shortcomings. Teachers were thus asked to reproduce, in a simplified and assisted way, the methodology of SLA research projects. The aim was to promote a new attitude towards interlanguage productions, based on understanding their internal logic and systematicity, in order to assist learners in their gradual approximation to the target language. The focus was mainly on grammatical structures, but teachers were also asked to look at lexical, textual and communicative features, in order to realise that grammar is just one dimension to be considered.

3.2. Context and participants

The project has been conducted for the past three years and is still under way. It involved 10 kindergarten, 7 primary and 2 middle school classes in different parts of Northern Italy, with a total of about 40 participating teachers, who were actively involved in data collection and in the creation, selection and fine-tuning of procedures for data elicitation and analysis. Altogether, about 120 NNS children of different linguistic backgrounds and 40 NS children have been included, aged between 5 and 12. The NNS children's proficiency in Italian varied from very basic to native-like. Some of them were enrolled in childcare services in Italy very early on, even before 3, and their competence in Italian was

virtually identical to that of their monolingual peers. Many others began their exposure to Italian in preschool, after 3, and a few started at 6 or later. It is not possible to provide here a more detailed description of their linguistic levels, as the assessment of their linguistic development was one of the aims of the project and the data collected thus far do not allow sorting them into precisely defined developmental levels. Actually, 'grading' pupils and sorting them into levels was not one of the aims of the scheme, and it was in fact discouraged, as will be discussed in the final section. The primary goal was to make teachers understand the logic of interlanguage systems, and asking them to classify children into bands or levels would have distracted them from the task of interpreting their productions.

The group of teachers was heterogeneous, too. All of them were already in-service and the majority had a considerable number of years of experience. Preschool teachers in Italy do not specialise in specific curriculum areas. Primary school teachers can in principle teach all areas of the curriculum, although most of them used to specialise in two broad areas, humanities and mathematics-sciences. This has been changed by a recent reform (2009), which is promoting a return to a model where a single teacher is in charge of all subjects. Most of the primary school teachers involved taught in the humanities area, as was the case for the single middle school teacher participating in the project. None of them had previous training in interlanguage analysis or applied linguistics. They took part in the training scheme voluntarily, as part of their elective training courses or in projects for experimenting effective ways of teaching Italian to language minority children.

3.3. Collecting interlanguage samples

The first step in the training scheme consisted in making teachers aware of the concept of 'data density' (Pienemann, 1998), i.e. the fact that some communicative tasks tend to promote or require the production of certain grammatical features more than others. They needed to understand that if a structure is not produced this does not necessarily mean it should have been - in other words, 'absent' is not equivalent to 'missing'. This implies a sensitivity to the relationships between activities and linguistic structures, which are quite obvious when teachers think of grammar exercises and drills, but tend to be overlooked in the case of oral communication tasks.

Secondly, it is also important to reflect on which grammatical structures are more informative about a learner's development, i.e. what are good 'diagnostic features'. A feature with high diagnostic value is one with a relatively slow development, appearing early but continuing to be challenging even to more advanced learners. This way, a communicative task providing a number of contexts for producing that structure would be relevant for assessing learners at very

different proficiency levels, whose performance may range from not supplying the structure at all, to producing it in a few sterreotypical contexts, to using it correctly in most cases except for the most irregular / exceptional ones, to complete mastery. Based on previous research (synthesized in Giacalone Ramat, 2002), the following structures were selected as having good diagnostic value for L2 Italian.[1]

Past tense marking
A number of studies have described a developmental sequence for past tense marking in Italian interlanguage (for a review, Banfi & Bernini, 2003). The past participle marker -to emerges as the first grammatical morpheme productively attached to the verb stem, expressing a perfective meaning, typically (but not exclusively) applied to past time contexts. The auxiliaries have/be required for passato prossimo ('present perfect') are at first produced erratically, then become more consistent and their choice more target-like, with be applied to unaccusative verbs and with pre-verbal clitic object pronouns (Pallotti & Peloso, 2008). When the auxiliary be is used, the past participle must agree in person and number with its subject or object, a feature emerging relatively late in L2 Italian (Chini, 1995). Verb conjugation in the imperfect (imperfetto) always appears after the emergence of perfective marking, both in the synthetic form expressing habitual, iterative aspect (lui mangiava molto, lei leggeva tutte le sere 'he used to eat much' 'she used to read every evening') and in the compound form in conjunction with gerundive used in progressive contexts (lui stava mangiando, lei stava leggendo 'he was eating' 'she was reading').

Clitic pronouns
Italian, like other Romance languages, has a series of clitic pronouns appearing only in object positions. They vary for person, number, gender and case (direct vs. oblique). Their high frequency in the input makes them appear relatively early, often as part of unanalysed formulas, but the full system is mastered only after many years (Berretta, 1986; Giannini, 2008; Maffei, 2009).

Noun phrase agreement
In Italian all the elements of the noun phrase must agree in gender and number with the head noun, as in la casa bella ('the-f.sg. house nice-f.sg') or i

1 While choice of diagnostic structures is inherently language-specific, it is also the case that similar structures prove to have good diagnostic values in different languages. For example, agreement within and across phrases and past tense marking have been included by Bartning and Schlyter (2004) in their model of French L2 development. See also Forsberg and Bartning (this volume).

> ragazzi italiani ('the-m.pl. boys Italian-m.pl.). Some adjectives have four different endings, others only two, one for singular and one for plural, regardless of gender; there are also smaller classes with invariable adjectives and other inflectional patterns. The problem of agreement is compounded with that of noun gender assignment, which can be inferred in many cases from the noun's phonological ending, but has to be learned by heart in another substantial proportion of cases. The first signs of agreement appear very early in article-noun sequences, which might often be formulaic chunks. Agreement with other elements of the noun phrase appears later and develops gradually, with application ratios growing slowly and approaching 100% accuracy only in the most advanced learners (Chini, 1995).

It should be stressed that the choice of these tasks was partly determined by the nature of the language under investigation - other languages with different diagnostic features may require a different selection of stimuli. These tasks were found to be effective also because they could be performed - of course in quite different ways - by children of all ages and at a variety of proficiency levels, from near-beginners to highly fluent speakers, thus allowing comparisons across groups of learners.

The communicative tasks used aimed at achieving a good density of these diagnostic features, plus of course several other grammatical structures such as the marking of other verb tenses or the use of prepositions. Some tasks were selected also because they had been effectively used in previous research on children learning Italian as a second language, allowing for comparison with already existing data bases. This was the case of the picture story *Frog, where are you?* (Mayer, 1969), employed in a number of studies on Italian L1 and L2 development (e.g. Serratrice, 2007), and of the cartoon Reksio, utilized in the project *Construction du discours par des apprenants de langues, enfants et adultes* (Watorek, 2004; for L1 Italian, Giuliano, 2006). The tasks were performed by the children individually in a separate room, thus requiring the presence of a second teacher or researcher during ordinary class activities.

In the course of the three years of experimentation, and in the various locations where the training scheme was implemented, a number of different communicative tasks were piloted, some of which were also constructed by the teachers. In the following pages, only those used more consistently will be described, followed by a critical discussion about their strengths, weaknesses, possible variants and suggestions for improvement.

Free conversation

The interview begins with an ice-breaking conversation about the child, his or her family, school experiences and other similar topics. In this phase some more complex questions can be asked, eliciting decontextualised speech about family,

friends or objects having a special significance to the child. Unless the child is particularly shy or lacking linguistic resources, the interviewer should ask mostly open-ended, generic questions (like *and then?* or *What do you do every morning at school?*) or use mirroring techniques for letting the child continue without prompting him or her, i.e. by repeating parts of her previous utterances (*You said you have a little cat, how nice*) which displays active listening without leading the dialogue with questions and answers. Free conversations tend to have a relatively low data density, as the interviewer often does most of the talking by asking series of questions which may receive only minimal answers (see also Pienemann, 1998, pp. 297–303). However, in some cases they may offer relevant data for analysis and they may be used to sensitize teachers to the differences between their usual instructional conversations and tasks which allow students to perform more autonomously.

Past events narration

The introductory conversation can naturally lead to the narration of past events or states. In our protocol, this has proven to be the most effective and reliable way of eliciting past tenses. In fact, the question *What did you do last Sunday?* straightforwardly provides a number of obligatory contexts for past. The interviewer asks questions involving perfective (e.g. *what did you do last Sunday/yesterday/during the holidays?*) and imperfective aspects (*what did you use to do when you were going to the creche/nursery/when you were five years old?*). In order to elicit a variety of person markings, the conversation should not concern the child alone but also other persons such as friends or family.

A more structured variant of this task was having the children do an activity in class (e.g. preparing a fruit salad), take pictures of the various phases and using them as prompts to stimulate recall of various sub-actions with questions like *What did we do yesterday in class? Okay, we washed the fruit and then?* In this way, their productions on past tense events are more similar and can be compared.

Picture-story retelling

The child looks at pictures representing a series of events and then tells the interviewer the story. The picture-story *Frog, where are you?* (Mayer, 1969) was employed in some cases as it is clear, sufficiently long and complex to provide relatively rich data samples and can be understood by children at all ages after four. Narratives elicited with this story, especially by native or near-native speakers, tend to be rather long, which is an advantage in a research context but can pose problems in school contexts, where time for transcription is limited. For this reason another story was employed, a series of six pictures representing a father and a child going to a lake, catching a fish and taking it home; when the father is about to kill the fish with a knife the child starts crying and they go

back to set it free, but as soon as the fish is thrown into the lake it is swallowed by another bigger fish.[2] This story elicited narratives of about 100 words, which despite their brevity contained a number of interesting grammatical features, such as noun phrase agreement, clitic pronouns, person marking, prepositions. Obviously the tokens were very few, but this can also be seen as an advantage, as teachers found these stories easy to transcribe, analyse and compare. It is of course impossible to make an accurate and reliable estimate of the frequency and distribution of linguistic structures with such short texts, but they are nonetheless relevant for an analysis based on the emergence or presence of structures and for a first interpretation of their functioning.

In order to avoid use of non-verbal communication, children were told to hold the book in front of their eyes, without showing it to the interviewer or pointing to the pictures. Sometimes they would ask for a specific word, especially in the more complex Frog story. The interviewer would first prompt the child to find the solution autonomously, also using a paraphrase or a related term; if this attempt failed, the word was provided for the sake of maintaining a relaxed conversation flow.

Picture stories are effective for assessing various linguistic features, including lexical variety, syntactic complexity, expression of space and motion or participants' intentional states. As regards verb conjugation, they work well for assessing person marking, while they pose problems for time-reference evaluation. In fact, the task itself does not orient the speaker towards a particular interpretation, and narrations in the past tense - (*Once upon a time) there was a boy and he was looking at a frog* - or in the present - (*here, under my eyes) there is a boy and he is looking at a frog* - are both acceptable. In our data we observed a tendency to favour past tense narration in younger learners and present tense in older ones, with many children from 5 to 7 freely alternating present and past in the same story, thus making an analysis based on obligatory contexts impossible - in other words, with these data one can list the forms that are used, not those that are missing.

This problem might be overcome by giving interviewees an explicit prompt at the beginning, like *You must start the story with 'there was a boy who lived in a small house'*. However, these prompts don't seem to work even with adult speakers, who may end up using past tense marking in a story that is supposed to be told in the present, and vice versa (Robinson, Cadierno, & Shirai, 2009).

---

2 The story comes from the series *Vater und Sohn* by Erich Ohser (better known under the pseudonym of Plauen), published by Südverlag (www.vaterundsohn.de).

Video retelling
Following previous research by Watorek (2004), a cartoon of the series Reksio was employed to elicit narrative texts. In order to make communication more effective the interviewer was not supposed to watch the video together with the child, but left the room or attended to some other task, like writing or reading documents. With this stimulus children tended, more than with picture stories, to tell the story in the past, in the form *(In the video I've just seen) there was a boy....*, although some children also used the present tense or alternated between the two. This procedure is effective for assessing the ability to construct a coherent narrative, with appropriate reference to entities, time, space and the characters' psychological states, plus more general linguistic features like lexical choice, inflectional morphology or use of determiners and prepositions.

Spot the difference
Children were asked to describe differences between two similar pictures. This proved to be a more effective procedure than describing a single picture, which was initially piloted but eventually discarded. First of all, children found it more motivating. Furthermore, their productions tended to have more comparable sizes, as there was an optimal length given by the total number of differences to be reported, whereas free descriptions may dramatically vary in length, with some of them covering several pages of transcript. Thirdly, the task worked particularly well for eliciting complex Det-Adj-N noun phrases, because they were made communicatively necessary by the nature of the differences themselves - pictures were specifically drawn and coloured so as to contain e.g. two grey knives in one and three white knives in the other.

3.4. Transcription

Transcribing oral data was the most challenging part of the project, although it was seen as an important step for helping participants familiarise themselves with the dynamics of oral communication in the L2. Teachers were trained to use the software Soundscriber (www-personal.umich.edu/~ebreck/sscriber.html) on digital audio files and some managed to transcribe all of their data by themselves. However, most teachers' typing skills were very modest and their time for the project was limited, so that they ended up transcribing only very short texts. The six-picture story worked particularly well in this regard, as it produced only a few lines of transcript that still allowed for comparison across children and made teachers aware of the value of interlanguage analysis. This held true also for the video retelling, despite its yielding slightly longer stories. Longer tasks, like the Frog story or the description of a complex picture, often produced texts whose transcription required an amount of time beyond that which could be asked of in-service practitioners. In these cases, data were not

transcribed or they were transcribed by university students or staff hired by institutions promoting the teacher training schemes. The audio files were nonetheless archived by the schools to create portfolios for the children involved.

Transcription was kept at its simplest. All the words uttered were transcribed exactly as they were produced, including false starts, retracings and cut-off segments. Inaudible speech was represented by series of 'x', while best guesses were enclosed in parentheses. The interviewer's turns were systematically transcribed only if they contained speech, while those containing backchannels could be omitted. Pauses were marked by the symbol # or series thereof, with each token representing approximately a half second.

3.5. Analysis

Analysing data can be a very time-consuming activity, too. This is further complicated by the fact that teachers need to learn the features to focus on and appropriate ways to describe them, going beyond simple error spotting. In order to help them find regularities in interlanguage and discover its internal logic, several analytic grids were developed in the course of the project. They will be presented in the following pages, with a discussion about their merits, critical points and suggestions for further use.

The first grid (Table 1 in the Appendix) consists of a simple list of the essential areas to look at, leading to a more systematic observation than simply making disordered remarks on whatever feature meets the eye. In this table, a first basic distinction is made between communicative competence and linguistic competence. The former is further divided into efficacy (the ability to reach one's communicative goals) and fluency (the ability to do so smoothly, quickly and effortlessly). Since the main goal of the project was to focus on interlanguage, this aspect was intentionally left underspecified - further training on CEFR scales, for example, might stimulate more detailed accounts of communicative competence. The part on linguistic competence was divided into broad levels of language description, viz. noun and verb systems, syntax and the lexicon. Each of these levels contained a few sub-headings for the main categories requiring special attention - for example, in the noun system teachers were asked to systematically look at noun and adjective morphology, noun phrase structure (presence of constituents, agreement phenomena) and pronouns; in the verb system, at verb conjugation and inflection for tense, aspect and modality. They were also invited to illustrate each of their remarks with a few examples, indicating the relevant transcript line (examples of analyses based on this grid can be found in Ledda & Pallotti, 2005). It turned out that this grid worked well in training sessions, where teachers' analyses were assisted by an experienced trainer. However, when they were left on their own, most teachers

were able to write just a few minimal remarks for each heading, most of which were in the form 'uses X' or 'makes errors on Y'.

A more detailed version of the grid was thus designed including an exhaustive list of the features to be looked at (Table 2 in the Appendix). The main headings and sub-headings remain basically the same as in the previous version, with the addition of a macro-area 'textuality'. However, each area is outlined in much greater detail, with a number of characteristics to be taken into consideration and a brief introduction to the phenomena to be looked at, which served to illustrate key terms and basic processes relevant for that aspect of language. For instance, different aspects of communicative competence are described using slightly modified versions of some CEFR descriptors for Spoken fluency (Council of Europe, 2001, p. 129), Phonological control (p. 117), Qualitative aspects of spoken language use (p. 28), Sustained monologue (p. 59).[3] The section on nominal morphology recalls the basic difference between number and gender, the arbitrariness of the latter for all inanimate and most animate nouns and the existence of different inflectional classes. Turning to noun phrase construction, teachers were asked to consider various types of agreement among different constituents of the noun phrase and to look for possible systematicities and differences in singular/plural or masculine/feminine phrases, and so forth for all the other categories.

Teachers reacted positively to this second grid, as it was clearer to them what features should be focussed on and with which analytic categories. However, some of them responded to some items with a simple yes or no. For example, from the list of possible verb forms (present simple, present perfect, imperfect, subjunctive etc.) they just ticked the ones being produced. This is not wrong in itself, and the table actually invites such answers in some cases. The problem may be that in this way one might not arrive at a real understanding of an interlanguage's internal logic and rationality, which are more complex than a simple list of L2 features. In other words, if training is to be effective, these longer and more detailed checklists should not be used as inventories of items to be ticked, but should rather be seen as a memory aid for conducting a careful analysis.

---

3 CEFR scales were developed for describing adults' performance and their application to young children is not straightforward. In this project the adaptation consisted of choosing only selected descriptors from a few relevant scales, slightly modifying some of them. However, rating children on CEFR levels was not one of our goals, and the descriptors were actually given without any reference to CEFR scales and levels. For an application of CEFR scales to young learners see Papp and Salamoura (2009); see also chapters by Alanen, Huhta, and Tarnanen and Martin et al. in this volume discussing the CEFLING project on adolescents aged 12–16.

In order to stimulate a more qualitative look at interlanguage strategies, a third tool was proposed, focussing on one structure at a time (Table 3 in the Appendix). For each linguistic structure with particular diagnostic value, e.g. clitic pronouns, past tenses or noun phrase agreement, teachers were asked to select in their data examples that might help them understand how the learner is using that structure. According to emergentist / functionalist theories of language acquisition (Ellis, 2006), new rules emerge in interlanguage with a gradual spread from more prototypical cases, which may be learnt as lexicalized phrases or chunks, to slightly more abstract patterns, based on relatively simple generalizations on frequently occurring form-function mappings, to more abstract and complex patterns, incorporating a variety of features at the same time, and allowing for exceptions and irregularities (see also Martin, Mustonen, Reiman, & Seilonen, this volume). For example, verb marking for imperfective aspect will first appear on prototypical verbs encoding states, like *be* or *have,* then spread to activities and only finally to verbs expressing punctual notions, where speakers display their skill by producing unusual combinations of verb actionality and aspectual marking (Andersen & Shirai, 1996; Giacalone Ramat, 2002).

Hence, not all tokens of a grammatical structure are equal, some of them displaying (or allowing to infer) more proficiency than others. Similarly, not all errors are equivalent - some may indicate complete ignorance of the structure while others may be due to imperfect, partial or not completely automatized knowledge. This third analytical tool tries to capture this state of affairs, asking teachers to note down examples displaying complete or partial lack of knowledge, or examples indicating general knowledge but problems with irregular, unusual, complex cases, or examples where application to such cases may lead to the conclusion that the structure has been thoroughly acquired.

This approach allows one to zoom into specific structures and to assess their acquisition one by one. The same level of detail may be applied to a quantitative analysis, which then becomes very similar to those used in many SLA research projects, as teachers are requested to focus on one structure at a time and to count its various target- and non-target like forms. Table 4 in the Appendix exemplifies this type of grid with *passato prossimo* ('present perfect') in Italian. The first four lines are used to score various types of correct realizations, ranked in a tentative order of difficulty from the easiest ones, involving the unmarked participle ending *-to*, to more complex cases of participles inflected for gender and number or irregular participles (although acquisition of some of the latter may be facilitated by their high frequency). Subsequent lines classify various types of forms deviating from L2 norms, in what is evidently a blend of interlanguage-based (what forms are produced and what interlanguage rules are followed) and accuracy-based (the correspondence of these forms to L2 rules) descriptions.

Despite its seeming complexity, teachers found this table rather easy to compile, as they were guided by the pre-formulated descriptions, needing only to assign each token to one of the categories. This grid allows a fine-grained distributional analysis: rather than just counting correct and wrong forms, one can obtain a picture of what types of grammatical strategies are being employed, which may be revealing of developmental levels and acquisitional paths. Furthermore, one can obtain distribution ratios, expressed as percentages, which can be used to analyse data both longitudinally and cross-sectionally.

A problem with this procedure is that compilation can be rather time-consuming: transcripts must be searched for tokens of the target structure, which are then scored in the appropriate lines. While this scoring turned out to be rather fast, as in most cases it did not involve complex decisions or interpretations, it nonetheless required a careful look at transcripts for each diagnostic feature - hence the need to limit these to not more than two or three. A second problem is that the number of tokens for each category tends to be rather small, unless substantial data samples are collected. Quantitative results that can be obtained from this table should thus be interpreted cautiously due to their limited reliability. However, one should also bear in mind that the purpose of this analysis is to ascertain the presence, absence and logic of certain interlanguage strategies, rather than producing generalizable quantitative statements about L2 development.

4. Implications for teaching and testing

A set of procedures has been presented which were piloted to analyse children's interlanguage in a variety of school settings, as part of in-service teacher training schemes. The lessons that can be gained from this pilot study, and which may be extended to other contexts, can be grouped under two main headings - implications for teaching and implications for testing.

As regards teaching, the scheme had a positive impact on the participants' professional development. Teachers were overall satisfied, with some of them even enthusiastically reporting that this close attention to their students' productions radically changed their attitudes and practices. This had an impact on the teaching of Italian to native speakers as well, both in multilingual and in monolingual classes - analysing what they knew and what they did not, and showing that differences between them and non-native speakers were often limited to just a few areas, helped teachers reconceptualize many aspects of language education in multilingual classrooms. Traditional activities based on classifying and labelling linguistic structures gave way to more functionally-oriented ones, focussing on the areas that careful observation showed to be weaker.

This led to using an active approach to language education, which most of the time took the form of cooperative learning with mixed-level groups. Some of these innovative didactic activities were published on the Web as resources to share with colleagues (www.comune.re.it/interlingua). Thus interlanguage analysis has an important role to play as part of teacher training, both pre-service and in-service, as it leads to more learner- centred activities and to the realisation that any effective pedagogic intervention should start from understanding learners' competences, strategies and processes.

The value of interlanguage analysis for formative and diagnostic assessment is thus undeniable. In this context, issues of reliability and data robustness become less crucial - the logic of a child's interlanguage can be inferred, or at least acknowledged, even with small speech samples, and the very act of transcribing and analysing a few lines may already promote such a change of attitude. If time is at a premium, accurate analysis can be conducted only on those children who need more careful monitoring, e.g. newcomers or those with special difficulties. For these and all the others, collected data (digitised oral productions, written texts etc.) can in any case be seen as part of a portfolio documenting individual learning paths, regardless of whether and how they are accurately transcribed and systematically analysed.

A second area for which this study may be relevant is language testing. Other chapters in this volume report on research aimed at matching the development of communicative proficiency with profiles of second language acquisition. The problem arises of how such linguistic profiling can be practically incorporated into language testing.

It is unlikely that procedures like the ones discussed above may be directly used in large-scale, standardised testing. Firstly, they present all the problems (but also the strengths) associated with performance assessments, i.e. assessments in which '(a) examinees must perform tasks, (b) the tasks should be as authentic as possible, and (c) success or failure in the outcome of tasks, because they are performances, must usually be rated by qualified judges' (Norris et al., 1998, p. 8). In the case at hand, these judges should have a very special type of qualification, i.e. the ability to analyse interlanguages, which requires extensive training. Furthermore, an interlanguage analysis as is usually done by SLA researchers, or even in the simplified adaptation exposed in previous pages, takes a considerable amount of time.

If linguistic profiling is to be incorporated into large-scale language assessment, less time-consuming alternatives must be devised. A possibility would be using written data, which makes transcription unnecessary. However, the equivalence of interlanguage production in the oral and written mode has to be demonstrated, not assumed. Alternatively, raters can be instructed to assess oral language performance 'on the fly', while listening to it directly or from a record-

ing. The task can be assisted by the computer, as with the Rapid Profile software, developed by M. Pienemann and associates (http://groups.uni-paderborn.de/rapidprofile/), which provides an interface where the rater can input scores for a small set of diagnostic features. An even simpler solution would be having raters formulate holistic judgements on interlanguage structures, based on a checklist like the ones presented above but applied directly to learners' productions without transcribing them. The checklist would probably be formulated in terms of a scale, with different descriptors ordered in a series of levels, corresponding to a developmental sequence. Such rating scales could concern individual structures (e.g. tense marking, noun-phrase agreement, articles) or group several grammatical structures to provide a global picture of interlanguage development at various levels (but this rests on the assumption that it is indeed possible to identify relatively stable 'levels' of interlanguage development comprising a number of features; see Bartning & Schlyter, 2004, for such an attempt).

Whatever choice is eventually made regarding the implementation of interlanguage analysis into large-scale testing, it is essential that raters have a sound understanding of what an interlanguage is, how it works and how it should be analysed. The procedure presented here may offer some ideas for their training.

First of all, raters need to learn to separate the two areas of communicative proficiency and linguistic development. While it is true that the two dimensions are often related and grow side-by-side, there is also a considerable degree of independence, so that the two constructs are separated in most models of language proficiency. The first thing that raters need to learn is to keep the two dimensions apart, at least for analytic purposes. They should also be aware that constructs such as 'error compromising/not compromising communication' are spurious, in that they mingle a linguistic dimension (accuracy) with a communicative one (adequacy) (see also Kuiken, Vedder, & Gilabert, this volume).

A second aspect that needs to be discussed in a training scheme for raters is the difference between linguistic development and accuracy. Teachers and testers alike are frequently prone to the 'comparative fallacy', which entails describing interlanguage development in terms of errors and conformity to L2 norms. Especially in the initial-intermediate stages of acquisition, it makes little or no sense to count errors and other deviations from L2 norms, while it is more productive to recognise that an interlanguage has reached a complexity and sophistication level higher than another, something that can be quite unrelated to the number of errors produced. In order to do so, raters need to become aware of how interlanguages develop over time in order to express more complex grammatical functions, and the grids presented in this chapter may help them achieve such an awareness. The importance of transcribing oral data should not be overlooked in this respect, as it is an effective way of focussing one's attention on important details of linguistic production, including the use of prosody for marking infor-

mation structure at the phrase, sentence or text level, or the subtle interactions of phonology and morphology in word endings. Once this understanding of interlanguages as autonomous systems is firmly consolidated, one may discuss whether accuracy-based analyses could be used (though not exclusively) for characterising intermediate-advanced varieties, as Norris and Ortega (2003) suggest. A grid like the fourth presented in this chapter points to this direction, as it allows one to recognise the presence and distribution of both a variety of interlanguage forms and of structures conforming to L2 norms.

Learning to perform an adequate interlanguage analysis takes a substantial amount of time, based on the experience reported here. While most teachers involved in the project reacted very positively to the 'new' way of looking at their pupils' productions, when demonstrated by an experience trainer with an academic background, they encountered some difficulties in doing the analysis themselves. The main reason is that although most of them taught Italian as their main subject, they lacked an up-to-date and scientifically appropriate metalinguistic terminology. A relevant proportion of linguistic education in Italy deals with metalinguistic description, but this is done with a multitude of traditional categories - some of them misleading or ill-founded, some no longer in use in contemporary linguistics - while other crucial ones are missing, including *aspect, determiner, morpheme, phrase*. More importantly, traditional linguistic analysis at school consists in the mechanical application of metalinguistic labels (such as 'abstract noun', 'present perfect', 'concessive clause') to written texts, with little or no understanding of the general mechanisms responsible for the production of linguistic structures. Traditional labels may work relatively well when applied to the description of standard written Italian, but they fail when other languages or varieties are to be described. In such cases, concepts from general and functional-typological linguistics are essential, because they allow one to focus on linguistic processes, on how language works, rather than simply classifying individual items in a sentence. In other words, what teachers (and probably most raters) lack is an understanding of how their own language works, let alone others, including interlanguages. Their metalinguistic awareness is limited to a set of labels plus an ordinary native speaker's sensitivity to ungrammatical constructions, with a very limited capacity to explain why they are ungrammatical or what the logic behind grammaticality is. What needs to be stimulated is thus a different attitude towards language data, based on reasoning and understanding rather than on mere tagging and classifying.

Without such an attitude and the associated analytical competence, extensive or exclusive reliance on rating scales, even if based on SLA research findings, might prove to be limited or even misleading. In our experimentation, teachers were not given scales with level descriptors for different stages of interlanguage development. In fact, on the few occasions in which teachers received

CEFR scales for communicative competence, we noted that they were very happy with assigning learners to such prefabricated scales, which they did not take to be so different from traditional grading scales - they would say 'she is at B1' instead of 'she scored 6/10'. This however distracted them from the real objective, which was understanding students' linguistic strategies and their interlanguages' logic. Prefabricated descriptor scales might thus be used at some point in teachers' training - e.g. at the start for sensitizing them to the existence of 'typical' linguistic configurations at different developmental levels, or at the end as checklists - but their use should be limited. The same holds true for raters' training. While, for practical reasons, in their professional activity they may end up using prefabricated descriptor scales containing typical traits of different levels of interlanguage development, it is important that they reach a sound understanding of how interlanguages work in order to be able to apply such scales meaningfully.

References

Andersen, R., & Shirai, Y. (1996). The primacy of aspect in first and second language acquisition: The pidgin-creole connection. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 527–569). San Diego: Academic Press.

Bailey, A. (2008). Assessing the language of young learners. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of Language and Education: Vol. 7. Language testing and assessment* (2nd ed.) (pp. 273–284). New York: Springer.

Banfi, E., & Bernini, G. (2003). Il verbo. In A. Giacalone Ramat (Ed.), *Verso l'italiano: Percorsi e strategie di acquisizione* (pp. 70–115). Rome: Carocci.

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies, 14,* 281–299.

Berretta, M. (1986). Per uno studio dell'apprendimento dell'italiano in contesto naturale: Il caso dei pronomi atoni. In A. Giacalone Ramat (Ed.), *L'apprendimento spontaneo di una seconda lingua* (pp. 329–352). Bologna: Il Mulino.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33,* 1–17.

Cameron, L. (2001). Assessment and language learning. In L. Cameron, *Teaching languages to young learners.* Cambridge: Cambridge University Press.

Carpenter, K., Fujii. N., & Kataoka, H. (1995). An oral interview procedure for assessing second language abilities in children. *Language Testing, 12,* 157–181

Chini, M. (1995). *Genere grammaticale e acquisizione.* Milan: Angeli.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching, 5,* 161–170.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence'. *Language Testing, 21,* 259–283.

Ellis, N. (2006). Cognitive perspectives on SLA. *AILA Review, 19,* 100–121.

Gass, S., & Mackey, A. (2007). *Data elicitation for second and foreign language research.* Mahwah: Lawrence Erlbaum.

Giacalone Ramat, A. (2002). How do learners acquire the classical three categories of temporality? Evidence from L2 Italian. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense–aspect morphology* (pp. 221–248). Amsterdam: Benjamins.

Giannini, S. (2008). Proprietà formali e distribuzionali dei clitici in Italiano L2. In R. Lazzeroni, E. Banfi, G. Bernini, M. Chini, & G. Marotta (Eds.), *Diachronica et synchronica. Studi in onore di Anna Giacalone Ramat* (pp. 231–253). Pisa: ETS.

Giuliano, P. (2006). *Abilità narrativa ed emarginazione sociale.* Napoli: Liguori.

Hasselgreen, A. (2003). *Bergen 'Can Do' project.* Strasbourg: Council of Europe. Retrieved from http://www.ecml.at/documents/pub221E2003_Hasselgreen.pdf

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22,* 337–354.

Hendriks, H. (2008). Presenting the English Profile Programme: In search of criterial features. *Research Notes, 33,* 7–10.

Ioannou-Georgiou, S., & Pavlou, P. (2003). *Assessing young learners.* Oxford: Oxford University Press.

James, C. (1998). *Errors in language learning and use: Exploring error analysis.* Harlow: Longman.

Klein, W., & Perdue, C. (1992). *Utterance structure. Developing grammars again.* Amsterdam: Benjamins.

Klein, W., & Perdue, C. (1997). The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research, 13,* 301–348.

Ledda, F., & Pallotti, G. (2005). L'interlingua. In G. Pallotti & AIPI (Eds.), *Insegnare e imparare l'italiano come seconda lingua.* Rome: Bonacci.

Maffei, S. (2009). Osservazioni sullo sviluppo dei pronomi personali. In M. Palermo (Ed.), *Percorsi e strategie di apprendimento dell'italiano lingua seconda* (pp. 103–119). Perugia: Guerra.

Mayer, M. (1969). *Frog, where are you?* New York: Puffin.

McDaniel, D., McKee, C., & Smith Cairns, H. (1996). *Methods for assessing children's syntax.* Cambridge: MIT Press.

McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics,* **25***,* 243–263.

McKay, P. (2006). *Assessing young language learners.* Cambridge: Cambridge University Press.

Norris, J., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu: Second Language Teaching and Curriculum Center.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761).

Oxford: Blackwell.

Pallotti, G. (2007). An operational definition of the emergence criterion. *Applied Linguistics, 28,* 361–382.

Pallotti, G., & Peloso, A. (2008). Acquisition sequences and definition of linguistic categories. *The Open Applied Linguistics Journal, 1,* 77–89.

Papp, S., & Salamoura, A. (2009). An exploratory study into linking young learners' examinations to the CEFR. *Research Notes, 37,* 15–22.

Perdue, C. (1993). *Adult language acquisition: Cross-linguistic perspectives.* Cambridge: Cambridge University Press.

Philp, J., Oliver, R., & Mackey, A. (2008). *Second language acquisition and the younger learner: Child's play?* Amsterdam: Benjamins.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: Benjamins.

Rea-Dickins, P. (Ed.). (2000). [Thematic issue on assessing young learners]. *Language Testing, 17*(2).

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education: Vol. 7. Language testing and assessment* (2nd ed.) (pp. 257–271). New York: Springer.

Rea-Dickins, P., & Rixon, S. (1997). The assessment of young learners of English as a foreign language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of Language and Education: Vol. 7. Language testing and assessment* (pp. 151–161). Dordrecht: Kluwer.

Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics, 30,* 533–554.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10,* 209–231.

Serratrice, L. (2007). Referential cohesion in the narratives of bilingual English-Italian children and monolingual peers. *Journal of Pragmatics, 39,* 1058–1087.

Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375–409). San Diego: Academic Press.

Taylor, L., & Maycock, L. (Eds.). (2007). [Thematic issue on Cambridge Young Learners English (YLE) tests]. *Research Notes*, 28.

Valdez Pierce, L., & O'Malley, J. M. (1992). Performance and portfolio assessment for language minority students. *NCBE Program Information Guide Series*, 9, Spring.

Watorek, M. (2004). Construction du discours par des apprenants de langues, enfants et adultes. *Acquisition et Interaction en Langue Étrangère, 20,* 130–171.

Williams, C. (2008). Challenges to parsing English text: The language of non-native speakers. *Research Notes, 33,* 10–15.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity.* Honolulu: Second Language Teaching and Curriculum Center.

APPENDIX

Table 1. Interlanguage observation grid

Communicative competence	*Linguistic competence*			
	Noun system	**Verb system**	**Syntax**	**Lexicon**
Communicative efficacy	Noun and adjective morphology	Verb conjugation	Formulas	Variety, richness
Fluency	Noun phrase construction	Verb tense, aspect and mood	Word order in different types of constructions	Communication strategies
	Pronouns			

Table 2. List of descriptors and topics for the systematic observation of interlanguage*

What aspects are systematic? What regularities emerge? What **can** children do? This is not a list of items to be ticked with a simple yes or no, but a guide for systematic observation and analysis.

Communicative competence
Fluency
Does the learner express him- or herself easily, fluently, effortlessly?
- Can manage very short, isolated, mainly pre-packaged utterances, usually stimulated by teacher's prompts.
[...]
- Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech.
- Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.
[...]
- Has acquired a clear, natural, pronunciation and intonation.
Communicative efficacy
Can the learner convey ideas effectively? Achieve the goals he/she aims at? Avoid misunderstandings?
In conversations
- Can communicate with a few words and memorized patterns.
[...]
- Is entirely fluent in interaction, being able to manage it effectively.
In stories and descriptions
- Can tell a story or describe something in a simple list of points
[...]
- Can give clear, detailed descriptions of complex subjects.

>>>

* Due to space limitations, the original table (Gabriele Pallotti – Stefania Ferrari) has been abridged by omitting some descriptors of communicative competence.

Linguistic competence

NOUN SYSTEM

Noun and adjective morphology

Observe how nouns are inflected for gender (masculine and feminine) and number (singular and plural). Recall that number inflection has a meaning (it depends on the number of referents being talked about) while gender inflection is almost always arbitrary and must be learned by heart (what is masculine in the sun and feminine in the moon? In German the exact contrary is true).

Nouns ending in -e give special problems as they can be both masculine and feminine.

- Singular nouns: masculine and feminine?
- Plural nouns: masculine and feminine?
- Gender of nouns ending in -e?

Noun phrase construction

How is gender and number agreement marked? What elements - e.g. articles, demonstratives, possessives, adjectives - contribute to forming noun phrases, as in *i bambini intelligenti, le ragazze simpatiche, il cerchio giallo, la tazza rossa*?

Note agreement between article and noun (i*l bambino, i coltelli*), noun and adjective (*bambino allegro, coltelli gialli*) and article, noun and adjective (*il bambino allegro, i coltelli gialli*).

Several types of determiners exist beside the article: quantifiers (*qualche matita, molti colori*), numerals (*tre, cinque*), possessives (*il suo zaino, le loro borse*), demonstratives (*questa ragazza, quel libro*).

- Article/noun agreement
- Article noun/adjective

[...]

- Agreement in singular phrases
- Agreement in plural phrases

Are demonstratives used?

Are possessives used?

Pronouns

What pronominal forms are used? Note both free pronouns (*io, tu, lui, lei, noi*...) and clitics, which can express a direct (*me, te, lo, la, li*) or indirect object (*mi, ti, gli, le, ci, vi, gli*).

Also note if there are combined pronouns (*glielo, ce li, me la*) and clitics' position with respect to the verb (sometimes you may hear *io prendoli, voglio lo vedere*).

Finally, note clitic usage typical of substandard Italian: *a lei gli/ci dico.*

- Presence and usage of free pronouns
- Presence and usage of direct object clitics
- Presence and usage of indirect object clitics
- Combined pronouns
- Pronouns' position

>>>

VERB SYSTEM

Verb conjugation

How are different persons expressed? With one fixed form, with several forms or with the entire paradigm?

- Are verbs inflected? How?
- Some persons
- All persons (required by communicative demands)

Verb's tense, aspect and mood

How are notions of tense, aspect and mood expressed? What tenses, aspects and moods of the Italian verb system are used?

- Present
- Imperative
- Past participle
- Present perfect
- Imperfect
- Conditional
- Future
- Subjunctive
- Gerund
- *Stare* + gerund (progressive)
- Simple past

SYNTAX

Formulas

Are fixed formulas used, i.e. sentence chunks memorized as if they were a single word (e.g. *come si chiama? come stai? non ce l'ho, dammi, non lo so*)? Number, appropriateness, variety.

Negation

- No + X. (*no mangiare questo, no io così, no pane*)
- Non + X (*non mangio questo, io non faccio così, non c'è il pane*)
- Non ... mica, neanche ... (*non ha mica detto così, non ha neanche un soldo*)
- With indefinites (*niente, nessuno* ...)

Word order in different types of construction

How are sentences constructed? According to a canonical word order subject-verb-complement or with more complex orders? Observe for example:

- Post-verbal subject (*è arrivato Mario, sono caduti loro, si è spenta la luce*)
- Dislocations (*il libro non l'ho visto; non l'ho visto, il libro; a Roma ci sono già stato*)

>>>

Subordination

Are subordinate clauses produced? Which ones?

- Simpler ones (causal, temporal, final)
- More complex ones (relative, hypothetical, concessive) (if communicative situation requires them)

TEXT ORGANIZATION

How are sentences and parts of text connected?

- use of temporal (*poi, allora, dopo, mentre, alla fine*), argumentative (*però, invece, eppure*), meta-textual (*insomma, e tutto questo..., in poche parole*) connectives
- cohesion across different text parts, marked by pronouns and other pro-forms (*questo lo faccio solo la domenica*)

LEXICON

Variety, richness

Is the lexicon varied? Are terms used appropriately and precisely?

- Has a very basic repertoire of simple expressions for giving personal information and satisfying concrete needs
- Can use basic structures and memorized expressions or groups of few words to speak of him/herself or other persons, about ordinary actions, places and objects owned
- Controls sufficient language structures and lexicon to express him/herself, with some hesitations or circumlocution
- Can express him/herself clearly and briefly, but in a communicatively appropriate way, about everyday topics
- Has a rich linguistic repertoire, including a wide range of specific and appropriate terms, which can vary for style and register

Communication strategies to fill lexical gaps

Are particular communication strategies used to compensate for lack of specific terms?

- Repetition
- Reformulation/paraphrase (*la casa delle api, l'animale che salta*)
- Lexical invention (*il camionaio, matrimoniare*).
- Request for clarification/teacher's help
- Other

Table 3. Qualitative analysis of one structure (e.g. present perfect)

Examples displaying lack of knowledge	
Examples displaying difficulties and uncertainty	
Examples displaying knowledge	
Examples displaying general knowledge but problems in irregular/complex cases	
Examples displaying excellent knowledge, proficiency	
Examples of probably formulaic uses	

Table 4. Passato prossimo (present perfect)*

1) Read the transcript and underline in red all tokens of present perfect. When a form is not correct, besides underlining it add a cross next to it.
2) Use the 'score' column to mark a line every time you see the type of structure described in that line; when you reach five lines, draw a line across them to facilitate counting. At the end, you will report the sum in the 'total' column. If an example contains more than one error, score more than one line in different columns. For example, if the child says 'noi ha arrivato' you will score one on the line 'auxiliary *have* instead of *be*', one on the line 'no subject-auxiliary agreement' (noi ha) and one on the line 'no subject-participle agreement' (noi … arrivato).
   Examples should be provided for some of the errors, for structures demonstrating good knowledge, for self-corrections and for possible doubts.
   Don't score cases in which the form is repeated immediately after being uttered by another person.
   If the transcript is very long you can carry out quantitative analysis only on one part (but beginning and end should be clearly marked).

Passato prossimo	*Examples*	*Score*	*Total*
Corretti con participio –to (abbiamo mangiato, sono arrivato, è rimasto) *Correct with -to participle* ('we have eaten, I have arrived, he has remained')			
Corretti con participio –ta, -te, -ti (è tornata, siamo stati, l'ho vista) *Correct with –ta, -te, -ti participle* ('she has			
returned, we have been, I have seen her')			
Correct with irregular participle			
Correct with pluperfect			
Ausiliare + verbo non passato (ha mangia, sono torno) *Auxiliary + non past verb* (he has eat, I have return)			
*Wrong auxiliary choice (*to have *instead of* to be)			
*Wrong auxiliary choice (*to be *instead of* to have)			
Lack of auxiliary have			
Lack of auxiliary be			
No agreement between subject and past participle			
No agreement between subject and auxiliary			
Other non-standard uses(e.g. analogic constructions on irregular verbs)			
Doubtful, unclassifiable or uninterpretable cases			

* Due to space limitations, the original table has been abridged by giving the Italian original wording and examples in a few cells only.

Discourse connectives across CEFR-levels: A corpus based study

Cecilie Carlsen
University of Bergen

The chapter "Discourse connectives across CEFR-levels: A corpus based study" focuses on the use of discourse connectives, such as *and, but, so, then,* and *however,* in written learner texts of Norwegian as a second language. The *Common European Framework of Reference for Languages* (CEFR) makes specific predictions about the use of such discourse connectives in learner language, i.e. that that the range of different connectives expands across proficiency levels, that more advanced learners make use of less frequent connectives than learners at lower levels, and that learners gain increased control of connectives as they progress. The overall research question of the study reported in this chapter is whether the predictions made in the CEFR about learners' use of discourse connectives are supported by authentic learner data. The data used is a computer learner corpus of written Norwegian developed at the University of Bergen, Norway. This corpus has the great advantage of being linked to the CEFR. The study reported here is one small contribution to the huge task of validating the CEFR against real learner data, an overall aim of the SLATE network.

1. Introduction[1]

The present chapter focuses on the use of discourse connectives, such as *and, but, so, then,* and *however,* in written learner texts of Norwegian as a second language. The *Common European Framework of Reference for Languages*, CEFR, (Council of Europe, 2001) makes specific predictions about the use of such discourse connectives in learner language, elaborated in the illustrative scale of *Coherence and Cohesion* (p. 125). The CEFR predicts that the range of different connectives expands across proficiency levels, that more advanced learners make

1 I would like to thank Daniel Apollon at the University of Bergen for invaluable help with the correspondence analysis, the editors and two anonymous reviewers for useful comments on earlier drafts of this chapter, and Tania Horak, at Lancaster University, for proof reading. Any remaining errors are my own.

use of less frequent connectives than learners at lower levels, and that learners gain increased control of connectives as they progress. The overall research question of the study reported here is whether the predictions made in the CEFR about learners' use of discourse connectives are supported by authentic learner data. The predictions are tested against a computer learner corpus of written Norwegian (ASK)[2] developed at the University of Bergen, Norway. This corpus has the great advantage of being linked to the CEFR, which allows us to investigate what learners can and cannot do at different CEFR-levels[3]. The study reported here is one small contribution to the huge task of validating the CEFR against real learner data, an overall aim of the SLATE network.

There have been relatively few studies of coherence in writing in Norwegian as a second language. One such study is Høyte (1997), who investigates the relation between learners' use of connectives and test-scores, and finds a weak positive correlation between scores and the use of varied connectives. Høyte does not, however, investigate further what connectives are used at low versus high levels of proficiency. In another study Palm (1997) compares native speakers and non-native speakers of Norwegian, and finds that the latter group over-uses the common connective *fordi* [because], but her study is based on a limited number of informants. Similarly, McGhie (2003) investigates the use of causal connectives in oral and written production of learners of Norwegian and finds that learners use only a few of the available connectives to express these rhetorical relations. Like Palm (2007) she finds an overuse of *fordi* [because] in learner language and a lesser use of its counterpart *derfor* [therefore]. Her study, however, includes only five informants. Qualitative studies are necessary in order to achieve an in-depth understanding of the semantic-pragmatic content and poly-functional use of connectives (Blakemore, 2002; Mosegaard-Hansen, 1998). The study presented in this chapter is however of a different kind: The approach is corpus-based and quantitative throughout. The research purpose is to look at the use of a range of different connectives in learner language at different levels of proficiency to reveal patterns of over- and underuse which may not be easily generalized from the results of studies based on small data samples. The study focuses on the use of 36 different connectives

2 ASK is an acronym for the three constituent morphemes of Norwegian AndreSpråksKorpus [SecondLanguageCorpus] (see Tenfjord, 2007, p. 207).

3 During 2008/2009 I collaborated with Felianka Kaftandjieva at the University of Sofia, Bulgaria, in linking ASK to the CEFR. A group of 10 experienced raters was involved in the re-assessment of corpus-texts. 200 texts were scored by 10 raters, the remaining 1022 texts by two parallel groups of 5 raters, each group scoring 511 texts (see Carlsen, 2010, for details).

in learner language at seven levels of proficiency and compared with the use of connectives of native speakers of Norwegian. This approach does not allow an in-depth investigation of the semantics and use of each connective[4]. It is the first study of connectives in Norwegian learner language using a learner corpus linked to the CEFR.

2. Theoretical background

2.1. Text coherence and discourse connectives

Coherence is often described as that: "[...] which makes a discourse more than the sum of the interpretations of the individual utterances" (Sanders & Spooren, 1999, p. 235). Coherence in a written text refers to the linking of ideas to make it meaningful to readers (Lee, 2002). The skilled writer uses a variety of devices to construct coherent texts such as reference, substitution and ellipsis (Halliday & Hasan, 1976; Kehler, 2004). The use of explicit linking words or linking phrases is one way of signalling coherence relations. A text may however be coherent without explicit marking of coherence relations, but such relations often are marked linguistically, as Knott and Dale (1994) and Spooren and Sanders (2008, p. 2005) point out.

The present study does not set out to investigate all available coherence devices, but limits its focus to explicit linking words such as *and, but, because*, *however, despite, furthermore*, referred to in the following as discourse connectives, or simply connectives (see Blakemore, 2002; Schiffrin, 1987). The class of connectives consists of different linguistic elements and is therefore difficult to define strictly grammatically. Most studies dealing with connectives therefore are confined to a functional definition (Mosegaard-Hansen, 1998). Fraser (1996, p. 190) defines connectives as elements "which signal a relation between the discourse segment which hosts them, and the prior discourse segment". I will use a similar definition here, but similarly to Spooren and Sanders (2008, p. 2005), I also include elements that link larger text sections and paragraphs. Discourse markers that do not primarily have a linking function, such as *I mean, sort of, right, well, oh, you know, kind of* etc., often called pause fillers or hesitation markers mostly used in spoken language, are not included in the present study.

---

4 This is the focus of a quantitative and qualitative follow-up study in which I look at the relative effect of proficiency-level and cross-linguistic influence on the use of a more limited number of connectives in the writing of Spanish learners of Norwegian.

2.2. Coherence and the use of connectives in learner language

Cross-linguistic studies comparing the use of discourse connectives between two or more languages have shown that different languages tend to use different connectives to a somewhat different degree and with somewhat different meaning (Fabricius-Hansen, 2005; Fløttum, Dahl, & Kinn, 2008; Östman, 2005; Stenstrøm, 2006). It is therefore not surprising that constructing coherent texts poses problems to language learners, even at advanced levels (Connor, 1996; Lee, 2002). Even so, the use of discourse connectives in learner language has not been the focus of much research interest (Hinkel, 2001, p. 113; Müller, 2005, p. 1). The existing research on the topic is hard to compare due to different definitions of connectives, and the results are inconclusive.

In research focusing on the relation between the use of connectives and levels of proficiency, some studies have found small or no differences between learners at different levels of proficiency (Castro, 2004; Johnson, 1992), while others have found that highly rated essays are cohesively denser than poorly rated ones (Witte & Faigley, 1981). Similarly, some early studies of learner English (Evensen, 1985; Rygh, 1986) found a higher frequency and diversity of connectors in texts produced by advanced learners than in texts by learners at lower levels of proficiency.

Other researchers have compared the use of connectives in native and non-native speaker texts. Connor (1984) found no significant difference in general cohesion density between native speakers and advanced learners, while others have found that non-native writers of English overuse explicit cohesion markers as compared to native English writers (Field & Oi, 1992).

Several researchers investigating the use of high-frequency connectives have come to the conclusion that these are overused by learners. In a corpus-based study, Paquot (2008) for instance compares the use of five exemplifying lexical items between non-native speakers (the International Corpus of Learner English) and native speakers (two different native corpora) and finds a striking overuse of the connective phrase *for example* by non-native speakers. Similarly, in a corpus-based study of Swedish learners of French (InterFra corpus) Hancock (2005) finds that the high-frequency connective *parce que* [because] is overused by learners even at advanced levels. Müller (2005), on the contrary, finds that native speakers of English use the simple causal marker *so* twice as much as non-native speakers.

Finally, Hinkel (2001) compares the use of coordinating conjunctions in texts written by native speakers of English and learner groups with different first languages (L1s), and finds that some L1-groups have a similar use to native speakers, some groups significantly underuse, while other groups overuse connectives, which points to the importance of studying discourse features in relation to cross-linguistic influence as well.

2.3. Coherence, connectives and the CEFR

In the CEFR discourse competence is treated as one aspect of pragmatic competence and defined as "the ability of a user/learner to arrange sentences in sequence so as to produce coherent stretches of language" (Council of Europe, 2001, p. 123). Coherence and cohesion are mentioned among other criteria which need to be met in order to achieve straightforward and efficient communication. Four illustrative scales are available for discourse competence (Council of Europe, 2001, pp. 124–125). In this study the focus is on the illustrative scale of *Coherence and Cohesion*, which describes the use of organisational patterns and cohesive devices in the construction of coherent discourse (see Appendix, Table A1 for a reproduction of the illustrative scale of coherence and cohesion).

The scale of coherence and cohesion reflects a basic distinction in the description of language development in the CEFR, i.e. that between quantity and quality. The use of connectives across proficiency levels is, on the one hand, described in relation to the relative range of different connectives used: At the lower levels, only the "very basic connectors" are expected, "simple connectors" are expected at the A2-level and at the A2+ level "the most frequently occurring connectors". At the higher levels, the range of different connective devices is assumed to increase. It is worth noticing that it is only at the B2+ level that "a variety of linking words" is expected. Below B2+ level only "a limited number of cohesive devices" are expected. At this point, an important distinction needs to be made very clear: The CEFR predicts greater range but does *not* predict greater connective density at higher levels. It does not predict that the more advanced the learners get, the more overt signals of coherence relations they use. Since the main purpose of the study reported here has been to test the predictions of the CEFR, connective density has not been the focus of my study.

On the other hand, the CEFR describes the use of connectives in relation to the degree of control and efficiency with which they are employed: At the A1, A2 and B1 levels limited reference is made regarding the control of connectives, other than "can link…". At the B2-level, connectives are described as linking utterances into a "clear, coherent discourse, though there may be some "jumpiness" in a long contribution". At the B2+ level, there is explicit reference to the use of connectives as being "efficient", at C1 as being "controlled" and finally at the C2-level: "full and appropriate".

3. Aim of the study

The overall research question of the study reported here is whether the CEFR's description of learners' use of connectives is supported by empirical learner data. Based on the level descriptors of the illustrative scale of coherence and cohesion, three main predictions about the use of connectives may be deduced. Firstly, the *range* or repertoire of cohesive devices is assumed to grow across CEFR-levels. If the scale's predictions are correct, learners at higher levels of proficiency should utilise a greater variety of different connectives than learners at lower levels. Secondly, learners at lower levels of proficiency are assumed to rely heavily on the use of common, high-frequency connectives, while learners at higher levels are assumed to use *low-frequency* connectives as well. And finally, learners at higher levels of proficiency show qualitatively better *control* of cohesive devices than learners at lower levels[5]. The above predictions have been reformulated into three hypotheses, which are tested empirically in this study:

H1: Texts at higher levels contain a *broader range* of different cohesive devices than texts at lower levels

H2: Texts at higher levels contain more *low-frequency* connectives than texts at lower levels

H3: Texts at higher levels show a *greater degree of control* of the cohesive devices used than texts at lower levels

4. Data and methodology

In order to investigate the predictions made in the CEFR about the use of connectives, it was necessary to include a wide range of different connectives in the study. Since no definitive list of discourse markers exists (Blakemore, 2004, p. 221), I had to develop a list based on earlier taxonomies of connectives (Halliday & Hasan, 1976; Knott & Dale, 1994) and descriptions in Norwegian grammars (Faarlund, Lie, & Vannebo, 1997; Hagen, 1998). I wanted to include a range of different connectives representing different rhetorical functions as

5 A fourth prediction could also be made; i.e. the reference to specific connectives (*and, then, but, because*) at the A1- and A2-levels. Since the target language is Norwegian and not English, and since there are no A1-texts in the ASK-corpus at present, this prediction has not been tested here.

mentioned in Halliday and Hasan's (1976, pp. 242–243)[6] taxonomy, e.g. *additive, adversative* and *causal*, single- as well as multi-word units, and connectives with different degrees of frequency. The frequency of the different connectives was established by investigating their use in texts written by native speakers of Norwegian selected from the control corpus of ASK, which contains texts written by 200 native speakers of Norwegian on the same prompts and under the same circumstances as the other texts in ASK (native speaker (NS)-ASK). Connectives that were not used in the NS-ASK were excluded from investigation. The connectives were divided into three frequency groups: High-frequency, n = 5 (relative frequency > 1.0), Medium-frequency, n = 19 (relative frequency < 0.9999 and > 0.0100), and Low-frequency, n = 12 (relative frequency < 0.0100). The total number of connectives included in this study is 36 (see Table 1 below for the categorization and translation of the Norwegian connectives into English[7]).

The data used in this study are selected from the electronic learner corpus of Norwegian (Norsk Andrespråkskorpus, ASK, see Tenfjord, 2007). ASK contains texts written by learners of Norwegian with 10 different L1s (Albanian, Bosnian/Serbian/Croatian, Dutch, English, German, Polish, Russian, Spanish, Somali and Vietnamese). The informants are adult foreigners living in Norway, making the corpus one of second and not foreign language. The texts are authentic test responses (essays) selected from two different standardized tests of Norwegian as a second language, one at the intermediate level and one at the advanced level. At the intermediate level, learners are asked to "write a text" about everyday themes, for example traditions, values, friendship, the place you live etc. At the advanced level, learners are asked to discuss and develop an argument in relation to themes such as education, integration, welfare, pollution, labour etc. The text types or genres are somewhat different at the two levels, mainly descriptive/expository at the intermediate level and expository/argumentative at the advanced level. The different genres are therefore likely to be reflected in the rhetorical function of the connectives used: More adversative

6 Halliday and Hasan's taxonomy also includes temporal connectives, which has been largely ignored in my study: Many of them have non-connective homonyms in Norwegian, which makes them notoriously difficult to study in a quantitative study such as the present one. Secondly, distinguishing between items referring to external events and connectives referring to the order of the text itself requires a qualitative approach.

7 The translation is tentative and merely linguistic since different languages use connectives in somewhat different ways, as cross-linguistic studies referred to here show.

and causal connectives are to be expected at the higher levels of proficiency. Genres should not, however, affect the hypotheses tested in this study since the use of varied connectives, low-frequency connectives, and control of connectives may be observed within each rhetorical group of connectives.

The corpus texts are automatically tagged for word class and morphological traits and manually error-coded, which allows searches of words and word-combinations, of incorrect as well as of correct forms. During the year of 2008/2009 two thirds of the texts in ASK (1222 texts) representing all texts of 7 L1s were subject to a re-assessment on the CEFR-scale by a group of 5-10 experienced raters who are very familiar with the CEFR levels (Carlsen, 2010). A series of different reliability estimates were calculated, such as *Homogeneity index* (Mean +0.84), *Correlation with the rest* (+0.90), and *Inter-rater correlation* (Mean +0.82), all well within an acceptable range in terms of rater agreement. Whole levels (A2, B1, B2, and C1) as well as in-between levels (A2/B1, B1/B2, and B2/C1) were used[8]. The size of the different CEFR-groups varies from 137,885 words in the B2-group to 6,115 words in the A2-group (see Appendix, Table A2 for CEFR-group size).

The study reported in this chapter uses a quantitative method and is to a large degree based on investigations of frequency of use. The mentioned difference in size of the CEFR-groups makes it necessary to use relative rather than absolute frequencies. Relative frequencies are calculated automatically in ASK by dividing the absolute frequency of occurrences by the number of words within each level group. The occurrences of connectives across the CEFR-levels (see H1 and H2) were investigated through the means of *correspondence analysis*, which is an exploratory technique designed to analyse the relation between rows and columns in a two- or multi-way table. The results are usually displayed as a scatter plot, in which the relations between, for example, variables on the one hand and observations on the other are visualized jointly as points in a common coordinate system. If the original data set is high dimensional, the reduction obtained in a much lower dimensional space may offer substantial advantages for interpreting the latent structure of the data set. Correspondence analysis is therefore a useful tool to get an overview of patterns in a data set.

The final hypothesis, H3, should ideally have been investigated qualitatively. This has not been practically possible due to the large number of connectives included. I have therefore used a quantitative approach based on the error-cod-

8 There were no texts in ASK found to be at levels below A2 or above C1. ASK is however currently being expanded by adding more texts at the A1/A2, A2, and A2/B1 levels.

ing inherent in ASK. Based on the results of the correspondence analysis, there are good reasons to ignore the high frequency connectives, which, as we will see, are used similarly across CEFR-levels, and focus attention on the medium- and low-frequency connectives. To test H3, only connectives that are actively used by all groups have been included, since control of use cannot be observed unless there are actual occurrences in the learner texts. Error searches were restricted to *W- wrong word choice,* while errors in spelling, morphology or syntax were ignored. The results are presented as errors across the number of total occurrences and as relative frequencies of errors, calculated by dividing the number of errors by the number of total occurrences within each CEFR-group (see Appendix, Table A3).

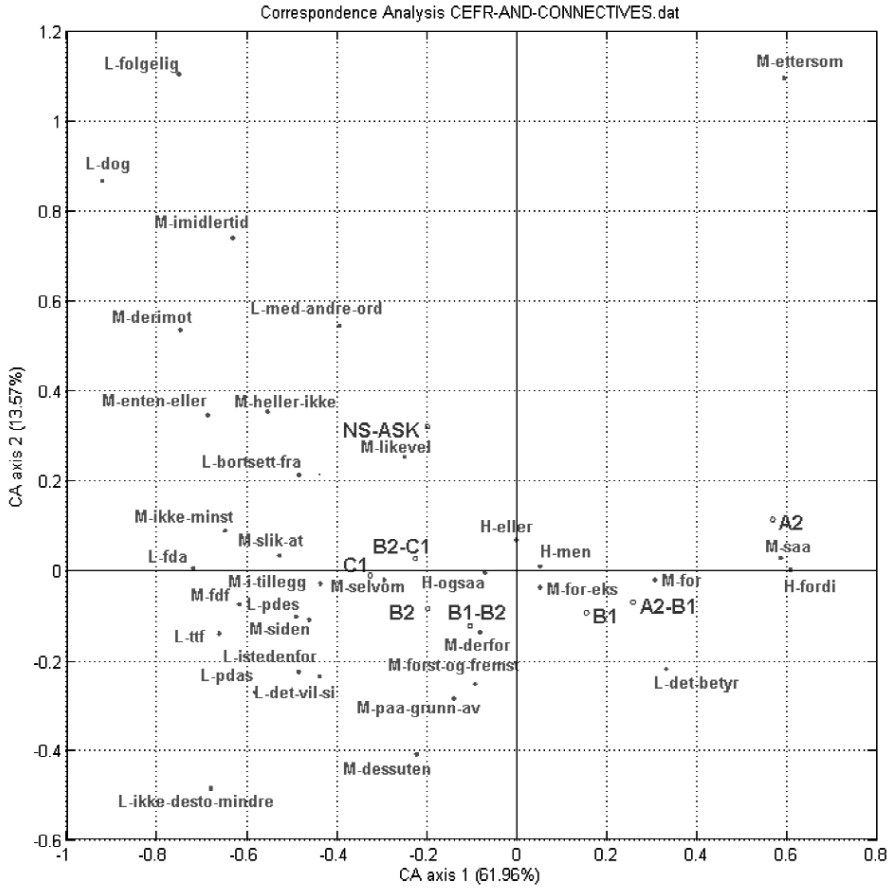
5. Analysis and interpretation

The results of the correspondence analysis are presented in Figure 1 below[9]. Since connectives are used somewhat differently across different languages, a translation of the connectives in the figure to English is not straightforward, even though it would make the figure more understandable to the general reader. I have therefore chosen to keep the Norwegian labels in the figure, making a translation available in Table 1 presented below. The table also explains the abbreviations of some of the connective phrases necessary in order to make Figure 1 more readable. The most important information in relation to the CEFR and the hypotheses of this study is not on the level of the individual connectives, but rather on the group-level based on frequency: H (high-frequency), M (medium-frequency) and L (low-frequency), added in front of each connective in Figure 1.

The scatter plot in Figure 1 displays the relation between the two variables of the dataset, CEFR-levels and connectives. The CEFR-levels order the use of connectives along a gradient stretching roughly from east (lowest proficiency values) to west (highest proficiency values). The vertical line running through the barycentre (centre of mass, 0) between B1 and B1/B2 discriminates efficiently between patterns of use typical of lower level groups (A2 to B1) and higher level groups (B1/B2 to C1). The clustering of upper level groups (B1/B2-C1) indicates only minor differences within these groups when it comes to the use of connectives. The more advanced learners' use of connectives is sim-

9 The most frequently used connective by all groups, *og* [and], is not included in the figure. It is used similarly across proficiency groups.

Figure 1. Results of the correspondence analysis



ilar to that of the native speakers (NS-ASK), with the exception of three connectives located in the upper-left corner of the scatter plot. Connectives near the barycentre, such as *men* [but], *eller* [or], and *også* [too/also] are common to all groups and do not differentiate between higher and lower levels. 75.52% of the variance (inertia) of the original data set is explained by two dimensions (axis 1: 61.96%, axis 2: 13.57%) indicating a high degree of correlation between CEFR-levels and the use of connectives, and a clearly different use of connectives at lower versus higher levels of proficiency.

The prediction of H1 was that texts at higher levels would contain a *broader range* of different cohesive devices than texts at lower levels. This prediction is supported by the data. The scatter plot demonstrates that a great number of different connectives clustered around the higher CEFR-levels to the left in the

Table 1. Connectives in frequency-groups based on NS' use of connectives, with translation into English

FUNCTION	HIGH-FREQ.		MEDIUM-FREQUENCY		LOW-FREQUENCY	
ADDITIVE	*eller* *også*	or too/also	*for eksempel* *i tillegg* *enten..eller* *ikke minst* *heller ikke* *først og fremst* *for det første* *(fdf)* *dessuten*	for example in addition either….or not least nor first and foremost firstly besides	*det vil si* *med andre ord* *for det andre* *(fda)* *bortsett fra* *det betyr*	i.e./this means in other words secondly except from this means
ADVERSATIVE	*men*	but	*likevel* *selv om* *derimot* *imidlertid*	still/neverthele ss even though on the other hand however	*istedenfor* *til tross for* *(ttf)* *dog* *på den ene siden* *på den andre siden* *ikke desto mindre*	instead of despite though on the one hand on the other hand nevertheless
CAUSAL	*fordi*	because	*derfor* *slik at* *så (saa)* *på grunn av* *siden* *ettersom* *for*	therefore so that so because of since since for	*følgelig*	consequently

picture, and only a few to the right, which indicates that learners at higher levels of proficiency use a series of different connectives, while the lower-level learners confine themselves mainly to just a few connectives. One important point should be made: The CEFR predicts that even at a B2-level only a limited number of cohesive devices are used. The results of the present study clearly show, however, that learners even at a B1/B2 level use a range of different connectives, and to an extent which separates them sharply from the lower levels.

H2, which implies that learners at higher levels of proficiency use more *low-frequency* connectives than learners at lower levels, is also supported by the data. Around the lower level groups on the right side of the vertical line, there are mainly high- and medium-frequency connectives, while there are a great number of low-frequency connectives clustered around the more advanced learners to the left. The high-frequency connectives, except H-*fordi* [because], and H-*også* [too/also], are all clustered around the barycentre and common to all groups. Low-frequency connectives are generally under-used at A2, A2/B1

and B1-levels, with the exception of the low-frequency connective phrase L-*det betyr* [this means] which is overused at the A2-level.

According to the final hypothesis, H3, learners at higher levels of proficiency show a *greater degree of control* of the cohesive devices used (see Appendix, Table A3). The results of the error-searches show that there are in fact relatively few cases where learners use connectives wrongly, in the sense that they use one connective where they should have used another one instead. However, the relative frequency of errors is strongly correlated with CEFR-level (Chi-square test: $p<0.001$ level), yielding support for the prediction made in the CEFR that the degree of control of cohesive devices rises across proficiency levels. An interesting point worth mentioning here is the fact that the error-pattern differs across the rhetorical function of the connectives: The additive connectives are used correctly in all instances by the B2, B2/C1 and C1 groups, and also in the groups of A2 and A2/B1, and there are only occasional errors in the B1 and B1/B2 groups. The adversative and causal connectives on the other hand, have a somewhat higher error rate. The connectives that seem to cause the most problems to learners in my study are the adversative connectives *derimot* [however/contrary] and *selv om* [even though], as well as the causative connective *derfor* [therefore].

In this study the main focus has not been on the individual connectives, yet the study has shed some light on the use of particular connectives across CEFR-levels. Most of the 36 connectives can be divided into three groups according to the pattern they show across the CEFR-levels: Some connectors show a steep negative correlation across levels of proficiency. They are largely over-used at low levels of proficiency and fall gradually as one approaches the C1-level. This is the case for the high-frequency connectives *men* [but], and *fordi* [because] (the latter is used six times as frequently by the A2-group as by the C1 group), for the medium-frequency connectives *så* [so], and *for* [because], as well as for the low-frequency connective phrase *det betyr* [this means] (see Appendix, Figures A1 and A2). Other connectives show an opposite pattern, i.e. they are only used to a limited degree or not used at all at low levels of proficiency and increase with higher levels of proficiency. This is the case for a range of medium-frequency connectives, for example *i tillegg* [in addition], *slik at* [so that], *ikke minst* [not least], *derfor* [therefore] and *enten…eller* [either…or] (see Appendix, Figures A3 and A4). Finally, quite a few connectives increase in use from A2 to B2 - B2/C1 levels and then drop as one approaches the C1-level. Many low-frequency connectives show this pattern, for example *istedenfor* [instead of], *på den ene siden* [on the one hand], *på den andre siden* [on the other hand], *ikke desto mindre* [nevertheless], *til tross for* [despite], and *med andre ord* [in other words] (see Appendix, Figures A5 and A6).

6. Discussion and conclusions

The results of the study reported here are not surprising as they largely support the predictions made in the CEFR. Still, some of the findings may need further comment. As predicted by H1, learners at higher levels of proficiency tend to use a greater range of different connectives in their writing than learners at lower levels. This finding is in line with Evensen (1985) and Rygh (1986). The data of the present study do however show that learners use a range of different connectives earlier than predicted in the CEFR: Learners even at B1/B2-level use a range of different connectives contrary to the predictions of the CEFR, which only expects this at the B2 + level. This finding may indicate that a revision of the CEFR-scale is warranted at this point.

H2 was also largely supported by the data. Indeed, learners at lower levels tend to overuse high-frequency connectives, like *fordi* [because], *men* [but], *eller* [or] and medium-frequency connectives like *så* [so], and *for* [because]. This phenomenon is well-known from other studies of connectives (Hancock, 2005; Paquot, 2008). It is tempting to borrow a term from a former colleague and associate of the SLATE-network, Angela Hasselgreen, who refers to this phenomenon in learners' use of vocabulary as "lexical teddy bears" (Hasselgreen, 1994). High frequency connectives or "connective teddy bears" give novice learners a degree of comfort and security particularly useful in a test-situation like the one from which the texts in the ASK-corpus are selected. It is likely that the overuse of high-frequency connectives is due in part to learners using one and the same connective expressing different rhetorical functions. The obvious limitation of a quantitative study as the one presented here is the lack of information about the use and content of the different connectives in learner language, and it is therefore important to complement studies like the present one with qualitative investigations.

The data also support H3 indicating that learners gain increased control of the connectives they use across levels of proficiency. There are only a few mistakes in their use but still, the lower-level groups make significantly more errors than advanced learners. The data of the present study do not lend themselves to investigating the reasons why there are more errors in the use of adversative and causal connectives than in the use of additive connectives. The rhetorical functions expressed by adversative and causal connectives are more complex than a mere adding of information, which may explain some of the difference. In addition, the adversative and causal connectives included in the study of H3 have more specific content and use, which makes errors easier to spot. Again, these questions need to be addressed empirically in a qualitative study.

Finally, the results of the study indicate that some low-frequency connectives are used gradually more frequently towards higher levels of CEFR, but

drop at the highest levels. A possible explanation may be that while learners at B2 and B2/C1 levels use explicit low-frequency markers consciously when structuring their texts, learners at C1 level use other devices to express coherence. As mentioned earlier, the use of explicit coherence markers like connectives is but one way of constructing coherent texts. The mere presence of connectives does not necessarily make a text coherent, and indeed, it is possible to construct a coherent text with limited use of explicit markers of coherence relations. The finding illustrates that coherence-relations need to be investigated qualitatively as well as quantitatively to grasp the full picture.

References

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse connectives.* Cambridge: Cambridge University Press.

Blakemore, D. (2004). Discourse markers. In R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 221–240). Malden: Blackwell.

Carlsen, C. (2010). Linking a learner corpus to the Common European Framework of Reference. Manuscript submitted for publication.

Castro, C. D. (2004). Cohesion and the social construction of meaning in the essays of Filipino college students' writings in L2 English. *Asia Pacific Education Review, 5,* 215–225.

Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Research on Language & Social Interaction, 17,* 301–316.

Connor, U. (1996). *Contrastive rhetoric. Cross-cultural aspects of second-language writing.* Cambridge: Cambridge University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Evensen, L. S. (1985). Discourse-level interlanguage studies. In N. E. Enkvist (Ed.), *Coherence and composition. A symposium* (pp. 39–65). Åbo: Åbo Akademi.

Faarlund, J. T., Lie, S., & Vannebo, K. I. (1997). *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.

Fabricius-Hansen, C. (2005). Elusive connectives. A case study on the explicitness dimension of discourse coherence. *Linguistics, 43,* 17–48.

Field, Y., & Oi, Y. L. M. (1992). A comparison of internal conjunctive cohesion in English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*(1), 15–28.

Fløttum, K., Dahl, T., & Kinn, T. (2008). *Academic voices.* Amsterdam: John Benjamins.

Fraser, B. (1996). Pragmatic markers. *Pragmatics 6*(2), 167–190.

Hagen, J. E. (1998). *Norsk grammatikk for andrespråkslærere.* Oslo: Ad Notam Gyldendal.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hancock, V. (2005). Étude intonodiscursive des sequences introduites par parce que dans le francais parlé des apprenant avancés suedophones. In J. M. Delefosse (Ed.), *Actes du colloque international organisé les vendredi 25 et samedi 26 juin 2004 à Paris. Acquisition, pratiques langagières, interactions et contacts (APLIC)* [CD-ROM]. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-10149

Hasselgren, A. M. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with vocabulary. *International Journal of Applied Linguistics*, *4*, 237–258.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning, 12*, 111–132.

Høyte, T. (1997). *Tekstkoherens og tekstkvalitet* (Unpublished master's thesis). University of Bergen, Norway.

Johnson, P. (1992). Cohesion and coherence in compositions in Malay and English. *RELC Journal, 23*(2), 1–17.

Kehler, A. (2004). Discourse coherence. In L. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 241–265). Malden: Blackwell.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes, 18,* 35–62.

Lee, I. (2002). Teaching coherence to ESL students: A classroom inquiry. *Journal of Second Language Writing, 11*(2), 135–139.

McGhie, M. (2003). Årsaksrelasjoner i tilegnelse av norsk som andrespråk. In S. Lie, G. Nedrelid, & H. Omdal (Eds.), *MONS 10. Utvalde artiklar frå det tiande møte om norsk språk* (pp. 207–217). Kristiansand: Høyskoleforlaget.

Mosegaard-Hansen, M.-B. (1998). The semantic status of discourse markers. *Lingua, 104,* 235–260.

Müller, S. (2005). *Discourse markers in native and non-native English discourse.* Amsterdam: John Benjamins.

Östman, J.-O. (2005). Constructions in cross-language research. Verbs as pragmatic particles in Solv. In K. Aijmer & A.-M. Simon-Vandenbergen (Eds.), *Pragmatic markers in contrast* (pp. 237–257). Amsterdam: Elsevier.

Palm, K. (1997). *Argumenterende skriving. En analyse av andrespråkstekster og førstespråkstekster fra videregående skole* (Unpublished master's thesis). University of Oslo, Norway.

Paquot, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier & S. Granger (Eds.), *Phraseology. An interdisciplinar perspective.* Amsterdam: John Benjamins.

Rygh, I. L. (1986). Connector density? an indicator of essay quality? In L. S. Evensen (Ed.), *Nordic research in text linguistics and discourse analysis* (pp. 203–213). Trondheim: Tapir.

Sanders, T., & Spooren, W. (1999). Communicative intentions and coherence relations. In W. Bublitz, M. Lenz, & E. Ventola (Eds.), *Coherence in spoken and written discourse* (pp. 235–350). Amsterdam: John Benjamins.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Spooren, W., & Sanders, T. (2008). The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics, 40,* 2003–2026.

Stenström, A.-B. (2006). The Spanish discourse markers o sea and pues and their English correspondences. In K. Aijmer & A.-M. Simon-Vandenbergen (Eds.), *Pragmatic markers in contrast* (pp. 155–172). Amsterdam: Elsevier.

Tenfjord, K. (2007). ASK and you will find what you seek. In C. Carlsen & E. Moe (Eds.), *A human touch to language testing* (pp. 198–208). Oslo: Novus Press.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication, 32,* 189–204.

APPENDIX

Table A1. The illustrative scale of Coherence and Cohesion (Council of Europe, p. 125).

	COHERENCE AND COHESION
C2	*Can create coherent and cohesive text making **full and appropriate use** of a variety of organisational patterns and a **wide range of cohesive devices**.*
C1	*Can produce clear, smoothly flowing, well-structured speech, showing **controlled use** of organisational patterns, **connectors and cohesive devices**.*
B2	*Can use a **variety of linking words efficiently** to mark clearly the relationships between ideas.*
	*Can use a **limited number of cohesive devices** to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.*
B1	***Can link** a series of shorter, discrete simple elements into a connected, linear sequence of points.*
A2	*Can use **the most frequently occurring connectors** to link simple sentences in order to tell a story or describe something as a simple list of points.*
	*Can link groups of words with **simple connectors** like "and", "but" and "because".*
A1	*Can link words or groups of words with **the very basic linear connectors** like "and" or "then".*

Table A2. Number of words in the different CEFR-groups in ASK

Number of words	CEFR-level
6137	A2
52221	A2/B1
92334	B1
89617	B1/B2
137413	B2
41383	B2/C1
11614	C1

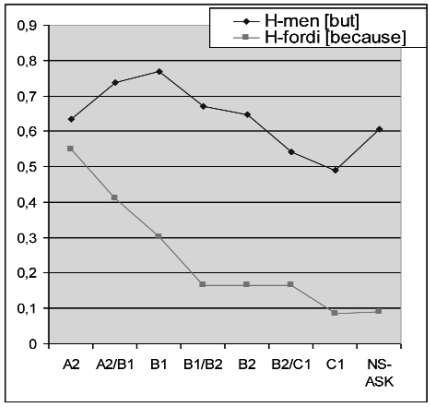
Table A3. Erroneous use of a sample of the 36 connectives across CEFR-levels*

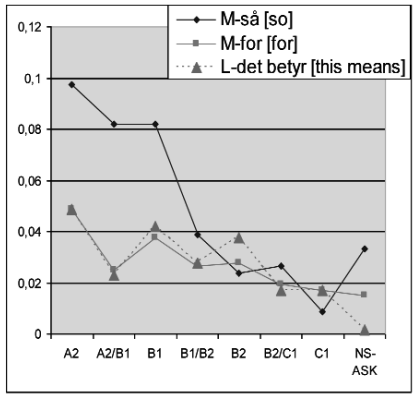
FUNCTION	**L- and M-freq. connectives used by all groups**	**A2, A2/B1**	**B1, B1/B2**	**B2, B2/C1, C1**
ADD.	M-for eksempel/ **for example**	0/56	1/177	0/181
	[rel.freq.]	[0]	[0,0056]	[0]
	M-i tillegg/ **in addition**	0/11	0/58	0/116
	[rel.freq.]	[0]	[0]	[0]
	M-først og fremst/ **first and foremost**	0/16	1/81	0/93
	[rel.freq.]	[0]	[0,0123]	[0]
	M-dessuten/ **besides**	0/15	1/77	0/79
	[rel.freq.]	[0]	[0,0129]	[0]
	L-det vil si/ **this means**	0/5	0/25	0/41
	[rel.freq.]	[0]	[0]	[0]
	L-det betyr/ **this means**	0/14	0/65	0/61
	[rel.freq.]	[0]	[0]	[0]
ADV.	M-likevel/ **still, however**	1/17	1/59	0/102
	[rel.freq.]	[0,0588]	[0,0169]	[0]
	M-selv om/ **even though**	2/31	2/134	2/167
	[rel.freq.]	[0,0645]	[0,0149]	[0,0119]
	M-derimot/**on the other hand**	1/1	1/9	2/37
	[rel.freq.]	[1.0000]	[0,1111]	[0,0540]
	L-til tross for/ **despite**	0/1	0/15	0/15
	[rel.freq.]	[0]	[0]	[0]
CAUS.	M-derfor/ **therefore**	3/83	1/246	1/272
	[rel.freq.]	[0,0361]	[0,0040]	[0,0036]
	M-på grunn av/ **because of**	0/33	1/141	0/123
	[rel.freq.]	[0]	[0,0070]	[0]
	M-ettersom/ **since**	0/2	0/1	0/4
		[0]	[0]	[0]
ERRORS TOTAL				
REL. FREQ.		**1,1594**	**0,1846**	**0,0695**

* The total number of occurrences is presented to the right of the slash, while the number of errors in use of the particular connective is presented to the left, for example *selv om* [even though] is used 31 times in the lowest proficiency group, 134 times in the B1 and B1/B2 groups and 167 times by the most proficient learners. In each proficiency group there are two occurrences where the connective is used incorrectly in the sense that a different connective should have been chosen instead. The number of errors (wrong choice of connective) is divided by total occurrences of the particular connective within the CEFR-level group to obtain the relative frequency of errors. Since the connective *selv om* is used five times as often in the texts of the more advanced learners, the relative frequency of errors is smaller [0.0119] in this group than in the lowest proficiency group [0.0645]. The last row displays the total occurrence of errors in each group divided by the actual occurrences within each group, i.e. it is the sum of the relative frequencies for each group.

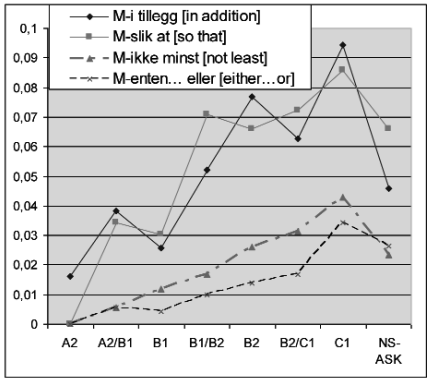
Even though the number of errors is not high for any group, there is a positive correlation between levels of proficiency and control when defined as the lack of errors in use. The difference between level groups for total errors is significant at the p>0.001 level (Chi-square "Goodness of Fit" Test).

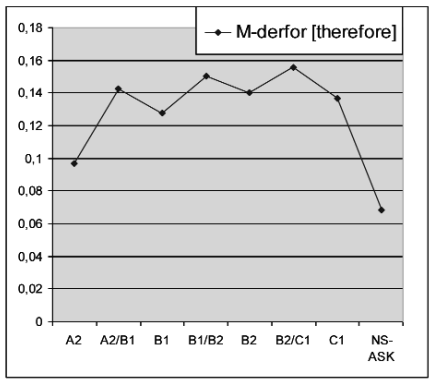
Figures A1 and A2: Connectives overused at lower levels (NS-ASK to the right)



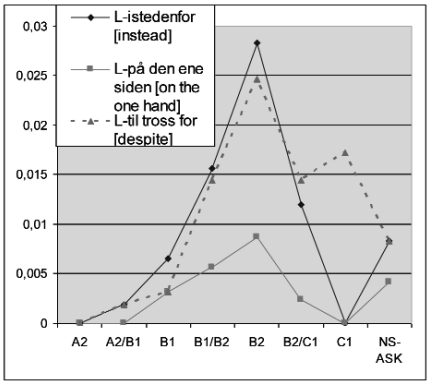


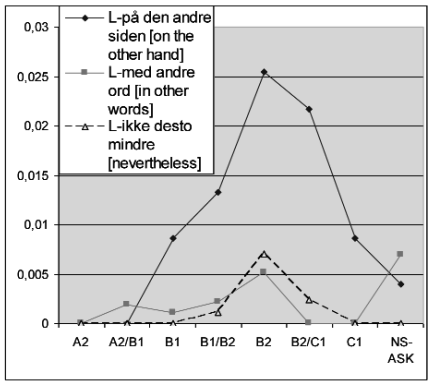
Figures A3 and A4: Connectives underused at lower levels (NS-ASK to the right)





Figures A5 and A6: Connectives rise in use, drop at B2-B2/C1 levels (NS-ASK to the right)





The development of vocabulary breadth across the CEFR levels.

A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe.

James Milton
Swansea University, UK

This chapter attempts to attach measurements of vocabulary breadth, the number of words a learner knows in a foreign language, to the six levels of the Common European Framework of reference for Languages (CEFR). The details of the Framework document (Council of Europe, 2001) indicate that vocabulary breadth ought to be a useful metric in the description of the levels and that, broadly, it would be expected that as language level increases so would the learner's knowledge of vocabulary and the sophistication with which that vocabulary can be used. The evidence we have from vocabulary size tests is reviewed and confirms this assumption, and suggests the actual volumes of vocabulary that are associated with each CEFR level. This information should be very useful to learners, teachers and other users of the CEFR is helping to link language performance to the CEFR levels. The evidence also appears to suggest that vocabulary breadth may vary from one language to another but it is not yet clear whether this reflects differences between the languages themselves, or differences in the construction of the corpora from which vocabulary size tests are derived.

1. Introduction

This chapter addresses the principal aim of SLATE, which is to determine 'which linguistic features of learner performance (for a given target language) are typical at each of the six CEFR levels?' (see Hulstijn, Alderson, & Schoonen, this volume; see also "Aims of SLATE," n.d.). It attempts to identify, the scale of vocabulary knowledge which is typical at each of the six levels of the Common European Framework of Reference for foreign languages (CEFR). It addresses, therefore, an issue which the creators of the CEFR themselves raise in pointing out that 'users of the Framework may wish to consider … what size of vocabulary (i.e. the number of words and fixed expressions) the learner will need…' in seeking to attain a particular level of performance (Council of Europe, 2001, p. 150). And the CEFR document further suggests, 'an analysis

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 211-232

of the … vocabulary necessary to perform the communicative tasks described on the scales could be part of the process of developing a new set of language specifications' (Council of Europe, 2001, p. 33). In addressing this issue, therefore, this chapter also addresses the second of the research issues SLATE identifies and attempts to contribute to a linguistic tool kit for diagnosing learners' proficiency levels by examining the number of words in their foreign language that learners at each CEFR level typically know. This is potentially very useful for teachers and learners and will make the process of assigning learners to CEFR levels quicker and, potentially, more accurate. It should help, too, to make the CEFR more robust by adding detail to the levels descriptors.

This chapter will begin by considering what the CEFR framework says about vocabulary knowledge and the way it is expected to develop as learners improve in competence. Broadly, this suggests that language learners, as they progress through the levels of the CEFR, will grow increasingly large, and increasingly complex, lexicons in the foreign language. This relationship between vocabulary knowledge and overall competence in a foreign language is supported by research that suggests that vocabulary knowledge is key to both comprehension and communicative ability (e.g. Stæhr, 2008). While vocabulary knowledge and general linguistic performance are separable qualities, given that the number of words a learner knows is not the sole determiner of how good he or she is in communication, they are not entirely separate qualities. A learner's vocabulary can be expected to become measurably larger and more sophisticated as communicative competence increases. The potential for this as a diagnostic tool is obvious since if vocabulary knowledge can be measured, then learners may be quickly and easily linked to the relevant CEFR level. Such a measure would not provide details of every aspect of linguistic performance, of course, but might in addition to providing a placement within the framework for vocabulary knowledge be a useful general measure. The methodology for measuring vocabulary knowledge will be explained and this involves an understanding of what is meant by 'word' in this context. Current methodology allows the numbers of words learners know in a foreign language to be estimated with some confidence, and these measurements appear particularly useful in making broad assessments of learner level. The measurements we have of vocabulary size and which are linked to the CEFR levels will be presented and examined.

2. Vocabulary within CEFR descriptors

Some of the early materials relating to the CEFR contained great detail about the vocabulary associated with performance at some of the six levels. At what is now called the B1 level, given several names at the time such as *Threshold* and

Niveau Seuil, there are several word lists available for this level (for example, Coste, Courtillon, Ferenczi, Martins-Baltar, & Papo, 1987; Van Ek & Trim, 1991). These lists typically contain about 2000 words. At what is now A2 level, called *Waystage* at the time in English, materials also included wordlists (for example Van Ek, 1980) and these were, as might be expected, smaller in size than the B1 level lists with about 1000 words. In each case the wordwere derived from notional functional areas which were deemed appropriate to these levels, such as clothing and what people wear, personal identification, and routines in daily life. Adumbrating the words that should be known in word lists had the serious drawback, however, of prescribing the language for each level in a way that restricted the flexibility of the system and its ability to be applied across the huge variety of language courses and language learning that takes place in Europe, and even across the different languages that are used in Europe. The 2001 CEFR document makes the argument that 'descriptors need to remain holistic in order to give an overview; detailed lists of micro-functions, grammatical forms and vocabulary are presented in language specifications for particular languages (e.g. Threshold level, 1990)' (Council of Europe, 2001, p. 30). The word lists have not been abandoned or disowned in anyway by the CEFR, therefore, but a different and more all-inclusive approach to language description has been adopted. Current descriptions of the CEFR level have, therefore, defined the levels in terms of skills, language activities or communicative goals (Council of Europe, 2001). The current descriptions are flexible and inclusive and by being general they can apply across different languages more readily than the separate lists for individual languages were capable of doing.

The new levels descriptors sometimes include reference to the vocabulary that might be expected of learners performing certain skills and this is illustrated in samples of A1 and B1 level descriptors, provided in Table 1, which are taken from the Council of Europe's (2001) description of the CEFR. These include, in the A1 listening and reading descriptors, reference to the recognition and comprehension of 'familiar words', and in the B1 reading descriptors reference to the understanding of 'high frequency or everyday job-related vocabulary'. The terminology is couched in a form to give a broad characterisation but may be hard to apply in practice. What are these familiar words and what is everyday vocabulary?

The CEFR document also includes details of the vocabulary range and vocabulary control which are expected of learners at each level of the hierarchy. The vocabulary range criteria are presented in Table 2. This is likewise a series of general characterisations, for example, how broad should a lexical repertoire be before it is broad enough to fit the C level descriptors? Would a few thousand words be sufficient or is the learner expected to know the several tens of thousands which native speakers are reputed to have (D'Anna, Zechmeister, &

Table 1. A1 and B1 level descriptors from Council of Europe (2001, pp. 26–27)

LEVEL	LISTENING	READING	WRITING
A1	I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can write a short, simple postcard, for example, sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form
B1	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc.	I can understand texts that consist of mainly high frequency or everyday job-related language. I can understand the description of events, feelings and wishes in personal letters.	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.

Hall, 1991; Goulden, Nation, & Read, 1990)? Again, at what point does a learner's vocabulary knowledge pass from being sufficient for self-expression, at B1 level, to being good at B2 level? A further question arises as to how learners are to demonstrate this knowledge when the tasks presented to them, written essays or oral interviews for example, only allow them to produce a few hundred words, and most of these will be highly frequent and common to most learners (Milton, 2009). Daller and Phelan (2007) demonstrate that raters can be quite inconsistent in applying these kinds of criteria. While judgements of vocabulary range appear to be one of the more reliably applied sets of criteria in this data, it appears that raters can be misled by non-vocabulary factors such as accent in making their judgements (Li, 2008).

The value of the CEFR lies in the ability of its users to apply these criteria consistently and accurately but in the absence of more detailed criteria this may be difficult to do in practice. This difficulty is implicitly recognised in the CEFR document with the suggestion that vocabulary size details might usefully be added to the descriptors. The potential value of a form of assessment which is able to put some numbers, or more precise measurements, to these characterisations is very clear. If a learner possesses many thousand words, including idiomatic and colloquial expressions, and is comparable to a native speaker in his or her foreign language vocabulary knowledge then this would be good evidence that he or she would be at C2 level, at least in terms of vocabulary range. A learner with only a few hundred foreign language words would probably be at A1 level in terms of vocabulary range and almost inevitably would be much more limited in their skill in using the foreign language. It is exactly the kind of development which the writers of the CEFR foresee and

Table 2. Vocabulary range criteria from Council of Europe (2001, p. 112)

	VOCABULARY RANGE
C2	*Has a very 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 or her field and most general topics. Can vary formulation to avoid 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/her 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.*

which SLATE is embracing in its diagnostic tool kit. A set of figures for the size of vocabulary learners possess, and a straightforward method for determining these, would appear to be a very useful addition to the more general descriptive criteria on vocabulary range in particular.

3. Vocabulary knowledge and language skill

While the idea that the bigger and better your vocabulary in a foreign language is, the better you will be in your foreign language seems obvious, it is worth asking what research evidence we have to demonstrate that this is in fact the case. There is now a quite extensive body of research evidence which supports this idea and even provides some information as to the scale of vocabulary needed for different levels of performance and even which words are required to attain the highest levels in the CEFR framework.

The principle underlying these studies is an instrumentalist view of vocabulary that words are the primary carriers of meaning (Vermeer, 2001) and that as a consequence vocabulary provides the 'enabling knowledge' (Laufer & Nation, 1999) to be successful in other areas of language communication and proficiency. These studies show repeatedly that estimates of vocabulary knowledge correlate with reading comprehension (for example, Beglar & Hunt, 1999; Laufer, 1992; Qian, 1999; Stæhr, 2008), with writing ability (for example, Astika, 1993; Laufer, 1998; Stæhr, 2008), with listening comprehension (Milton, Wade, & Hopkins, 2010; Stæhr, 2008; Zimmerman, 2004), and with oral fluency (Milton et al., 2010; Zimmerman, 2004). The correlations are usually quite good and are typically between 0.6 and 0.8. Large vocabularies, there-

fore, typically associate with good performance in the communicative skills, and low vocabularies associate with poor performance. Perhaps not surprisingly therefore, the research also shows that tests of vocabulary knowledge can discriminate between groups of learners at different ability levels (Meara, 1992) and can be a useful tool for assigning learners to the correct level in an institutional program (Laufer, 1992; Laufer & Nation, 1999; Schmitt, 1994). The research also shows the presence of thresholds in vocabulary knowledge; volumes of vocabulary knowledge which appear essential if certain levels of comprehension, communication or performance in a foreign language are to be attained. Vocabulary thresholds have been suggested for reading comprehension (Alderson, 1984) and for reading and writing ability (Stæhr, 2008).

Recent research in this area goes beyond searching for correlations between vocabulary knowledge and the scores on individual languages skills and seeks to use regression analysis to calculate the scale of the contribution which vocabulary knowledge makes to performance in these skills. These studies establish that vocabulary knowledge, and vocabulary size in particular, is a major contributor to communicative performance in a foreign language. Stæhr (2008), for example, examines the relationship between examination grades on listening, reading and writing papers, and the vocabulary size of the testees, using scores on Nation's (1990, 2001) Vocabulary Levels Test as an indicator of vocabulary knowledge. His results suggest a link between vocabulary knowledge and all three elements of exam performance and a strong link with reading in particular. The correlations are given in Table 3.

Table 3. Spearman correlations between vocabulary size scores and listening, reading and writing scores (N=88) (Stæhr, 2008, p. 144)

	Listening	Reading	Writing
Vocabulary size	.69**	.83**	.73**

** Correlation is significant at the 0.01 level

Stæhr (2008) goes on to divide his exam results into two groups; below average, and average and above average. He carries out a binary logistic regression analysis using this division and concludes that as much as 72% of variance in the ability to score an average mark or above on the reading test can be explained by vocabulary size – the number of words a learner knows. Vocabulary may be less important than this in writing and listening but the contribution of vocabulary knowledge still appears sizeable. Staehr's results suggest that up to 52% of variance in the ability to score average or above in writing, and 39% of variance in listening, can be explained through vocabulary knowledge.

Milton et al. (2010) investigate the contribution of two types of vocabulary knowledge, orthographic vocabulary size and phonological vocabulary size, to the scores and sub-scores on the International English Language Testing System (IELTS) test (see http://www.ielts.org/default.aspx). The IELTS test provides sub-scores for each of the four skills: reading, writing, listening and speaking. The orthographic vocabulary size test used is X_Lex (or XLex), the Swansea Levels Test (Meara & Milton, 2003). The phonological vocabulary size test used is Aural Lex, known as ALex (Milton & Hopkins, 2005), which is an aural version of XLex where the tests words are heard but not seen. Details of this test are given later in this chapter.

The Spearman correlations which emerge from these scores are provided in Table 4 and reveal an interesting pattern to the way vocabulary knowledge interacts with scores on the IELTS skill sub-scores.

Table 4. Spearman correlations between vocabulary size scores and IELTS scores (N=30) (Milton et al., 2010, p. 91).

	ALex	reading	listening	writing	speaking	overall
XLex	.46*	.70**	.48**	0.76**	.35	.68**
ALex		.22	.67**	.44*	.71**	.55**

** Correlation is significant at the 0.01 level
* Correlation is significant at the 0.05 level

The modest correlation between ALex and XLex suggests that two different aspects of vocabulary knowledge are being tested and that they are not strongly connected. Orthographic vocabulary (XLex) scores correlate well with reading and writing skills, while phonological vocabulary (ALex) scores correlate well with speaking scores. Both vocabulary scores correlate with listening scores perhaps because the test for this skill involves both reading and listening. It is the orthographic vocabulary (XLex) scores which correlate particularly well with overall IELTS scores and which therefore appear to link best with overall language performance. Linear regression analysis suggests that nearly 60% of variance in the IELTS writing scores, 48% of variance in reading scores, and 58% of variance in overall IELTS scores can be explained by differences in orthographic vocabulary size. 51% of variance in listening scores can be explained by a combination of orthographic and phonological vocabulary scores. Using a binary logistic regression, where learners are divided into groups scoring IELTS 5 or better or below 5, 61%, variance in speaking scores can be explained through differences in phonological vocabulary (ALex) scores.

But research also suggests which words might be particularly relevant in these frameworks. Stæhr (2008) suggests that knowledge of the most frequent 2000 words in English in particular represents a threshold which must be crossed if learners are to gain an average score or above on the language tests he uses, and from this he suggests that the most frequent 2000 words are essential for learners to progress to intermediate level and beyond, presumably the B and C levels in the CEFR, and this is supported by Nation (2001, p. 16). Nation's (2006) study of coverage in English and comprehension further suggests that knowledge of the most frequent 5000 words, and overall vocabulary knowledge of perhaps 8000 or 9000 words, is essential for the highest levels of fluency and understanding in English as a foreign language. Where vocabulary knowledge measures tie so closely to performance and skill in the foreign language, it might be expected that vocabulary knowledge will link to levels within the CEFR.

These are interesting results and very relevant to this chapter since they suggest a very strong association between a learner's vocabulary size, and in particular the number of words a learner recognises in written form, and the communicative level and performance that the learner attains. This lends weight to the idea that particular vocabulary sizes might be associated with the grades of the CEFR, and confirms the attention paid in the CEFR document to vocabulary range, and in particular vocabulary size, as measured by the tests used in Stæhr (2008) and Milton et al. (2010) in particular.

Even from a brief review of this kind two general truths emerge. One is, as the CEFR hierarchy suggests, that progress through the hierarchy is closely related to vocabulary knowledge and knowing more and more words in the foreign language. High level performers tend to have extensive vocabulary knowledge and elementary level performers do not. The second is that knowledge of the most frequent words in the foreign language appears crucial to successful performance.

4. Vocabulary knowledge

Attention is paid in the CEFR description to several aspects of vocabulary knowledge and the terms vocabulary range, vocabulary control and vocabulary size are all used. How do these terms fit into the terminology for vocabulary knowledge which is more commonly used by researchers in this area and are there tests for these qualities? A common characterisation in vocabulary studies is to consider vocabulary knowledge as a number of contrasting dimensions. On one dimension is vocabulary size, also called lexical breadth, which is 'the number of words a learner knows regardless of how well he or she knows them' (Daller, Milton, & Treffers-Daller, 2007, p. 7). A calculation of vocabulary size

hinges, therefore, on the number of words a learner can recognise as words in the foreign language, and it may not matter for this calculation whether a meaning or a translation can be attached to the word and whether the word can be used with any great subtlety. This is the type of knowledge which is widely used by researchers when searching for the connection between vocabulary knowledge and language skills such as reading and writing, as do Stæhr (2008), Nation (2006) and Milton et al. (2010) in the previous section, and is explicitly mentioned by the CEFR description as a potentially useful calculation (Council of Europe, 2001, p. 150). Much of the Vocabulary range criterion, with its characterisations of basic vocabulary and broad lexical repertoire appears to be a function of this size or breadth dimension. I would argue too that Vocabulary control, with its emphasis on the learner's ability to select the right word for the meaning intended, is also largely a function of vocabulary size.

Vocabulary size contrasts with other dimensions of vocabulary knowledge. It contrasts with the knowledge a learner may have about how these words may work, their nuances of meaning and subtleties of combination, which is known as vocabulary depth. Knowledge of vocabulary depth is often calculated by estimating the degree to which learners can appropriately combine words as in collocations (Gyllstad, 2007), or can be selected for their appropriateness in given situations as in the use of idioms and colloquialisms, but the concept may also include partial word knowledge and knowledge of polysemy (Read, 2000). The Vocabulary range criterion (Council of Europe, 2001, p. 112) includes elements of vocabulary depth in addition to vocabulary size or breadth by including, at C level, reference to idiomatic expressions, colloquialisms and connotative meaning. The vocabulary control criterion (Council of Europe, 2001, p. 112) also appears to include elements of vocabulary depth in its reference, again at C level, to the 'appropriate use of vocabulary'. There need be no ambiguity in the CEFR's characterisations here since vocabulary size test scores correlate well with vocabulary depth scores. The two qualities are very closely inter-related and a test of one dimension inevitably tests the other. Vermeer (2001) argues that depth is a function of vocabulary size and that effectively they are the same dimension.

Vocabulary size contrasts too with the ease and speed with which these words can be called to mind and used in communication, which is usually characterised as productive vocabulary knowledge or lexical fluency. If these three dimensions of vocabulary knowledge seem, superficially, easy to understand and potentially useful in characterising learner language, they have proved rather harder in practice to operationalise. Of the three dimensions only vocabulary size has a generally accepted definition which can give rise to standard tests capable of measuring knowledge in this field. Fortunately, a vocabulary size test would seem to capture most of what the CEFR terms Vocabulary range and Vocabulary control. However, it has taken some time for standard tests to

emerge, because measuring a foreign language learner's word knowledge requires decisions to be made about what should be counted as a word. Early counts of the number of words a person knows gave some very large figures; sometimes in the hundreds of thousands for native speakers (for example Seashore & Eckerson, 1940). Word learning on this scale appeared to set a super-human task for foreign language learners who aspired to native-like levels of competence and performance. But figures such as these are a product, at least in part, of the definition of what a word is. These counts tended to be made on the basis of dictionary counts where polysemous words or homonyms might have several entries. A word such as *bank* might, therefore, include several entries as a noun: the *bank* of a river or the *bank* which dispenses money. It might also include several entries as a verb: *to bank* as in the turn of an aircraft, to put money into a bank, or to rely on something. But it might also include separate and additional entries for the various derived and inflected forms of these words. *Bank* and its plural *banks* might, conceivably, have separate entries. So too might other related forms such as *banker* or *bankable*. There is a tendency in these, older, counts for every different form of a word to be counted as a different word, and for every different meaning of the same form to be counted differently.

A word count made in this way may not usefully characterise the nature or scale of learning that foreign language learners undertake. Inflected and derived forms of words, in English these include regular plurals made by adding –s and regular past tenses made by adding –ed for example, often appear to be rule based. Once a learner knows a rule, it can be applied to a huge number of other words in English without the need for separate learning of a new form. It may be more useful, therefore, to characterise a word as a base form, in the example above *bank*, and a variety of related forms: *banks*, *banking* and *banked* for example. A base form and its related forms are known as a word family and we have evidence that words are stored and used in this way by native speakers (Aitchison, 1987, Chapter 11, for example summarises this evidence). And we have evidence that words are learned in this way by foreign language learners. Schmitt and Meara's (1997) research among Japanese L2 learners of English suggests that inflectional suffixes in particular are learned comparatively early and that a base form and rule-based variants are a feature of the developing lexicon, even if inaccuracies in use persist and knowledge of derived forms are added much later in the course of learning. If the definition of a word family is broadly drawn to include a base form and all derivations and inflections then it is estimated that an educated native speaker might know some 17,000 to 20,000 words (D'Anna et al., 1991; Goulden et al., 1990). While this is still a huge volume of learning for a non-native speaker to aspire to, it is substantially more approachable in scale than the hundreds of thousands of words suggest-

ed by earlier counts. There are many reputable and useful word counts and word lists which have been compiled using a definition of word as a large word family (for example, Coxhead, 2000).

This is not the only way to define a word family, however, nor the most appropriate for the purpose of building vocabulary size counts into the levels of CEFR. As Gardner (2007, pp. 260–261) points out these constructs need not be absolutely rigid in their definitions but might usefully be varied, for example to match the levels of knowledge of the learners being examined. A feature of the large word family is that it includes many comparatively infrequent derivations in its definition of a word, and these derivations are rarely known by non-native speakers at least until they achieve advanced levels of performance (Schmitt & Meara, 1997). A word count based on this definition seems likely to mischaracterise the knowledge of lower level learners, therefore. A rather more useful definition of a word family, which addresses this issue, is the lemma. A lemma includes a base form of a word and only the most frequent and regular inflections; related forms of a word which do not differ in part of speech from the base form. In English the lemma would include regular plurals and genitives in nouns, regular inflections -s, -ed, -ing and –en past participle forms in verbs, and comparative and superlative –er and –est endings in adjectives. Research evidence suggests these tend to be acquired early in the process of learning and this definition matches the kind of knowledge which elementary and intermediate level learners have. It is not surprising, therefore, that vocabulary size tests often draw on word frequency information where the words being counted are lemmatised words, for example, Nation's (1990, 2001) Vocabulary Levels Test and Meara and Milton's (2003) XLex. Nation's test has been described by Meara (1996) as the nearest thing we have to a standard test in the area, but the vocabulary size tests Meara has subsequently developed, such as XLex, appear to offer additional qualities to the tester. It seems likely, then, that these tests might be most useful for the basis of the size estimates that the creators of the CEFR feel would add useful detail to the levels descriptors.

5. Measuring vocabulary size

The previous sections have reported the use of vocabulary size tests which provide useful characterisations of the foreign language vocabulary size of learners. What are these measures and how do they work? Nation's (1990, 2001) Vocabulary Levels Test is one widely used test. It tests words in the 2000, 3000, 5000 and 10000 word frequency ranges, in addition to words from the University Word List (Nation, 1990) in order to estimate overall lexical competence. Each level tests 30 test words in a multi-item matching task where testees

are provided with six test words and a selection of three explanations which must be matched up with the test words. An example is given in Figure 1.

Figure: 1. Vocabulary Levels Test example taken from Nation (2001, p. 416)

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

1 business	
2 clock	_____ part of a house
3 horse	_____ animal with four legs
4 pencil	_____ something used for writing
5 shoe	
6 wall	

This format allows rather more than passive recognition of word forms to be tested and should allow an estimate of knowledge of words and their meanings to be formed. It is quite a complex test, however, where success relies not just on a learner's knowledge of the test words (on the left hand side) but also on knowledge of the words in the explanations (on the right hand side) and it is not completely clear which items are being tested. Further, each question contains multiple test items and the learner's knowledge of some of the items is likely to have an impact on the ability to work out the answers to other items where these are not known. We know that learners often try to maximise their scores by making educated guesses in these circumstances, it is called economy of practice, but we do not know the effects of guesswork on the scores that emerge. Kamimoto (2005), for example, reports think aloud protocols conducted among learners taking this test and the feedback he received suggests that a considerable amount of guesswork and calculation goes on in answering the questions and that the learner's choice of guessing strategy can produce considerable differences in score. The Levels Test might have much more variation according to guesswork than most users ever imagine. However, there is no explicit way in the test for taking account of this phenomenon or compensating for it. It is not completely clear how scores on this test might equate to vocabulary size although it would be a surprise if scores on this test and vocabulary size did not correlate well. It is also heavily weighted in its content to the infrequent ranges of vocabulary. It will be recalled that Stæhr (2008) draws attention to the importance of the most frequent 2000 words for learners but the most frequent 1000 words, the frequency band which includes so much structure and function vocabulary, is absent from consideration in this test. The majority of test items lie at or beyond the 5000 word level and learners will need to be highly

accomplished in English before they have significant knowledge in this area. A test in this format is probably unsuitable as a measure to tie into the levels of the CEFR since it seems unlikely that it will be able to distinguish well between learners at the elementary levels of knowledge and performance.

The better alternative is the XLex check list test format (Meara & Milton, 2003) which corresponds to tests of passive receptive vocabulary within a specified vocabulary range. The XLex tests estimate knowledge of the most frequent 5000 lemmatised words in a language and estimate overall knowledge of this vocabulary. They are Yes/No tests which present learners with test words, one by one, and learners have to indicate whether they know each word. There is a combination of real words and false words which are designed to look and sound like words in the target language. The number of Yes responses to the real words allows a preliminary estimate of the proportion of words known from the most frequent 5000 words, and the number of Yes responses to the false words allows this preliminary estimate to be adjusted for guessing and overestimation of knowledge. The tests give an overall score of words known out of the most frequent 5000 words. These tests can draw on frequency lists in different languages, where these exist, to allow comparable tests of vocabulary to be constructed. In Meara and Milton's (2003) XLex, for example, the English tests are based on data available from Nation (1984) and Hindmarsh (1980), Greek tests draw on the Hellenic National Corpus (http://hnc.ilsp.gr/en/ ; see also Milton & Alexiou, 2010), and the French tests draw on Baudot's (1992) frequency lists. The 5000 words content includes the most frequent 2000 words, the importance of which elementary and intermediate learners Stæhr (2008) has pointed to, and the most frequent 5000 words which Nation (2006) has pointed to as so crucial for advanced level comprehension and communicability. Studies on these tests in English format (Milton, 2005), in French format (David, 2008), and in Greek format (Milton & Alexiou, 2010) conclude that they are both reliable and valid measures of receptive vocabulary size. An illustration of the format of this type of test is given in Figure 2.

Figure 2. Example of a checklist test (French version from Milton, 2009, p. 257)

Please look at these words. Some of these words are real French words and some are invented but are made to look like real words. Please tick the words that you know or can use. Here is an example.

☒ chien

Thank you for your help.

☐ de ☐ distance ☐ abattre ☐ absurde ☐ achevé ☐ manchir

This format has certain features which make it attractive for calculating the vocabulary size that learners have in relation to CEFR levels. The test is simple in format and comparatively large numbers of items can be tested in a short space of time. Results on these tests are usually very reliable. The test format also has the benefit of allowing the creation of parallel or even equivalent forms of the test in different languages to be made rather more straightforwardly than would be the case with multiple choice or Levels Test formats. Examples of the tests can be found in Milton (2009, pp. 255–261).

6. Scores on vocabulary size measures and CEFR levels

The research literature contains several examples where vocabulary size estimates have been linked explicitly to the levels of the CEFR. Milton and Meara (2003) tested students taking and passing Cambridge exams at every level of the CEFR and estimated their vocabulary sizes using the XLex tests. These exams test all the 4 skills and vocabulary size is therefore linked to a learners' level in the widest possible sense. Their results are shown in Table 5.

Table 5. Approximate vocabulary size scores associated with CEFR levels (adapted from Meara & Milton, 2003, p. 8)

CEFR Levels	Cambridge exams	XLex (5000 max)
A1	Starters, Movers and Flyers	<1500
A2	Kernel English Test	1500 - 2500
B1	Preliminary English Test	2750 - 3250
B2	First Certificate in English	3250 - 3750
C1	Cambridge Advanced English	3750 - 4500
C2	Cambridge Proficiency in English	4500 - 5000

Milton and Alexiou (2009) used three different language versions of XLex and collected data from over 500 learners of English, Greek and French as second and foreign languages with a view to comparing the levels of vocabulary knowledge at each CEFR level. As broad a range of languages and learners as possible was sought and the authors drew on contacts from Vocabulary Research Group in order to collect the data from a variety of locations. The EFL learners in Hungary were drawn from 144 students in two schools in Szeged and the 88 EFL learners in Greece from a private language school in central Greece. The learners of French in UK were 155 students at a comprehensive school in South

Wales, and the learners of French in Spain were all 50 students of French at a Secondary school in northern Spain. The Greek learners of French were all 65 students of French at a private language school in central Greece. The learners of Greek as a second language comprised all 64 learners at a university centre in Thessaloniki. The CEFR levels were determined in this study by teachers who placed the learners being tested into streams for study at each of the CEFR levels. The mean scores and standard deviations collected from learners at each level are summarised in Table 6. The standard deviation scores are included to indicate the degree of overlap between the groups at each level.

Table 6. Summary of mean scores (and standard deviations) for each CEFR level in three foreign languages

CEFR Level	EFL in Hungary	EFL in Greece	French in Spain	French in Greece	Greek in Greece
A1		1477 (580)	894 (604)	1125 (620)	1492 (705)
A2		2156 (664)	1700 (841)	1756 (398)	2237 (538)
B1	3135 (434)	3263 (434)	2194 (717)	2422 (517)	3338 (701)
B2	3668 (666)	3304 (666)	2450*	2630 (251)	4012 (415)
C1	4340 (471)	3690 (471)	2675 (643)	3212 (473)	
C2		4068 (261)	3721 (416)	3525 (883)	

* in this field there is only one student score and no SD can be calculated

These figures are broad generalisations but one striking feature emerges. Progressively higher vocabulary scores are associated with progressively higher levels in the CEFR hierarchy. Milton and Alexiou (2009, pp. 200–204) use ANOVA and Tukey analyses to demonstrate that the differences between the mean scores at each CEFR level are statistically significant, and linear regression modelling to show that, with the exception of the Hungarian data, some 60 to 70% of variance in CEFR levels can be explained by differences in vocabulary size. It is apparent from this data that the assumption made in the CEFR literature, that as learners progress through the CEFR levels their foreign language lexicons will increase in size and complexity, is broadly true. There is individual variation and overlap between the scores that learners attain within the CEFR levels, however. It will be recalled that vocabulary size and communicative performance are separable qualities and one interpretation of this variation is that students with the same or similar vocabulary sizes may make different use of this knowledge to communicate more or less successfully.

A further notable feature of these estimates is just how much vocabulary is needed to progress beyond the most basic levels of performance. Something like 2000 words, of the most frequent 5000, might be needed to attain A2 level in all the languages tests, and this recalls Stæhr's (2008) comments, although about EFL in particular, about the importance of the most frequent 2000 words and the threshold that knowledge of these words represents in attaining above average marks in his writing and reading tests. It appears that approximately 3000 words, of the most frequent 5000 words, might be needed to progress beyond the elementary A1 and A2 levels and achieve any kind of independence in English and Greek as foreign languages. The advanced levels of performance on the CEFR, C1 and C2 level, are associated as Nation (2006) suggested with almost complete recognition of the most frequent 5000 words. Scores of 4000 or better are associated with these levels and these suggest that learners' overall vocabularies must be very large indeed, and this fits well with Nation's (2006) figures, derived from coverage calculations, that 8000 or 9000 words overall might be needed for full comprehension of written texts.

However, the figures associated with each level in the three different languages are different. The figures for French as a foreign language are consistent with each other and suggest that at each CEFR level the learners have smaller vocabulary sizes than would be the case for English. The data available for Greek as a foreign language suggests that achieving each CEFR level requires a rather larger vocabulary size than would be the case for either English or French. Milton and Alexiou (2009) are at pains to point out that the data available come from comparatively small samples and that caution should be exercised in reaching conclusions as a consequence. Nonetheless, there are reasons why differences of this kind might be expected when comparing the vocabulary sizes of learners across different languages. This raises the question of how vocabulary sizes can be meaningfully compared in this way.

7. Making cross-linguistic comparisons

One of the virtues of a checklist test format, where words are chosen from across frequency bands, is that it is possible to construct tests in different languages which are arguably comparable with each other. The tests used in the previous section can all claim to estimate the number of words the learners know out of the most frequent 5000 lemmatised words in those languages. It is not inevitable, however, that knowing the same number of words in several languages will mean you can perform identically in those languages. Languages differ in important ways and one effect of this may be that it is possible to do more with fewer words in one language than in another.

There are several reasons for thinking this which are discussed in Milton (2009). One is that languages can inflect and derive words very differently and this may affect frequency calculations. The most frequent words in English, for example, include pronouns and prepositions but not every language shares this quality. Agglutinative languages, like Hungarian, Finnish and Turkish, handle many of the functions and meanings which these words convey in English rather differently. Typically these meanings are conveyed by the addition of suffixes to the root form of a verb or noun with the result that a single word family might include many more word forms than would be the case in English. This should affect the volumes of vocabulary required for coverage and comprehension. It may be possible to do more with, say, 1000 words in Finnish than in English. A second reason might lie in the historical development of languages. Speakers of English, for example, often appear to have a variety of words available for much the same idea in a way that is quite different from other languages. For historical reasons English differentiates, for example, between many farmyard animals and the meat which comes from them, between *pork* and *pig* and between *sheep* and *mutton*, in a way that other languages do not. This implies that comprehension of English might require a larger vocabulary than would be needed for equivalent understanding in another language. Further, languages can differ considerably in their word formation processes. Some languages, like German, favour compounding and combine existing words to make a new word. By contrast, other languages favour creating or deriving words for new concepts and ideas. The effect of this word combining process may be compounded since not all languages are as clear as Western European languages in signalling where word boundaries occur. Chinese, for example, which combines ideographs to make words, does not mark word boundaries in writing so it may not immediately be clear where one word ends and the next begins, or whether an expression should be treated as two or three words or as one. Different decisions would, when systematised across a whole corpus, produce different word counts.

A further difference may result from cultural differences in how languages choose to express themselves across different registers. We have some empirical evidence from English and French that these differences can affect coverage. It appears that in French the most frequent vocabulary does the service of everyday language and, in addition, can also be used in formal and academic registers where, in English, a specialist academic vocabulary is required. Cobb and Horst (2004, p. 30) point to the coverage provided of academic texts by the most frequent 2000 words in French. The figure they quote of nearly 89% would be equivalent to the General Service Word List of 2000 words plus Coxhead's Academic Word List, a further 570 carefully selected rather than purely frequency-based words in English. It appears that it really may be possible to do more communicatively with fewer words in French than is the case in English, as the figures in Table 6 suggest.

This observation leads Milton and Alexiou (2009) to suggest that the figures for vocabulary size in the CEFR levels may be tied to coverage, and that this will provide a way of both explaining and predicting how vocabulary size will vary across the CEFR levels in different languages. They note (2009, pp. 207–208) that the figures for coverage in Greek suggest that larger vocabularies will be needed to achieve the CEFR's various communicative requirements at each level, and that the data from Table 6 confirms that this is the case. Milton and Alexiou (2008) also suggest, however, that this is not the only possible interpretation of this data and that these differences might equally well be the result of differences in the ways the various corpora used to derive information on coverage were created. Generally, language corpora aspire to draw on a wide variety of general source material such as newspapers, novels and periodicals so as to be as representative of the normal use of the language as possible. But there are no conventions to indicate how large a sample from these genres is appropriate for the corpus to be representative of the language as a whole. A corpus compiled from a large number of small samples drawn from a wide variety of these sources may well contain a greater diversity of words than corpora drawn from a smaller number of texts where the material is likely to be thematically much more unified. This may explain why the Greek corpus (http://hnc.ilsp.gr/en/), at least in the incarnation used for the Greek XLex tests, appeared to contain a much larger proportion of singly occurring words than the other two corpora. The English language corpora used to create the EFL XLex tests (Hindmarsh, 1980; Nation, 1984) include material drawn from EFL teaching texts, a feature which is absent from both the French corpus (Baudot, 1992) and Greek corpus used in this study. This may have influenced the number of words that learners recognise, since the English XLex tests potentially contain words more closely related to the material they have encountered in class although it is not clear that this will inevitably be the case. None of these corpora, at the time they were used, contained a spoken sub-corpus and while this is unlikely to have affected the words which emerge as most frequent, it is likely to affect the comparative frequencies of these words and, potentially, the frequency bands they occur in therefore. Language corpora, constructed in a much more strictly equivalent manner, are needed before doubts about the comparability of corpora, and the tests derived from them, can be allayed.

8. Conclusions

The evidence from studies of vocabulary size has confirmed that vocabulary size measurements can be tied to the levels of the CEFR with some confidence, although it seems at present as though the actual vocabulary sizes may depend

on the language being tested and perhaps the source of the words which form the test. The higher up the hierarchy of the CEFR learners progress, the more words they are likely to require and the greater vocabulary size they will have. The relationship between vocabulary size and the CEFR levels is sufficiently strong, notwithstanding some individual variation, for figures for vocabulary size to be attached to the CEFR level. This relationship is not just a matter of academic interest and one of the most useful benefits in linking vocabulary knowledge and the CEFR in this way is to add detail to the framework in the manner that the CEFR itself anticipates. Users of the system often find it difficult to match learners or materials to the levels with any precision and different people, different examiners, even different national examination systems, can apply the CEFR's levels descriptors very differently. The interest SLATE has in creating a diagnostic tool kit which links linguistic features of performance to the CEFR levels looks to be important, therefore, and the use of vocabulary size measurements, and the tests to derive such measurements, appear to add detail to the CEFR rather as the creators of the framework anticipate. The presence of these kinds of more precise descriptors should help users of the system in different schools or countries apply grading criteria more consistently and confidently. Vocabulary size measurements have much to add to the CEFR hierarchy therefore. It seems that vocabulary size can be a very useful part of a linguistic tool kit and even by itself is a good predictor of CEFR level (Milton & Alexiou, 2009).

References

Aims of SLATE. (n.d.). Retrieved from http://www.slate.eu.org/aims.htm

Aitchison, J. (1987). *Words in the mind*. Oxford: Blackwell.

Alderson, J. C. (1984). Reading in a foreign language: A reading or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1–24). London: Longman.

Astika, G. G. (1993). Analytical assessment of foreign students' writing. *RELC Journal, 24*(1), 61–72.

Baudot, J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain*. Montréal: Les Presses de l'université de Montréal.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2,000 word level and university word level vocabulary tests. *Language Testing, 16,* 131–162.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In B. Laufer & P. Bogaards (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). Amsterdam: John Benjamins.

Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M., & Papo, E. (1987). *Un niveau seuil*. Paris: Didier.

Council of Europe. (2001). *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213–238.

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge.* 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.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior, 23,* 109–122.

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal, 36,* 167–180.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28,* 241–265.

Goulden, R., Nation, I. S. P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11,* 341–363.

Gyllstad, H. (2007). *Testing English collocations – developing receptive tests for use with advanced Swedish learners* (Doctoral dissertation). Lund University, Sweden.

Hindmarsh, R. (1980). *Cambridge English lexicon.* Cambridge: Cambridge University Press.

Kamimoto, T. (2005, September). *The effect of guessing on vocabulary test scores: A qualitative analysis.* Paper presented at The European Second Language Association 2005 conference (EUROSLA 15), Dubrovnik, Croatia.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London: Macmillan.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19,* 255–271.

Laufer, B., & Nation, P. (1999). A productive-size test of controlled productive ability. *Language Testing, 16,* 33–51.

Li, H. (2008). *Measuring and assessing English L2 spoken vocabulary* (Unpublished doctoral dissertation). Swansea University, UK.

Meara, P. (1992). *EFL vocabulary tests.* University College Swansea, Centre for Applied Language Studies.

Meara, P. (1996). The dimension of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P., & Milton, J. (2003). *X_Lex, the Swansea Levels Test.* Newbury: Express.

Milton, J. (2005). Vocabulary size testing in the Arabic-speaking world. In P. Davison, C. Coombe, & W. Jones (Eds.), *Assessment in the Arab world* (pp. 337–353). Dubai: TESOL Arabia.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol: Multilingual Matters.

Milton, J., & Alexiou, T. (2008). Vocabulary size in Greek as a foreign language and the Common European Framework of Reference for Languages. *Journal of Applied Linguistics, 24,* 35–52.

Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards et al. (Eds.), *Vocabulary studies in first and second language acquisition* (pp. 194–211). Basingstoke: Palgrave.

Milton, J., & Alexiou, T. (2010). Developing a vocabulary size test for Greek as a foreign language. In A. Psaltou-Joycey & M. Mattheoudakis (Eds.), *Selection of papers for the 14th International Conference of Applied Linguistics (GALA)* (pp. 307–318). Aristotle University of Thessaloniki, GALA.

Milton, J., & Hopkins, N. (2005). *Aural Lex.* Swansea University.

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, & M. D. López-Jiménez (Eds.), *Further insights into non-native vocabulary teaching and learning* (pp. 83–97). Bristol: Multilingual Matters.

Nation, I. S. P. (Ed.). (1984). *Vocabulary lists: Words, affixes and stems.* English University of Wellington, English Language Institute.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Boston: 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? *The Canadian Modern Language Review, 63,* 59–82.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review, 56,* 283–307.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Schmitt, N. (1994). Vocabulary testing: Questions for test development with six examples of tests of vocabulary size and depth. *Thai TESOL Bulletin, 6,* 9–16.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework – word associations and verbal suffixes. *Studies in Second Language Acquisition, 19,* 17–36.

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabulary. *Journal of Educational Psychology, 31,* 14–38.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36,* 139–152.

Van Ek, J. A. (1980). *Waystage English.* London: Pergamon Press.

Van Ek, J. A. & Trim, J. L. M. (1991). *Threshold 1990.* Strasbourg: Council of Europe.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22,* 217–234.

Zimmerman, K. J. (2004). *The role of vocabulary size in assessing second language proficiency* (Unpublished master's thesis). Brigham Young University, USA.

Linking L2 proficiency to L2 acquisition: Opportunities and challenges of profiling research[1]

Jan H. Hulstijn
University of Amsterdam

At the SLATE meeting of December 2006, Franceschina, Alanen, Huhta and Martin (2006) proposed the so called *DEMfad* agenda, meaning that development of a particular linguistic construction can best be studied by examining the *f*requency, *a*ccuracy and *d*istribution of form-meaning constructions longitudinally, from *e*mergence to *m*astery in a given *d*omain[2]. However, we would not only like to examine the acquisition of specific elements of L2 grammar in isolation, but also in combination with an understanding of how the lexico-grammatical system at large is developing. Researchers therefore need to ascertain the "linguistic profile" of L2-learners, at any point of L2 development, in terms of the accuracy, under-use and over-use of morpheme sequences. Many of the papers brought together in this volume provide evidence of the feasibility of such a profiling enterprise. This is an exciting, promising and non-trivial feat, first of all for language testing practice and possibly also for syllabus design. As the contributions of Salamoura and Saville and of Alanen, Huhta and Tarnanen show, it is possible to build learner corpora of written productions, have productions rated at a given CEFR level, annotate them, and then conduct lexical and grammatical analyses yielding lexico-grammatical profiles. Comparisons of the profiles appear to show that the profiles of two adjacent developmental levels often differ from each other in the frequency of occurrence of a whole range of (correct or incorrect) morphemes rather than in the total absence versus presence of particular morphemes or in the partial versus fully correct application of particular rules of grammar.

---

1 I would like to thank the three editors and Gabriele Pallotti for their useful comments on an earlier version of this text.

2 The six underlined letters make up the DEMfad label; see the contribution of Martin, Mustonen, Reiman, and Seilonen in this volume for a full explanation).

Thus the good news is that, although second language acquisition is a matter of gradual progress (not a matter of jumping from one stage to the next), it appears to be possible to assign, with high probability, an L2 writing product to a particular CEFR level. With sufficiently big learner corpora and with increasingly sophisticated software tools, we can now envisage, for the not too distant future, that computers will reliably rate L2 productions. Computer software is presently already capable of automatically rating constrained responses, such as single-utterance responses; it does so by sheer "ignorant brute force", without parsing responses into traditional linguistic categories. The profiles resulting from the sort of work described in this volume will also allow computer rating of longer or non-constrained productions. Much laborious work has to be conducted, however, until this type of computer rating has materialized because it is unlikely that universal profiles will be found for each CEFR level for each target language and the software has to take learners' L1 into account. Finally, as Salamoura and Saville point out, profiling research may help construct detailed scales of the lexical and grammatical mastery required for a given target language. This would produce scales of much more detail than the global, language-neutral scales for lexis and grammar in chapter 5 of the Common European Framework of Reference, CEFR. (Council of Europe, 2001). Once reliably established, profiles can be used in L2 instruction (in addition to the Profile Deutsch, English, Spanish, etc projects of the Council of Europe). A research team at Lund University has already developed a program that is capable of producing grammatical profiles in a corpus of texts written by L2 learners of French (Granfeldt, Nugues, Ågren, Thulin, Persson, & Schlyter, 2006).

The danger of circularity in establishing CEFR-related profiles

It is clear then that language assessment stands to benefit from profiling research. But will SLA research benefit as well? Before answering this question, let me briefly bring to mind what the CEFR is and what it isn't. The CEFR attempts to describe communicative language proficiency in terms of real-world like language activities (chapter 4) and language competencies (chapter 5). It has no ambition of linking language proficiency to language acquisition in any detail, beyond the general observation (Council of Europe, 2001: 17) that acquisition of an L2 is a matter of development in what learners can do with their L2 and how well they can do this. The CEFR contains well over 50 scales, all using the same six level symbols (A1, A2, B1, B2, C1, and C2). Some scales do and others do not refer to linguistic accuracy. For instance, the B1 level of the global scale (p. 24) is defined with minimal reference to linguistic accuracy: "Can understand the main points of clear standard input on familiar matters

regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans." Only the word "simple" may be interpreted as referring to linguistic quality. However, the qualitative scale on pages 28–29 specifies accuracy at the B1 level in the following way: "Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations."

For profiling research to be successful in the ways to be outlined below it is crucial to avoid the danger of circularity, as several authors in this volume emphasize. If written or oral productions are first rated with the quality scale of chapter 3 (Table 3, pp. 28–29), containing references to linguistic accuracy, and the researcher is then going to analyse and compare productions at different CEFR levels in terms of accuracy, overuse, and underuse of certain grammatical phenomena, the study may fall victim to circularity. Thus, to avoid circularity in profiling research, when written or oral productions are being rated, rating scales should be used with no, or minimal reference to accuracy of linguistic forms.

The potential of profiling research for understanding second language acquisition

What might profiling research mean for our understanding of second language acquisition? I can think of two benefits, which I present below. First, however, I should say that I doubt whether profiling research will breathe new life into theories and research on so called developmental sequences and natural orders in SLA (see Ellis, 2008, Ch. 3, for a recent overview of the literature on developmental sequences). There are several obstacles for such research. First, most research on developmental sequences pertains to the initial stages of L2 acquisition, i.e., resulting in lower levels of language proficiency. To be able to see what happens during the initial stages of SLA, two snapshots at levels A1 and A2 of the CEFR are unlikely to yield the required detail. Furthermore, for a proper study of development, cross-sectional data need to be supplemented by longitudinal data and I wonder whether funding can be obtained for longitudinal studies involving sufficiently large numbers of L2 learners to produce learner corpora large enough for reliable profiling analyses. Finally, most research on developmental sequences or natural acquisition orders is based on data collected from L2 learners not receiving L2 instruction, whereas most of the data collected in CEFR-linked profiling research have been collected from "instructed"

L2 learners. Frankly speaking, I should perhaps add that I belong to the few skeptics with respect to theories and research on developmental sequences. I do not see which non-trivial developmental phenomena of L2 acquisition there presently are to explain, beyond the A1 level in instructed L2 learners. In 2001, Goldschneider and DeKeyser published an influential re-analysis of oral production data from twelve studies conducted between 1973 and 1996, together involving 924 subjects. Multiple regression analysis showed that 71% of the total variance in acquisition order of six functional morphemes of English is explained by the combination of five determinants: perceptual salience, semantic complexity, morphonological regularity, syntactic category, and frequency of these morphemes in the input. When we add L1 transfer as an explanatory factor for acquisition order of some L2 structures, I wonder what else is there to explain?

I see two benefits of profiling research for theories of SLA. First, analyses of large learner corpora consisting of L2 productions rated at different CEFR levels may throw new light on the study of L1 transfer in SLA, because it allows looking at multi-word strings (integrating lexical and grammatical phenomena) in large corpora collected from L2 learners with a large variety of linguistic backgrounds. It would, for instance, be interesting to examine whether profiles of learners with different first languages differ mainly in terms of underlying morphosyntax, as a result of L1 transfer, while hardly differing in terms of morpho-phonological surface forms, as implied by the conservation hypothesis of Van de Craats, Van Hout, and Corver (2002; see also Van de Craats, 2009). If there is one thing that more than 40 years of SLA literature has shown, then it is evidence of massive L1 transfer in SLA and L1 transfer is therefore likely to affect the profiles too, especially at the lower CEFR proficiency levels. Furthermore, by analysing strings of any length and composition, profiling research may provide an integrated lexico-grammatical picture of SLA. In most of the SLA literature, spanning some 40 years, the study of grammatical development is separated from the study of lexical development and the study of phonological development is separated from the study of morpho-syntactic development. Integrative, cross-domain analyses are likely to widen the scope of SLA theories.

Second, a promising line of research emerges when we focus on explaining the levels of L2 proficiency attained in terms of (a) input and (b) learner attributes such as age and level of education. Profiling research ought to look at native speakers as well. It would be good to administer the writing and speaking tasks, presented in the papers of this volume, also to L1 speakers of different ages and at different levels of education or profession. Since Chomsky (1965), many linguists have lost interest in investigating variability in oral and written L1 profi-

ciency, assuming that all adult native speakers share a common core grammar. One may wonder, however, to what extent adult native speakers do share a common lexicon and grammar (Hulstijn, 2007, 2010). In a recent study, Mulder and Hulstijn (2010) observed large variability in the accuracy and speed with which adult native speakers of Dutch, differing in age and the level of their education or profession, performed a variety of language tests. In speaking, a considerable number of participants produced utterances violating some very basic rules of Dutch grammar, such as subject-verb agreement. Thus, when L2 profiling research of the types represented in this volume aims to establish accuracy, underuse, and over-use of lexical and grammatical features in L2 learners, researchers need to have at their disposal corpora of oral and written language produced by L1 speakers of different ages and at different levels of education, for comparison. As the authors of the CEFR pointed out, the CEFR levels do not constitute steps towards native-speaker proficiency (Council of Europe, 2001: 36). The C2 level is attained by only a minority of native speakers. In other words, the CEFR levels implicitly incorporate learners' intellectual functioning into the proficiency scales, reflecting the vertical socio-economic structure in European societies. It would be fascinating to examine at which CEFR level the language proficiencies of adult native speakers begin to differ. Is A2 the highest level shared by adult native speakers or do they share as much as B1? Our view of L2 proficiency at different CEFR levels might undergo some fundamental changes when we take differences in native speakers' language proficiency into account.

References

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass: MIT Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Ellis, R. (2008). *The study of second language acquisition* (2nd edition). Oxford, England: Oxford University Press.

Franceschina, F., Alanen, R., Huhta, A., & Martin, M. (2006, December). *A progress report on the Cefling project.* Paper presented at SLATE Workshop, Amsterdam.

Goldschneider, J. M., & DeKeyser, R. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50.

Granfeldt, J., Nugues, P., Ågren, M., Thulin, J., Persson, E., & Schlyter, S. (2006). "CEFLE and Direkt Profil: A new computer learner corpus in French L2 and a system for grammatical profiling". In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (pp. 565-570), Genoa, Italy. Retrieved from http://www.cs.lth.se/home/Pierre_Nugues/Articles/lrec2006/lrec2006_dp.pdf

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

Hulstijn, J. H. (2010). *Explanations of associations between L1 and L2 literacy skills.* Manuscript submitted for publication.

Mulder, K., & Hulstijn, J. H. (2010). *Linguistic skills of adult native speakers, as a function of age and level of education.* Manuscript submitted for publication.

Van de Craats, I. (2009). The role of 'is' in the acquisition of finiteness by adult Turkish learners of Dutch. *Studies in Second Language Acquisition*, *31*, 59–92.

Van de Craats, I., Van Hout, R., & Corver, N. (2002). The acquisition of possessive have-clauses by Turkish and Moroccan learners of Dutch. *Bilingualism: Language and Cognition*, *5*, 147–174.

Language testing-informed SLA? SLA-informed language testing?

J. Charles Alderson
Department of Linguistics and English Language, Lancaster University

The aim of SLATE is to bring together language testing researchers and researchers of second language acquisition to create synergies to our mutual benefit. This volume bears testimony to the interesting and varied research that SLATE has inspired, and there is much to be learned from this, both by SLA researchers and by language testers. In this invited evaluative chapter, I will attempt to draw some of the lessons and to answer the questions of my title.

The first, and to me impressive, lesson is that second language acquisition is about much more than English. The various chapters here report on studies into the development of aspects of proficiency not just in English, but also in Finnish, French, Dutch, Italian, Norwegian and Spanish. That clearly reflects the (West) European nature of SLATE and is to be applauded. However, there is a notable lack of studies of other European languages, particularly the Central and East European Slavic and Baltic languages as well as Greek, Hungarian, Romanian and others. Even German, Portuguese and Swedish are missing. Let us hope that as SLATE's work extends and becomes better known, this situation will change.

The second lesson is that, so far, SLATE has not paid much attention to the relationship between the learners' first language and their target second or foreign language, and it is to be hoped that, in future, attention can be paid to this important matter. Currently, most informants in SLA studies seem to be from so many different L1s that it is impossible to draw conclusions about cross-linguistic transfer or influence. Future studies will need to be specifically designed to yield such information, rather than to hope that it might emerge from opportunistic samples.

In relation to this issue of L1, it is interesting to note that one or two of the studies reported looked at the performance and proficiency of native speakers of their target language, but it was not always clear whether like was being compared with like, i.e. informants of similar ages, educational background and so

EUROSLA MONOGRAPHS SERIES 1
Communicative proficiency and linguistic development, 239-248

on. The very notion of a native speaker has, of course, been questioned and problematised (see Davies, 2003, for example), but there are obvious benefits to seeing how (similar) native speakers perform on the measures used by SLA researchers, even though we are conscious of the comparative fallacy (Bley-Vroman, 1983).

The third lesson is that second language acquisition and language testing need to conduct research with a range of different informants, not just the ubiquitous university student. Obviously such captive populations are convenient and attractive for reasons of practicality, but it is important that younger learners, learners outside formal education, migrants in second language settings, and learners who are not simply taking a test, be studied. It is to the credit of authors of this volume that SLATE members have already begun the study of younger learners and those who are not simply conveniently available because they have taken a public examination.

The fourth point, albeit not perhaps a lesson, is the importance of the Common European Framework of Reference, CEFR, (Council of Europe, 2001) in virtually all of the chapters. This is not surprising, given both SLATE's explicit aims to examine the relationship between the communicative approach of the CEFR and the linguistic development of learners, and the growing importance of the CEFR in Europe and beyond. Not everybody agrees that this is a desirable state of affairs (Fulcher, 2004; McNamara, personal communication), and there are concerns that the CEFR as it currently stands is probably not suitable for, and not intended for, younger learners. Nor is it suitable on its own for the development of language tests, or even of textbooks and curricula (Alderson et al., 2006; Byrnes, 2007b; Hulstijn, 2007; Little, 2007; Weir, 2005; Westhoff; 2007, and others in the MLJ Perspectives edited by Byrnes, 2007a). However, there can be no doubt that the existence of the CEFR has given an important impetus to language education, to language testing and examining in particular, and to research into the development of language competence. That the CEFR needs to be adapted to particular contexts not only should go without saying, but is explicitly stated in the CEFR itself, particularly in the boxed texts that frequently begin "Readers might like to consider to what extent...". It is regrettable, but it was predictable, that claims are made about the CEFR level of curricula, textbooks, tests, and more, that have no empirical basis, but which are often produced for marketing or political purposes. But that should not detract from the importance of the CEFR; indeed it emphasises the importance of research that investigates and challenges the claims of both the language education profession (or industry) and of the CEFR itself. Such research, as is beginning to be attested in SLATE, can only enhance our understanding of language proficiency and its development.

It is, however, important that such research be properly conducted, based upon knowledge of best practice in, and theories of, second language acquisition and language testing. Too much research has based itself on unsatisfactory measures of "proficiency" like years of study in school, first, second and subsequent years of study at university, the Vocabulary Size Placement Test of DIALANG, or a cloze or C-test. However, several authors in this volume have been careful to avoid the circularity of using CEFR linguistic scales alone.

> When rating with reference to the CEFR levels, our aim was to rate learners' performances on the basis of their ability to do things with the language. Paying too much attention to linguistic features could introduce circularity in the reasoning underlying a study such as Cefling: proficiency levels are determined on the basis of linguistic features, and these features are, in their turn, used in defining the levels. (Alanen, Huhta, & Tarnanen, this volume)

Crucially, the fifth and substantive lesson is the importance of paying attention to the construct to be investigated, and of widening the range of constructs. Inevitably, for reasons of practicality, there is a tendency to investigate development through studying learners' written productions. There is no need to transcribe speech, and, increasingly, data can be available in digital form if the informants have word-processed their writing (as in computer-based testing, for example). There is less emphasis in the research reported in this volume on examining learners' oral performances, and no research looking at how learners' reading and listening abilities develop. This is hardly surprising, given the difficulty of studying what are essentially internal processes, but it is nevertheless to be hoped that future research will pay more attention to these so-called receptive skills.

SLA research has in the past tended to pay much more attention to morphosyntax than to other aspects of language, so the chapters on the development of vocabulary (Milton, this volume) and cohesive devices like discourse connectives (Carlsen, this volume) are especially welcome. It is to be hoped that other areas of language use might be studied in the future, like the development of the pragmatic features of politeness, for example, and sociolinguistic and cross-cultural competences. Although some studies still use convenient but crude indices like errors per T unit, or the number of subordinate clauses per clause, it was refreshing to see much more attention than in the past being paid to variables of more convincing construct validity.

But perhaps the most important lesson of all for me, as a language testing researcher, was the importance of research into SLA paying much more attention to its methodology, and the validity and reliability of the instruments and proce-

dures used. In this regard the chapter by Alanen et al. (this volume) was exemplary in its account of the design of their study. More and more SLA studies use electronic corpora as the data for their investigations, either specially created by the researchers, or pre-existing learner corpora, such as the International Corpus of Learner English (Granger, 1998), The Longman Learner Corpus (2008) or the Cambridge Learner Corpus (2010). The Cefling project, reported on in two chapters in this volume, is an interesting case of both. The researchers made use of a corpus of examination scripts from the National Certificates of Finland, but also created their own corpus of young learners' writing on specially designed tasks.

Pre-existing corpora, although convenient, may have their drawbacks. The International Corpus of Learner English was highly innovative in its time and gave rise to numerous interesting studies, but it has two rather serious limitations. First, there is no indication of the learners' proficiency level, be that according to the CEFR or any other measure. Rather, they are classified according to their year of university study. Unfortunately, that is not a valid measure of their level of development. Secondly, the tasks on which the data were based were "persuasive or argumentative essays", with no standardised rubrics, on a wide range of topics, and so the comparability of essays from different sources must be in some doubt. In addition, different genres are ignored, as are learners at different ages. Even the Cambridge Learner Corpus (CLC), the subject of one chapter in this volume, has problems, as shown by Kim (2009). Data in the CLC were collected over a long period of time, based upon the Cambridge Main Suite of ESOL Examinations. However, not only do the tasks in these examinations change from one administration period to the next, but all the examinations have undergone more or less substantial changes over time. This makes it difficult to be sure that the tasks are parallel. In addition, no grade is given for the writing tasks, but only for the overall grade achieved on the whole examination. Moreover, the examinations have not been formally linked through a standard-setting process to the CEFR, and therefore it is difficult to make statements about progression through the CEFR levels, rather than through the various Cambridge examinations. Similar problems exist with the Longman Learner Corpus.

Researchers who create corpora specially for SLA and testing research, such as the Young Learner part of Cefling, need to pay careful attention to their design, as Alanen et al. (this volume) describe. Learners should be asked to perform a variety of tasks, which have been specially designed for their capacity to elicit relevant performances, preferably related to the target language use situation. These need to be thoroughly scrutinised, piloted on suitable numbers of target learners, analysed and revised in light of the results. They should then be administered in conditions which are appropriate to the task, not necessarily, however, under examination conditions. The written or oral performances should be rated on

relevant scales, avoiding the danger of circularity, and the raters should be familiar with the scales, trained in their use, and benchmark performances should be available for the guidance of the raters. Such scripts should always be rated by at least two raters, and only those scripts should be incorporated into the corpus on which raters agree, within clearly specified margins of disagreement. Ideally, there would also be available an independent measure of CEFR-related proficiency.

Most corpora, and indeed SLA studies, are cross-sectional, but there is a strong case for longitudinal studies. Such studies need to allow sufficient time for relevant development to occur, and thus are probably better undertaken with learners in the range A1 to B1, initially at least, than with more advanced learners. Inevitably there are difficulties gathering data from a sufficient number of informants for the results to be of more general interest, but the existence of a group like SLATE may facilitate larger-scale studies than are possible in doctoral research. Much SLA research, even cross-sectional research, has used rather small datasets, which limits the value of such studies.

Hulstijn, Alderson, and Schoonen (this volume) contextualise and present the initial research questions of interest to SLATE. An important question is: to what extent are these research questions addressed in this volume?

The over-arching research question was:

Which linguistic features of learner performance (for a given target language) are typical at each of the six CEFR levels?

Clearly this has been or is being addressed in virtually all the chapters presented in this volume. However, this research is in its early stages. No definitive answers have yet been provided, and much more work remains to be done.

The more specific research questions and goals of research were:

1. What are the linguistic profiles at every CEFR level for the two productive language skills (speaking and writing) and what are the linguistic features typical of the two receptive skills (listening and reading) at every CEFR level?

The chapters in this volume begin to address the productive skill of writing, particularly at the A2/B1 divide. However, speaking has yet to be explored in detail across the CEFR levels, and the two receptive skills have as yet received no attention, as discussed above.

2. *To what extent do common or different profile features exist across the seven target languages investigated by researchers in the SLATE group (Dutch, English,*

Finnish, French, German, Italian and Swedish)? Do the profiles differ along language-family lines (Finnish versus the two Romance languages represented in SLATE (French and Italian), and the three Germanic languages Dutch, English and German,)? To what extent do the profiles reflect learners' L1?

Although research relevant to these questions has begun, the surface has barely been scratched. Cross-linguistic comparisons are only addressed in one chapter (Kuiken, Vedder, & Gilabert, this volume) but are not related to CEFR levels, language families have not yet been profiled or compared, and the study of learners' L1s with respect to the second/foreign language profiles has yet to begin.

3. *What are the limits of learners' performance on tasks at each of the CEFR levels?*

Research into what learners canNOT do is barely represented in this volume, but remains an important area for distinguishing performances on different tasks at different CEFR levels.

4. *Which linguistic features, emerging from our profiling research, can serve as successful tools in the diagnosis of learners' proficiency levels and of weaknesses that require additional attention and training?*

The diagnosis of learners' strengths and weaknesses at the different CEFR levels has to be an important aim of SLATE research, and several chapters mention the diagnostic potential of their research. However, it has to be admitted that not much progress has been made in this area, especially with respect to the CEFR levels rather than to individual examinations or examination suites. The area of diagnostic testing is, however, very under-developed (but see Alderson, 2005, 2007; Alderson & Huhta, 2005). It is the firm belief of this author that SLATE has much to contribute in the future, if its efforts are appropriately focussed.

5. *Are there commonalities and differences between the linguistic profiles of foreign-language learners (learning the target language in the formal setting of a school curriculum or a language course) and those of second-language learners (learning the target language without formal instruction)?*

Given that linguistic profiles have yet to be achieved through empirical research for any language, this is clearly an ambitious, albeit interesting, question that is not addressed in this volume. Which is not to say that it will not be possible to address it in the future, as part of a long-term agenda.

And so to return to the questions of my title: *Language testing-informed SLA? SLA-informed language testing?* The assumption behind the questions is that both are important and equally relevant aims for SLATE, as is implicit in the acronym. However, in order to start a constructive and fruitful dialogue between the two disciplines, it is important to recognize some fundamental differences between SLA and testing.

As many chapters in this volume testify, SLA research is mainly concerned with variability of linguistic performance - be it variability related to the tasks, to the L1 or to a host of individual factors. While this variability is certainly a concern for language testers as well, it is clearly the case that testing, by its very nature, or, at least, proficiency testing, is mainly concerned to measure what is stable in language ability. There is little point in measuring something, especially if the results of that measurement will affect lives, if it is highly likely that that ability will change in the next ten minutes or three months, or if the very fact of measurement will distort the results. We need to have as stable as possible a picture of what we are measuring. Testers have to look for stability - one aspect of which is reliability - and we need 'stability of judgement' as much as we need 'stability of performance' (or ability, if you must). Testers also seek generalisability (whether you see this as a facet of reliability or validity is immaterial to the present argument). Thus, we are not interested in knowing whether the learner can understand this particular story about Prince Charles in today's edition of the Daily Mirror as compared with a rather different story about the same person in the Daily Telegraph, as compared with their ability to understand an article about the Dutch Royal Family. We typically want to make more general statements about reading ability than that. Thus testers - proficiency testers at least - are much less concerned with the particular than with the general underlying ability (I would argue if I had the space that diagnostic testers might well be more interested in the particular than the general, but that is another paper – see Alderson, 2005).

SLA researchers, on the other hand, seem to be more interested in examining the particular - the use in English of the third person s, the use of the present perfect, of negation, or particular speech acts, and so on. This is in part no doubt because they are driven by linguistic theories and the predictions of such theories - hence the proliferation of studies on the development of grammatical morphemes - or by pedagogical applications of linguistic insights - hence the search for the effect of negative evidence. SLA research is more interested, I would argue, in some of the bits and pieces, the minutiae of linguistic competence or performance than they are in the much wider picture of the ability to communicate meaning in context. Language proficiency testers these days (not diagnostic testers, who scarcely exist as a breed yet) are almost obliged to exam-

ine how somebody would perform in a range of communicative settings, across a variety of tasks, texts, activities and all the other dimensions of communicative behaviour that are discussed, however briefly, in the CEFR. SLA researchers are not, *per se*, interested in what is in the CEFR, because there is no specific mention of specific languages or specific linguistic knowledge.

SLA researchers take 'the worm's eye view', immersed in the detail of the blades of grass around them, whilst language testers take not merely the bird's eye view (which might include trying to find worms to eat), but arguably the space satellite's view in order to discover the underlying features that determine the shape of the landscape. The different eye views might be incompatible - that remains to be seen - but they are certainly different, and I would argue that they serve different purposes. However, although much neglected and confusingly defined (see Alderson, 2005), diagnostic testing is one obvious possible meeting place between SLA and language testing. Diagnosis needs theories of language development, as provided, at least potentially, by SLA and SLA needs the insights provided by language testing's awareness of the variability in performance across test task facets.

How far has this volume advanced the informing of SLA through language testing principles and practice, and to what extent has testing been informed by SLA theory and results?

My impression is that the main focus of the volume has been on SLA questions and issues, and has not (yet) been relevant to language testing. Which is not to say that this will not be the case in future. Indeed, I would assert that, to the extent that SLA research heeds the principles and practice of language testing, it will eventually yield results that will be hugely relevant to language testing. But these results are much more likely to be relevant to the diagnostic testing of learners' strengths and weaknesses than they are to the communicative testing of language proficiency in target language use situations.

Language testing, at least in high-stakes contexts, has tended to concentrate on the testing of an individual's ability to communicate accurately and appropriately in relevant settings, because it has been concerned with assessing learners' proficiency, and not with diagnosing their levels, strengths and weaknesses. Language testing may thus at first sight not appear to offer much to SLA's interest in the development, specifically, of linguistic competences. However, I believe this would be a mistaken conclusion. SLA has much to learn about the development of valid, reliable and comparable language use tasks which can offer insights into the linguistic features of learners' performance and development. Indeed, I believe that this volume is evidence of the importance of SLA researchers paying attention to language testing's concerns. Thus, it would

appear to be of more relevance to SLA researchers than to language testers. Whilst some progress could be said to have been made towards language-testing informed SLA, the relevance to SLA-informed language testing has yet to emerge.

Lest this appear too pessimistic a conclusion, let me finish by claiming that, to the extent that SLA is informed by language testing principles and practice, SLA will indeed contribute important research insights that will eventually be relevant to and will inform language testing theory and practice, especially, but not exclusively, in the area of diagnostic testing. In addition, I would hope that, given a common interest in the CEFR, both as a statement about what it means to be able to learn and use any foreign language, and as a possible framework for understanding - or thinking about - what language development might mean, we might be able to find common ground. I look forward to that future.

References

Alderson, J. C. (2005). Diagnosing *foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C. (2007). The challenge of diagnostic testing: Do we know what we are measuring? In J. Fox et al. (Eds.), *Language testing reconsidered* (pp. 21–39). Ottawa: University of Ottawa Press.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, *22*(3), 301–320.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of The Dutch CEFR Construct project. *Language Assessment Quarterly*, *3*(1), 3–30.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17.

Byrnes, H. (Ed.). (2007a). Perspectives [Special section]. *The Modern Language Journal*, *91*(4), 641–685.

Byrnes, H. (2007b). Developing national language policies: Reflections on the CEFR. *The Modern Language Journal*, *91*(4), 679–685.

Cambridge Learner Corpus [A collection of exam scripts]. (2010). Retrieved March 25, 2010, from http://www.cambridge.org/fi/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/?site_locale=fi_FI

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Davies, A. (2003). *The native speaker: Myth or reality?* Clevedon: Multilingal Matters.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *1*, 253–266.

Granger, S. (Ed.). (1998). *Learner English on computer*. London and New York: Addison Wesley Longman.

Hulstijn, J. J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language. *The Modern Language Journal*, *91*(4), 663–667.

Kim, J. (2009). *Development patterns of Korean learners corresponding to morphosyntactic items* (Unpublished Doctoral dissertation). UK: Lancaster University.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, *91*(4), 645–655.

The Longman Learner's Corpus. (2008). Retrieved March 25, 2010, from http://www.longmanusahome.com/dictionaries/corpus.php#aa

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal*, *91*(4), 676–679.

About the authors

Riikka Alanen is a professor of applied linguistics at the Centre for Applied Language Studies of the University of Jyväskylä, Finland. Her research interests include L2 learning and teaching as task-mediated activity and the role of consciousness and agency in the learning process. She has written articles in Finnish and English about children's L2 learning process as well as their beliefs about English as a Foreign Language. She is also the co-editor of the volume *Language in Action: Vygotsky and Leontievian legacy today* (2007).

J. Charles Alderson is professor of Linguistics and English Language Education at Lancaster University. He is a specialist in language testing and the author of numerous books, book chapters and academic articles on the topic, as well as on reading in a second or foreign language. He is a former co-editor of the journal *Language Testing*, co-editor of the *Cambridge Language Assessment Series* and a recipient of the Lifetime Achievement Award of the International Language Testing Association (ILTA). See also http://www.ling.lancs.ac.uk/profiles/J-Charles-Alderson/

Inge Bartning is a professor of French at Stockholm University. She has taught and published in the domain of French syntax, semantics and pragmatics. In the last two decades her main interest has been in French L2 acquisition, in particular the domain of developmental stages, advanced learners and ultimate attainment of morpho-syntax, discourse and information structure. She is currently participating in a joint project '*High Level Proficiency in Second Language Use*' with three other departments at Stockholm university (see www.biling.su.se/~AAA). She has (co-)authored articles in *IRAL*, *Journal of French Language Studies*, *EUROSLA Yearbook*, *AILE*, *Langue française*, *Lexique*, *Revue Romane* etc.

Cecilie Carlsen holds a Master's degree in SLA-research from the University of Bergen (UiB). In 2003 she completed a PhD in language testing and assessment. From 2003–05 she was involved in the development of national tests of English for Norwegian school children, and since 2005 she has worked with the development and validation of Norwegian test for adult immigrants at Norsk språktest, UiB. Since 2008 she has been a postdoctoral research fellow at the UiB. As part of her project she has carried out a linking of a learner corpus to the CEFR, and she is currently using the corpus in an investigation of cross linguistic influence of rhetorical structures.

Fanny Forsberg is a Research fellow and lecturer of French at Stockholm University, where she received her PhD in 2006 with a thesis on formulaic language in L2 French. She is currently involved in research on ultimate attainment in L2 French and Spanish, with particular regard to the role of formulaic language and how it relates to activity types, L2 pragmatics and the CEFR-scale. Furthermore, her research interests include more generally spoken language, conversation analysis and L2 acquisition.

Roger Gilabert is a lecturer and researcher at the University of Barcelona, Spain, where he is a member of the Language Acquisition Research Group (GRAL). He has worked in the area of second language acquisition with a focus on L2 production. He has published a number of articles in the area of task complexity and L2 performance.

Ari Huhta is a researcher at the Centre for Applied Language Studies, University of Jyväskylä. He specialises in language assessment, and has participated in several national and international research and development projects, such as DIALANG (1996–2004) and DIALUKI (2010–). His interests include self-assessment and feedback, diagnostic assessment, assessment of speaking, development of assessment procedures for SLA research, and the study of the linguistic characteristics of CEFR levels.

Jan Hulstijn is a professor of second language acquisition at the University of Amsterdam, Faculty of Humanities, Amsterdam Center for Language and Communication (ACLC). Most of his research is concerned with cognitive aspects of the acquisition and use of a nonnative language (explicit and implicit learning; controlled and automatic processes; components of second-language proficiency). He held previous positions at the Free University Amsterdam and at the University of Leiden. He was associate researcher at the University of Toronto, Canada (1982–1983) and visiting professor at the University of Leuven, Belgium, (2002) and at Stockholm University, Sweden (2005). His webpage provides information concerning his research projects and publications (http://home.medewerker.uva.nl/j.h.hulstijn/).

Folkert Kuiken is professor of Dutch as a Second Language at the University of Amsterdam. He is a senior research member of the Amsterdam Center for Language and Communication (ACLC) and is coordinator of the Dual Master of Dutch as a Second Language and of the AILA Research Network on Task Complexity and Second Language Learning (TaCoSeLL). His research interests include the effect of task complexity and interaction on SLA, Focus on Form, and the relationship between linguistic complexity and communicative adequacy.

Maisa Martin is professor of Finnish as a second and foreign language at the Department of Languages, University of Jyväskylä. She is the director of the project Linguistic Basis of the Common European Framework for L2 English and L2 Finnish, Cefling for short. Her current research interests focus on the development of

various linguistic features in L2 Finnish. She is also the co-director of the program Challenges of Language Acquisition within the LANGNET, the Finnish doctoral program in language studies, as well as the chair of Council for Finnish Studies at Universities abroad.

James Milton is Senior Academic in Applied Linguistics at Swansea University. His research work focuses on vocabulary acquisition and measurement. Recent works include Measuring Second Language Vocabulary Acquisition (2009) published by Multilingual Matters and Modelling and Assessing Vocabulary Acquisition (2007) published by CUP.

Sanna Mustonen is a doctoral student in the University of Jyväskylä, at the Department of Languages. Her PhD thesis is part of the Cefling project: It addresses the question of how Finnish as a second language develops across the proficiency levels of CEFR, focusing on the spatial and abstract uses of locative cases. She has also been working on practical fields: language skill assessment in The National Certificates system, teacher training, and developing new teaching materials.

Gabriele Pallotti is associate professor of Language teaching methodology at the University of Modena and Reggio Emilia. He is a member of the Executive Committee of the European Second Language Association (Eurosla), and of the SLATE network. His research focusses on Italian as a second language, SLA methodology, and cross-cultural and intercultural discourse. He has coordinated several nation-wide projects with the Italian Ministry of Education and has lectured in many countries. His works have appeared in edited volumes and in *Applied Linguistics, Journal of Intercultural Communication,* and *Journal of Pragmatics.* He is the editor, with J. Wagner, of *L2 learning as social practice* (in press).

Nina Reiman is a doctoral student in Finnish as a second language at the University of Jyväskylä. Her research focuses on the development of transitive constructions.

Angeliki Salamoura holds a PhD in English and Applied Linguistics from the University of Cambridge (UK) and has a number of publications in this field. She has a strong background in quantitative research methods and has worked as a postdoctoral researcher on English language processing at the department of Experimental Psychology in Cambridge. She is currently a Research and Validation Officer at Cambridge ESOL and her role involves working on the English Profile Programme and the development and validation of bespoke assessment products. Her research interests include second language learning and processing, bilingualism, methodological approaches to linking examinations to the CEFR and quantitative research methodology.

Nick Saville is Director of the Research and Validation Group at Cambridge ESOL. He has specialised in language testing and assessment since 1987and holds a PhD from the University of Bedfordshire in the area of language test impact. He is the representative of Cambridge ESOL in ALTE - the Association of Language Testers in Europe - and has close involvement with other European initiatives, such as the Council of Europe's Common European Framework of Reference (CEFR) and related "toolkit". Currently he is a member of the Cambridge University team responsible for co-ordinating the English Profile Programme - i.e. Reference Level Descriptions for English to accompany the CEFR. Nick has published widely on issues related to language assessment, and is an associate editor of Language Assessment Quarterly.

Rob Schoonen is an associate professor of SLA at the University of Amsterdam, where he teaches psycholinguistics, general linguistics and research methodology. His main research interests are (second) language acquisition and language proficiency, language testing and research methodology. He has been Associate Editor of *Language Learning* and a board member of the International Language Testing Association (ILTA). He currently is a member of the SLATE executive board. He has (co-)authored articles in *Applied Linguistics, Applied Psycholinguistics, Journal of Educational Psychology, Language and Education, Language Learning, Language Testing* and *TESOL Quarterly*.

Marja Seilonen is a doctoral student at the University of Jyväskylä. The subject of her research is genericity in the written texts of learners of Finnish. She is a lecturer of Finnish as a second language at the University of Eastern Finland.

Mirja Tarnanen is a senior researcher at the Centre for Applied Language Studies of the University of Jyväskylä, Finland. Her research interests include learning and teaching of L2, and language assessment and literacy practices in language education. She has been involved in several national research and development projects of language assessment and language education. She has run national language testing system National Certificates for adults since 2003.

Ineke Vedder is a researcher at the University of Amsterdam (Center for Language and Communication, ACLC), and head of education (Department of Language and Literature). Her research interests include instructed second language acquisition, particularly Italian as a second language, task-based language learning and L2 writing. Her publications have appeared in various books and journals (e.g. *IRAL, Journal of Second Language Writing, International Journal of Educational Research, Studi Italiani di Linguistica Teorica e Applicata, Linguistica e Filologia*). She is co-editor of the present SLATE book and of two edited volumes, to appear in 2011. See for further information her webpage (http://home.medewerker.uva.nl/s.c.vedder/).