



Universität Paderborn

Fakultät für Kulturwissenschaften

Anglistik/Amerikanistik

Interfaces between Second Language Acquisition and the Common European Framework of Reference – Proposing a Scale for Grammatical Range

Dissertation zur Erlangung des akademischen Grades
eines Doktors der Philosophie (Dr. phil.) im Fach
Englische Sprachdidaktik

vorgelegt von
Katharina Hagenfeld

Gutachter/in
PD Dr. Anke Lenzing
Prof. Dr. Dr. h.c. Manfred Pienemann

Paderborn, im Oktober 2018

To Martin

List of Figures.....	i
List of Tables.....	iii
List of Abbreviations.....	v

1. INTRODUCTION 1

1.1 AIMS OF THE THESIS.....	1
1.2 RESEARCH QUESTIONS AND HYPOTHESES	4
1.3 THE STRUCTURE OF THE THESIS.....	5

2. THEORETICAL BACKGROUND..... 11

2.1 THE COMMON EUROPEAN FRAMEWORK OF REFERENCE	12
2.1.1 HISTORICAL OVERVIEW	12
2.1.2 THE STRUCTURE OF AND NOTIONS IN THE CEFR.....	25
2.1.3 DIMENSIONS OF LANGUAGE PROFICIENCY AND COMPETENCES	27
2.1.4 COMMUNICATIVE COMPETENCE – LANGUAGE PRODUCTION AND PROCESSES	36
2.1.5 LINGUISTIC COMPETENCES IN THE CEFR.....	38
2.1.6 GRAMMATICAL COMPETENCE IN THE CEFR	43
2.1.7 LANGUAGE LEARNING IN THE CEFR.....	47
2.1.7.1 The Learning and Teaching of Linguistic Competences.....	50
2.1.7.2 The Role of Learner Errors in the CEFR.....	50
2.1.7.3 Assessment in the CEFR.....	51
2.2 PROCESSABILITY THEORY.....	54
2.2.1 YARDSTICKS IN PROCESSABILITY THEORY.....	55
2.2.1.1 A Brief Outline of Levelt’s Blueprint for the Speaker	56
2.2.1.2 A Brief Sketch of Bresnan’s Lexical Functional Grammar.....	61
2.2.2 THE HIERARCHY OF PROCESSING PROCEDURES AND STRUCTURAL OPTIONS FOR ENGLISH.....	65
2.2.3 HYPOTHESIS SPACE	76
2.2.4 THE EMERGENCE CRITERION	78
2.2.5 HISTORICAL BACKGROUND TO PROCESSABILITY THEORY.....	81
2.2.6 TEACHABILITY, DEVELOPMENTAL READINESS AND LEARNER ERRORS	84
2.2.7 LINGUISTIC PROFILING AND RAPID PROFILE.....	88

3. BRIDGING SCALES AND STAGES..... 95

3.1 SELECTED STUDIES ON SLA AND THE CEFR	97
3.2 PRIOR STUDIES ON THE CEFR AND PT	109
3.3 GRAMMATICAL RANGE – AN INTEGRATIVE APPROACH TO THE CEFR AND PT	113
3.3.1 DIFFERENCES IN THE FRAMEWORKS: UNIVERSALITY, EMERGENCE, ACCURACY	115
3.3.2 COMPETENCE – THE CEFR AND PT	118
3.3.2.1 Competence in the CEFR	118
3.3.2.2 Competence and PT.....	125
3.3.2.3 Competence: Chances and Challenges in Combining PT and the CEFR	127
3.3.3 THE SHAPE OF THE EMERGING LINGUISTIC SYSTEM IN LEARNERS	128
3.3.3.1 Progression in the CEFR	129
3.3.3.2 Progression in PT	130
3.3.3.3 Progression and Processes: Chances and Challenges in Combining PT and the CEFR	131
3.4 APPLIED ISSUES IN THE CEFR AND PT – COMBINED ASSESSMENT	133
3.4.1 ASSESSMENT IN THE CEFR.....	133
3.4.2 ASSESSMENT IN PT	134
3.4.3 INTERFACES REGARDING ASSESSMENT IN PT AND THE CEFR	135

3.5 SUMMARY	137
<u>4. THE STUDY</u>	<u>137</u>
4.1 SOME WORDS ON THE RATIONALE.....	138
4.2 RESEARCH QUESTIONS AND HYPOTHESES	140
4.3 METHODOLOGY	143
4.3.1 OVERVIEW OF THE PROCEDURES AND ANALYSES	144
4.3.2 PILOT PHASE FOR THE PERCEPTION OF GRAMMATICAL INACCURACY	148
4.3.3 THE DATA.....	151
4.3.4 THE TASKS.....	153
4.3.5 THE AUDIO-FILES AND THE EDITING PROCEDURE	154
4.3.6 PILOTING THE EDITED SAMPLES.....	158
4.3.7 DISTRIBUTING THE FILES TO THE RATERS.....	159
4.3.8 THE RATER GROUPS.....	162
4.3.9 THE RATING SCHEMES AND THE NOVICE RATER TRAINING.....	170
4.4 RESULTS	178
4.4.1 THE EFFECT OF GRAMMATICAL ACCURACY – ORIGINAL AND EDITED SPEECH SAMPLES.....	179
4.4.2 QUALITATIVE COMPARISON OF ORIGINAL AND EDITED SAMPLES	182
4.4.2.1 Comparison of Original and Edited Files based on Mode.....	185
4.4.2.2 Comparison of Original and Edited Files based on Range	187
4.4.3 DISCUSSION OF RESULTS ON GRAMMATICAL ACCURACY.....	189
4.4.3.1 Discussion of Original and Edited Files across Rater Groups.....	190
4.4.3.2 Rater Cognition and Rater Experience.....	194
4.4.3.3 The Role of Uneven Profiles.....	196
4.4.3.4 The effect of the locus of presentation of files.....	199
4.4.3.5 The effect of the different types of manipulation of edited files	201
4.4.3.6 Summary of the Effect of Manipulation in Morphosyntax.....	203
4.4.4 RELATIONS BETWEEN CEFR LEVELS AND PT STAGES	204
4.4.4.1 Discussion of Results on Relations between CEFR and PT.....	207
4.4.4.2 A Combined Scale for Grammatical Range	210
4.4.4.3 Comparing Scales for Grammatical Range and General Linguistic Range	214
4.4.4.4 Grammatical Range for English.....	217
4.4.5 RATER EXPERIENCE AND VARIABILITY OF RATING RESULTS	219
4.4.5.1 Discussion of Results on Rater Experience and Variability of Rating Results	225
4.4.5.2 Rater Experience and Agreement.....	225
4.4.5.3 Positive and Negative Wordings of Descriptors	226
4.4.5.4 Different Rating Techniques	231
<u>5. CONCLUSION AND FUTURE DIRECTIONS</u>	<u>233</u>
<u>6. REFERENCES.....</u>	<u>237</u>
<u>7. APPENDICES.....</u>	<u>256</u>
7.1 EXEMPLARY TRANSCRIPTIONS	256
7.2 EXEMPLARY PROTOCOL PILOT STUDY	270

List of Figures

Figure 1: Overview of historical events shaping the CEFR, taken from http://www.coe.int/t/dg4/linguistic/historique_en.asp (last access 20.06.2018)	14
Figure 2: The three bands and six levels of the CEFR	28
Figure 3: Global Scale of the CEFR, taken from CoE (2001: 24).....	30
Figure 4: General competences in the CEFR, adapted and modified from	32
Figure 5: Communicative Language Competences in the CEFR, taken from Green (2012: 20) .	34
Figure 6: Overall Oral Production Grid, taken from CoE (2001: 58)	36
Figure 7: CEFR Scale for General Linguistic Range, taken from CoE (2001: 110)	39
Figure 8: Grammatical Accuracy Scale in CEFR, taken from (CoE 2001: 114)	44
Figure 9: A blueprint for the Speaker, taken from Levelt (1989: 9)	57
Figure 10: Subject-verb Agreement, example adapted and modified from Pienemann (1998)	60
Figure 11: Three parallel structures in LFG, taken from Pienemann, DiBiase & Kawaguchi (2005: 200).....	61
Figure 12: Feature unification in the s-procedure, taken from Pienemann, DiBiase & Kawaguchi (2005: 200).....	62
Figure 13: Levels of representation in LFG, taken from Lenzing (2013: 94).....	64
Figure 14: Incremental language generation, taken from Pienemann (1998: 68)	67
Figure 15: Locus of information exchange for morphology, taken from Pienemann (2008: 16).....	68
Figure 16: Hypothetical hierarchy of processing procedures, taken from Pienemann (2005a: 14)	69
Figure 17: Direct mapping of argument onto surface form, taken from Lenzing (2013: 216) ...	71
Figure 18: The Multiple Constraints Hypothesis, taken from Lenzing (2013: 8)	72
Figure 19: Hypothesis Space, taken from Pienemann (1998: 232)	77
Figure 20: Accuracy and order of acquisition, taken from Pienemann (1998: 137).....	80
Figure 21: Developmental Readiness, taken from Keßler (2006: 96).....	86
Figure 22: Diagnostic Task Cycle, taken from Keßler (2008: 301)	91
Figure 23: Rapid Profile 4.0 user interface.	93
Figure 24: Grammatical competence in the CEFR in Context.....	119
Figure 25: Bachman's View of Communicative Language Ability, taken from Bachman (1990:85)	122
Figure 26: Time management for familiarization activities, taken from CoE (2009: 23)	164
Figure 27: Self-Assessment Grid part one, taken from CoE (2001: 26f.).....	167

Figure 28: Self-Assessment Grid part two, taken from CoE (2001: 26f.)	168
Figure 29: Rating Criteria given by bA07.....	227
Figure 30: Rating Criteria Rater bA08	228
Figure 31: Rating Criteria by Amateur Rater aA04.....	229

List of Tables

Table 1: PT hierarchy for English as a L2, taken from Lenzing, Plesser, Hagenfeld & Pienemann (2013: 272), on the basis of Pienemann (2005a: 24).....	73
Table 2: New Verb Co-occurrence Frames at B2 level, taken from Salamoura & Saville (2010: 116).....	104
Table 3: Relationship between PT stage and CEFR level found by Lenzing & Plesser (2010) ..	110
Table 4: CEFR levels and PT stages by Michalska 2010, presented by Keßler & Plesser (2011:236), adapted and modified	111
Table 5: Overview of Statistical Test in Relation to Results.....	146
Table 6: Results Perception of Grammatical Inaccuracy in Pilot Phase	150
Table 7: Audio files divided into original and edited files	152
Table 8: Overview of the Task Set	153
Table 9: Example of Editing Procedure in the Transcription	155
Table 10: Sources of Learner Data.....	157
Table 11: Mode of Presentation of Audio Files to Raters.....	160
Table 12: Pattern of Sample Order for File Distribution.....	161
Table 13: Participants in Rater Groups.....	162
Table 14: Overview of Novice Rater Training	165
Table 15: Overview Expert Rater Affiliations.....	170
Table 16: Letters used by amateur raters and according CEFR level	171
Table 17: Global Oral Assessment Scale, taken from CoE (2009: 184).....	172
Table 18: Complementary Grid for Global Oral Assessment, taken from CoE (2009: 185)	176
Table 19: Results of Rating - Comparison of Original Files and Edited Files.....	180
Table 20: Results Ratings - Range and Mode for CEFR Ratings for PT stages.....	184
Table 21: Mode for Ko10 and Ke10	186
Table 22: Comparison of Mode of files K05 and K01 by the Amateur Rater Groups	186
Table 23: Original - Edit Comparison based on Range for Amateur Raters.....	188
Table 24: Original - Edit Comparison based on Range for Novice Raters.....	188
Table 25: Original - Edit Comparison based on Range for Expert Raters	189
Table 26: Edit-Original Rating in Amateur Rater Group	190
Table 27: Edit-Original Rating in Expert Rater Group	191
Table 28: Original-Edit Rating in Novice Rater Group	193
Table 29: Mode for Ko10 and Ke10	197

Table 30: Mode for Ko06 and Ke06	199
Table 31: Overview of Direction of Rating Tendencies for Original and Edited Files across Groups	200
Table 32: Comparison of edited versions generated by audio-engineer and through re-recordings	202
Table 33: Results Rating - General Overview of Relations between PT stages and CEFR levels	205
Table 34: Overview Previous Studies on the Relationship between PT and the CEFR	207
Table 35: Empirically Motivated Proposed Scale for Grammatical Range in Comparison to Grammatical Accuracy	211
Table 36: Comparison Descriptors for General Linguistic Range and Proposed Descriptors for Grammatical Range.....	215
Table 37: Overview of potentially combined descriptors to propose a scale for English, based on Grammatical Range (examples taken from Lenzing 2013: 144; based on Pienemann 2005: 24)	218
Table 38: Variability due to Rater Experience and Assessment Grid Use.....	220
Table 39: Results Agreement within Sub-groups.....	221
Table 40: Amount of assigned CEFR levels to audio-files across groups	224

List of Abbreviations

ALTE	Association of Language Testers in Europe
a-structure	Argument structure
CEFR	Common European Framework of Reference
COALA	Computer-assisted Language Analysis
CoE	Council of Europe
c-structure	Constituent Structure
EC	Emergence Criterion
ELP	European Language Portfolio
f-structure	Functional Structure
IL	Interlanguage
ILR	Interagency Language Roundtable
IPG	Incremental Procedural Grammar
L1	First language
L2	Second language
LARC	Language Acquisition Research Centre
LFG	Lexical Functional Grammar
NP	Noun Phrase
PT	Processability Theory
RP	Rapid Profile
S	Sentence
SLA	Second Language Acquisition
SLATE	Second Language Acquisition and Testing in Europe
S-node	Sentence node
S-procedure	Sentence procedure
SUBJ	Subject
S-V-agree	Subject-verb agreement
TH	Teachability Hypothesis
t-level	Threshold Level
MCH	Multiple Constraints Hypothesis
MMM	Multidimensional Model
V	Verb
VP	Verb Phrase
ZISA	Zweitspracherwerb Deutscher und Italienischer Arbeiterkinder

1. Introduction

Studies in the teaching and learning of foreign languages have gained more and more attention in the past 50 decades. Especially since the Council of Europe (CoE) started to promote multilingualism and to set the long-term goal for Europe that “[...] all EU citizens should speak two languages in addition to their mother tongue” (CoE 2006: 9), the study of foreign language learning and teaching has found its way into the teaching training curricula. The present study is situated in this tradition; more specifically, in the area of applied linguistics.

1.1 Aims of the Thesis

In this study, I examine two frameworks¹ that have gained importance in the context of foreign language pedagogy and foreign language acquisition. One framework that this study addresses is the Common European Framework of Reference (CEFR) that was published in 2001 by the CoE to provide guidance for language professionals in the form of a reference tool. The authors of the CEFR claim that the document puts forward an action-oriented, learner-friendly and undogmatic approach to issues related to language, language teaching and language testing (see CoE 2001: 1f). Researchers in the field of language acquisition as well as language testing criticize that research into Second Language Acquisition (SLA) has found little appreciation during the designing process of the CEFR. Hulstijn (2011: 204) points out that the CEFR levels, in their present form, are not fully based on empirical evidence taken from L2-learner performance. He further criticizes that they are not based on any theory rooted in the fields of linguistics or verbal communication. This has also been criticized by, amongst others Weir (2005), Alderson et al. (2006), Hulstijn (2007) and

¹ Please note that the term framework in the case of the CEFR and in the case of PT should be used in two distinct ways. With the CEFR, framework rather refers to the reference points that the CEFR documents provides which are deliberately non-dogmatic and do not favor a particular theory (see CoE 2001: 1). Pienemann (1998, 2005) in contrast, deliberately conceptualized his theoretical SLA framework in a modular way (see Pienemann 1998) to aim for theoretical parsimony (Pienemann 2005b). PT's framework thus is only focused on a specific area of psycholinguistic language development. I am well aware of the differences in theoretical frameworks and reference frameworks, but nevertheless use the term framework to refer to both the CEFR and PT.

Wisniewski (2017b). It is here that the second framework which this study focuses on comes into play. I aim to discuss Processability Theory (PT) (Pienemann 1998; 2005; Pienemann & Lenzing 2015), a psycholinguistic framework to SLA that predicts a universal developmental path which underlies morphosyntactic development in language development. I argue in this study that PT is able to add to the descriptive, theoretical and empirical basis of the CEFR in the areas of language production, and more specifically, in the area of grammatical competence. The focus of this study is an analysis of the CEFR in terms of grammatical competence through an SLA lens. I argue that grammatical competence presented in the form of a scale for grammatical accuracy in the CEFR (see CoE 2001: 112f. and 114) is neither learner-centered, nor theoretically-motivated or empirically-grounded. Thus, the aim of this study is to put forward a scale for *Grammatical Range* that combines PT and the CEFR (see chapter 4.4.6). I conceptualize the scale for *Grammatical Range* on the basis of an empirical study that correlates PT stages and CEFR levels. The correlations are based on oral production data of language learners who were assessed with Rapid Profile (Pienemann 1992), a semi-automatic diagnostic tool based on PT, and proficiency ratings with the help of the Overall Oral Assessment Grid provided by the CoE, based on the descriptors for Oral Production in the CEFR. I hypothesize that the combined scale for *Grammatical Range* is more learner-centered, adheres to recent research in SLA and is compatible with the universal, undogmatic notions put forward in the CEFR. The wording in scale for *Grammatical Range* remains as close to the original voice of the scale for Grammatical Accuracy as possible but integrates the universal processing procedures as spelled out by PT. Moreover, it does not contain any references to grammatical accuracy (see chapter 4.4.6).

One reason why the CEFR and PT are investigated in more detail is that I assume they share a particular feature in the perception of language professionals. I consider both frameworks to often be depicted in an insufficient way. Both frameworks are best known for one feature; the six level-global scale of language proficiency (see e.g. Little 2014) with regard to the CEFR (CoE 2001: 24) and the six stages of morpho-syntactic development in PT (see e.g.

Pienemann 2005a: 24). Often, all the concepts and notions that both the CEFR and PT are based on, are deemed equal at face value, especially regarding the scale or stages respectively. They are thus often criticized for being either, in the case of the CEFR, not specific enough, or, in the case of PT, too narrow. What readers and users tend to overlook in both frameworks is the massive body of operational, theoretical and empirical considerations that have gone into the development of both PT and the CEFR. This thesis, therefore, tries to explore both accounts in more detail and to explore to what extent PT might be used to add to theoretical and empirical gaps present in the CEFR. It is important at this point to note that I do not intend to assign equal status to the CEFR and PT. Each of the frameworks should be seen in their specific domain. The CEFR is a framework that is used as a reference tool for guidance for curriculum developers, teacher trainers and language testers. PT is a SLA theory that provides a universal account of explaining the developmental sequence observed in language acquisition. The CEFR is not intended to explain developmental stages and PT was initially not intended to be used for, e.g., curriculum design. However, I consider the CEFR open enough and PT powerful enough to be added to the CEFR so that it is worth examining interfaces more closely. I am well aware that there will be some theoretical issues that cannot be solved in this thesis. These relate mainly to the central assumptions in conceptualizing proficiency and competence (see discussions in e.g. Brindley 1998; Leclercq & Edmonds 2014). The definition and operationalization of these terms is a philosophical debate that has been going on for quite some time and will probably not be solved soon. However, I want to follow Brindley's argument that "we cannot wait for the emergence of empirically validated models of proficiency in order to build up criteria for assessing learners' second language performance" (Brindley 1989: 56). I argue that the same holds true, not only for assessment, but also for the provision of the CEFR as a reference tool. This is why my thesis attempts to add to the empirical and theoretical basis of the CEFR while being aware of these issues, so that a practical solution towards a more theoretically-sound and empirically-grounded scale for grammar might be found.

1.2 Research Questions and Hypotheses

In this study, I will explore the question “Is second language development reflected in the six level-scale of communicative proficiency described by the Common European Framework of Reference?” I argue that in order to find empirically-based interfaces, I first need to examine the role of grammar in Overall Oral Production because a) PT is mainly concerned with grammatical features, b) PT mainly focuses on the oral production of learners, c) grammar in the CEFR is but one component part of oral production and d) currently, there is no better way of empirically testing the CEFR scales for finding interfaces between SLA and the CEFR than by using rating procedures based on the CEFR scales. The two most important research questions are the following: 1) Are there correlations between PT and the CEFR? This question entails the exploration as to whether morpho-syntactic development, as explained by PT, is reflected in the CEFR. 2) Do rater experience and assessment grid use influence rating results? That is, do experienced raters behave differently from less experienced raters in terms of assessing learner language? My research questions are not limited to these two, as I am able to explore more issues due to my study design (see chapter 4.2). I put forward the following hypotheses: 1) There are correspondences between PT and the CEFR. I assume that the correspondences are stronger at the lower CEFR levels at which language production (i.e. lexicon and grammar) is more restricted and less elaborate². This hypothesis will be the basis for putting forward a combined scale for *Grammatical Range* based on PT and the CEFR. Regarding research question two, I hypothesize that 2) grammar is a crucial factor in determining the CEFR proficiency level of a language learner with and without an assessment grid and that less experienced raters are more prone to cling to grammar than more experienced raters. I assume that the reason for this is that grammatical accuracy is quite easy to assess since one can quickly determine whether a grammatical structure is incorrect. However, when

² Pienemann (1998: 232) explains this phenomenon with the concept of hypothesis space for development and variation in which he argues that the leeway of variational options, that might be produced by language learners, broadens when progressing in the developmental hierarchy.

it comes to the assessment of fluency, for example, a distinction between wrong or right cannot be made this easily since fluency is a rather fluid concept.

1.3 The Structure of the Thesis

The structure of this thesis unfolds as follows: Chapter 2 introduces the theoretical background of the study, i.e. chapter 2.1 describes the CEFR more closely. It starts out with providing some background information on the notion and aims of the CEFR. Chapter 2.1.1 describes the historical background to the CEFR with a special focus on the Threshold level (van Ek 1975, Trim et al 1980, Richterich 1983) as the most influential reference points to the CEFR. The advancement in the Threshold level is that it is based on an analysis of learners' needs. The aim of the Threshold level is to equip learners "[...] who want to be able to communicate socially on straight-forward every-day matters and lead a socially normal life when they visit a foreign country" (van Ek 1975: ii) with the necessary linguistic resources. I lay out the historical background to the CEFR in this thesis because a) as an appreciation of the massive body of conceptual work that has influenced the CEFR and, at the same time, b) to show that the different conceptual viewpoints sometimes lead to internal incongruities in the CEFR³. I assume that the incongruities are mainly based on the fact the CEFR tries to describe language use (and language proficiency) in a most holistic manner. Along with Brindley (1986; 1991; 1998) and Pienemann & Johnston (1987)⁴, I argue that language use (and language proficiency) is a matter too complex to be captured in one document only. Chapter 2.1.2 lays out the structure and notions of the CEFR in more detail. It describes the definition of communicative proficiency in the CEFR in the form of can-do statements, commonly arranged in six levels and three bands. The chapter describes the aim of the CEFR to

³ For example, the authors of the CEFR make explicit that they do not favor one model over another in order to remain undogmatic (CoE 2001: 1f.). However, they dedicate the majority of their view on communicative cultural competence to Byram's (1997) ICC model (see CoE 2001: 104f.).

⁴ Brindley (1986; 1991) and Pienemann & Johnston (1987) relate their criticism to the use of language proficiency rating scales; especially the Australian Second Language Proficiency Rating Scale. However, I assume that their reasoning also holds true for documents that aim at describing language proficiency holistically. In this case, it is the CEFR.

“overcome the barriers to communication among professionals working in the field of modern languages” (CoE 2001: 1). The reference levels provide “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe” (CoE 2001: 1). Chapter 2.1.1 highlights that the CEFR makes aspects of language proficiency explicit but that it does not develop them. Rather, the CEFR proposes to give language professionals the opportunity to maneuver and exploit the document for their context. Chapter 2.1.3 introduces the dimensions of language proficiency and competences depicted in the CEFR. It describes the action-oriented approach taken in the CEFR (see CoE 2001: 9), introduces the CEFR’s horizontal and vertical axes (qualitative and quantitative parts) and shows the arrangement of the scales in the form of the three bands and six levels (CoE 2001: 22). This chapter explores the descriptor formulation based on two surveys conducted in Switzerland more closely and gives a general overview over the terms competences and communicative competences provided in the CEFR. Chapter 2.1.4 explores language production and processes as part of communicative competences in more detail because these are the points that I argue PT can best relate to. PT is a psycholinguistic theory that mainly focuses on language production and that takes a processing view to the acquisition of second languages in terms of morpho-syntactic development. This chapter thus lays out the view taken in the CEFR on production and processes so as to be able to relate it to the assumptions made by PT in chapter 3. The same agenda is followed when introducing linguistic competences and grammatical competence in chapters 2.1.5 and 2.1.6. The chapter on linguistic competences shows that grammatical competence makes up one part of linguistic competences described in the CEFR. The other competence areas comprise: lexical, semantic, phonological, orthographic and orthoepic competences (CoE 2001: 109). The main components of linguistic competence are defined as “[...] the knowledge of, and the ability to use, the formal resources from which well-formed, meaningful messages may be assembled and formulated” (CoE 2001: 109). This chapter shows that many of the competences, despite grammatical competence, are represented in a variety of different scales on different aspects connected to use in that particular area,

such as the scales for vocabulary (see scales on Vocabulary Range and Vocabulary Control CoE 2001: 112). However, the only scale for grammar that is provided, is a scale for Grammatical Accuracy (CoE 2001: 114). Grammatical Competence and the scale for Grammatical Accuracy is described in chapter 2.1.6. In this chapter, I argue that the CEFR presents an insufficient picture of grammatical competence in only providing one scale for grammatical accuracy. I argue that this is especially problematic since the CEFR is mostly known for its scales and not its qualitative dimension. The core claim is that a focus on accuracy evokes the idea that the acquisition of grammar is mainly concerned with accuracy. However, ample research has shown that accuracy is not a measure of linguistic development (see e.g. Pienemann 1998: 137). Therefore, I assume that the provision of only one scale for Grammatical Accuracy is not learner-centered. My reasoning is that a combination of the scale for Grammatical Accuracy and the ideas of grammatical development, explained by PT, can give rise to a more learner-centered, theoretically-grounded and empirically-validated scale for *Grammatical Range* (see chapter 4.4.6). The ideas on language learning issues described in the CEFR and presented in chapter 2.1.7. of this thesis, are supposed to assist the arguments for combining the CEFR and PT, as laid out in chapter 3. The focus of chapter 2.1.7 is put on the learning and teaching of linguistic competences, as well as the role of learner errors in the CEFR. The latter is highlighted because the qualitative dimension of the CEFR states that learner errors are positive indicators of language acquisition. Yet, the CEFR fails to include these ideas into the scale for Grammatical Competence. Chapter 2.1.7 ends with a description of the ideas presented on language assessment in the CEFR. This chapter is provided in order to be able to follow the discussion on a combined CEFR-PT assessment given in chapter 3.

The remainder of chapter two presents aspects of the second framework in focus: Processability Theory. Chapter 2.2 briefly introduces the core ideas of PT; namely that PT takes a processing perspective to explaining the developmental schedule found in the acquisition of morpho-syntactic structures in SLA. PT argues that this developmental path can be explained by the make-up of the human language processor. The human language processor is largely

adopted by the ideas presented by Levelt (1989). Levelt's Blueprint for the Speaker (Levelt 1989: 9) thus forms one yardstick in PT and is laid out in chapter 2.2.1.1. The second yardstick in PT forms the formal theory of grammar; Lexical Functional Grammar (LFG) (Bresnan 2001). LFG is briefly sketched out in chapter 2.2.1.2. Bresnan (2001: vii) states that "LFG is a theory of grammar which has a powerful, flexible, and mathematically well-defined grammar formalism designed for typologically diverse languages." Thus, LFG represents the typologically plausible component of PT. After having described the two yardsticks in PT, I introduce the hierarchy of processing procedures explained by PT and how this hierarchy is applied to English. Thus, the processable options for English language learners are depicted. This chapter is especially important because I argue that the universal processing procedures, as spelled out by PT, can be integrated into the scale for grammatical accuracy of the CEFR in chapters 3 and 4. Chapter 2.2.3 introduces the concept of Hypothesis Space (Pienemann 1998: 232). Hypothesis Space illustrates that PT cannot only account for the universal developmental path in SLA but is also able to show that variation in language learning underlies systematic principles. The idea of variation is that it arises from the choices a learner has at any stage of development, given the constraints on processing. Given the two dimensions in SLA, development and variation, the question arises as to how it can be determined whether a linguistic structure has been acquired or not. This issue is accommodated for in chapter 2.2.4, which introduces the Emergence Criterion (EC) (Meisel et al. 1981). The EC is especially valuable in SLA research and language testing as it does not assume that accuracy can account for describing language development, but that the emergence of a linguistic structure should be at focus. This ties in with my argument that the one scale for Grammatical Accuracy presented in the CEFR paints an insufficient picture of grammatical competence. Therefore, I claim that the CEFR and PT should be combined in order to produce a scale for *Grammatical Range*. After having outlined the EC, the historical background to PT is presented. This chapter is analogous to chapter 2.2.1 on the historical background of the CEFR. Chapters 2.2.6 and 2.2.7 are concerned with applied issues regarding PT. In 2.2.6 the Teachability Hypothesis (TH) is introduced in connection to the

concept of developmental readiness. Pienemann (1985: 37) maintains that if a learner is developmentally ready to acquire a structure, i.e. if the structure to be taught is in accordance with the current developmental stage or slightly above it, “instruction can improve acquisition with respect to (a) the speed of acquisition, (b) the frequency of rule application and (c), the different linguistic contexts in which the rule has to be applied.” However, learner errors will be encountered frequently in the foreign language classroom, but the view of errors within the PT framework is not a deficient one. Rather, learner errors are seen as positive indicators for language development (Larsen-Freeman & Long 1991: 57). This is the view of errors that is adopted for the conceptualization of the scale for *Grammatical Range*. I argue that this positive view of learner errors reflects SLA in terms of morpho-syntactic development more truthfully than a focus on grammatical accuracy and that therefore, the scale for Grammatical Accuracy should be rearranged into a scale for *Grammatical Range* based on the universal assumptions made by PT. Issues regarding language assessment in terms of PT are discussed in chapter 2.2.7 on Linguistic Profiling and Rapid Profile (RP).

Chapter 3 constitutes a theoretical account to bringing the assumptions made by PT together with the descriptions of linguistic competences presented in the CEFR. This chapter engages with a theoretical account to finding interfaces between the CEFR and PT and aims at laying out the chances and challenges in combining the two frameworks. The chapter should be seen as background information to chapter 4 in which I present the scale for *Grammatical Range* based on my empirical study. My claim is that grammatical competence is insufficiently depicted and not very learner-centered in the current version of the CEFR. I claim this because the only scale for grammatical competence that is presented in the quantitative part of the CEFR is the scale for Grammatical Accuracy. However, accuracy does not mirror language development (see chapter 2.2.4). Since PT focuses on universal aspects of grammatical development, I assume that PT is able to inform the CEFR in the area of grammatical development. The modular approach taken in PT that puts the processing of linguistic features at the center, is able to be integrated into the CEFR because it focuses on only one discrete subtask of SLA and the CEFR is

structured in a way that it describes several subtasks (see chapter 2.1.2). While I do not argue that all of the ideas in the CEFR are compatible with PT, I assume that the CEFR is open enough to embrace features of language processing as proposed by PT (see chapter 2.2.2). In order to explore the issue of SLA-based interfaces to the CEFR, chapter 3.1 presents prior studies on the CEFR from the SLA field. Since there is only a small body of research that investigates oral language production and/or grammar in the CEFR, the studies presented in this chapter have various different foci. They mainly propose language-specific interfaces between SLA and the CEFR and fail to provide congruent results due to methodological issues and the different research foci. None of the studies present theoretically-motivated, empirical interfaces between SLA and the CEFR. Chapter 3.2 focuses, more specifically, on interfaces between PT and the CEFR. Only three studies have investigated the two frameworks prior to the present study. The studies did not explicitly focus on finding interfaces between the CEFR and PT but rather on inter-rater reliability issues or PT and CEFR combined assessment. However, the studies provide a first impression on where CEFR-PT interfaces might be found. After having presented prior studies, chapter 3.3 takes an integrative approach to the CEFR and PT from a theoretical perspective. It discusses the terms universality, emergence and accuracy as well as competence, progression and processes from a CEFR and PT angle. This chapter shows that there are quite significant differences in some of the notions, but I argue that these differences do not lead to an incompatibility of the two frameworks. Therefore, I assume that it is valuable to propose a scale for *Grammatical Range*. Chapter 3.4 discusses issues on assessment from both fields. I argue that a combined assessment with Rapid Profile and proficiency ratings based on the CEFR will lead to more reliable and valid results.

Chapter 4 lays out the details of the empirical study. The aim of the study is to explore whether PT can add to the descriptive basis of the CEFR in terms of grammatical ability. My assumption is that grammatical competence is underdeveloped, especially in the quantitative part of the CEFR, because only a single scale for grammatical accuracy is presented (see chapter 4.1 on the rationale). All the other component parts of linguistic competences are equipped

with more than only one scale (see e.g. the different scales for vocabulary CoE 2001: 112). Furthermore, I argue that grammatical competence in language learners cannot be captured by scale for accuracy because language learners necessarily make mistakes during their language acquisition process (these ideas are based on Pienemann's 1998 argument that accuracy does not mirror language development; see chapter 2.2.4 for more details). I therefore hypothesize that it is especially valuable to include the universal assumptions about SLA made by PT in the CEFR in terms of grammatical competence. Chapter 4.2 specifies my two main research questions and my hypotheses. Chapter 4.3 lays out the details of my methodology, including the innovative approach to determining the role of grammatical accuracy in oral language proficiency ratings. Chapter 4.4 presents the results of my analyses. The chapters are divided into statistical, quantitative results as well as more qualitative accounts to analyzing the data. I consider both accounts fruitful, especially for the second research question, as together they provide a better insight into how raters administer their ratings. A discussion of the results is given after each chapter individually. The thesis ends with the conclusion in chapter 5, the list of references (chapter 6), as well as an appendix that contains exemplary transcriptions of the learner data and details of my pilot study.

2. Theoretical Background

In this chapter, the constructs of the CEFR and Processability Theory are laid out as both frameworks form the core of this thesis. In the following paragraphs, I will first present the historical cornerstones and some background information regarding the development of the CEFR. I will then describe the sources that shaped the form of the CEFR as it is known today. Major principles in the CEFR are laid out. I decided to write this chapter closely to the primary sources as it is important to reflect the original voice of the CEFR, and not to alter its definitions. Chapter 3 about the interfaces between the CEFR and PT will then consider more

secondary sources and post-hoc interpretations of the CEFR descriptors by various scholars.

2.1 The Common European Framework of Reference

The Common European Framework of Reference is a document published in 2001 by the Council of Europe that defines (communicative) competences of language users, to promote plurilingualism and life-long learning across European member states. The CEFR is supposed to be regarded as a holistic, but never exhaustive reference tool that is intended to be used by language professionals (e.g. curriculum designers, language test providers, etc.) to integrate notions of learner-centered, communicative, and diversity-appreciative language learning and teaching (see CoE 2001: 2). The CEFR was written based on a review of research initiated by the Council of Europe (CoE) to provide illustrative descriptors for various language competences and situations. Those descriptors usually⁵ follow a structure of 6 levels organized in three bands. The three bands are strongly informed by the development of the Threshold Level by Van Ek (1975). It is worth tracing back some of the history that shaped the CEFR at this point, in order to better understand the point of departure of the CEFR for describing communicative competences. More details on the CEFR in its 2001-version will be given in chapter 2.1.2.

2.1.1 Historical Overview

The Common European Framework of Reference is a document that resulted from ongoing research projects which originated around four decades ago (Little 2007: 174) and were initiated by the Council of Europe.⁶ According to Little (2006: 174), the CoE was “[...] founded in 1949 to defend human rights, parliamentary democracy and the rule of law, develop agreements to

⁵ There are a few exceptions, such as the scale for mediation, that does not specify the lowest or the highest level, i.e. A1 and/or C2.

⁶ For more information about the CoE and further policy documents, see its official homepage: www.coe.int and http://www.coe.int/t/dg4/linguistic/historique_en.asp

standardize social and legal practices in the member states, and promote awareness of a European identity based on shared values” during times in which Europe was still under post-war influences. In order to achieve the goals listed above, and to guarantee mobility between European member states, it is evident that continuous education⁷ and language learning needs to be promoted. Little (2006: 174) argues that “[...] mutual understanding, effective educational and cultural exchange, and the mobility of citizens all require large-scale and successful language learning.” The CoE’s research projects yielded, amongst a resolution⁸, a number of recommendations that were defined by the Committee of Ministers and Parliamentary Assembly of the Council of Europe. Recommendation (82) 18, for example, states the aim “to achieve greater unity among its members’ [...] by the adoption of common action in the cultural field“ (CoE 1982: 1).

From the literature available, it is not easy to pinpoint the exact onset of the development of the CEFR as it was a continuous process. North (2007: 23) claims that the gradual process in which the CEFR levels have emerged started in 1913 with the Cambridge Proficiency Exam that was later merged into level C2. In his 2014 publication however, North traces the origin back to the 1960s in which he claims the “[...] history of the CEFR really starts [...]” (North 2014: 14). A diachronic approach to describing its development might be beneficial.

The official websites of the Council of Europe give the following overview of events that shaped the development of the CEFR:

⁷ See the Council of Europe’s Lifelong Learning Initiatives as an example for its support of continuous language learning, <http://pjp-eu.coe.int>.

⁸ To access the resolution, recommendations and other official documents, see http://www.coe.int/t/dg4/linguistic/20thsessioncracow2000_EN.asp#TopOfPage and http://www.coe.int/t/dg4/linguistic/Conventions_EN.asp#TopOfPage

Key moments in history	
1957	First intergovernmental conference on European co-operation in language teaching
1963	Launch of first major project in language teaching
1975	Publication of 'Threshold Level' specification
1989	New member states begin to join intergovernmental projects
1994	European Center for Modern Languages established
2001	European Year of Modern Languages Common European Framework of Reference for Languages European Language Portfolio

Figure 1: Overview of historical events shaping the CEFR, taken from http://www.coe.int/t/dg4/linguistic/historique_en.asp (last access 20.06.2018)

After the European Cultural Convention (ECC) that aimed at “furthering greater understanding of one another among the peoples of Europe” (CoE 1954: 1) was signed in 1954, the first intergovernmental conference on European co-operation in language teaching took place. This conference concluded with the establishment of various medium-term research projects. One of those is the Major Project in Modern Languages (1963-1972) that “promoted international co-operation on audio-visual methods and the development of applied linguistics, including support for the founding of the International Association of Applied Linguistics (AILA)” (http://www.coe.int/t/dg4/linguistic/historique_en.asp).

According to North (2014: 14), the focus shifted from the development of audio-visual aids for language education, to the specification of a “European-wide credit scheme for adult language learners of modern languages” in the projects that took place from 1971-1977. With the aim of coordinating such a credit scheme, the Rüschtikon Symposium was held in 1971 in Switzerland. North (2014: 14) describes its focus to be on three major aspects: “(a) new forms of organisation of linguistic contents, (b) types of evaluation with a unit/credit scheme and (c) means of a unit/credit scheme in the teaching/learning of modern languages in adult education”. The major achievement at that time was

the development of a functional-notational⁹ approach to describing learning objectives that led to the specification of a Threshold level. The Threshold level specified “in operational terms what a learner should be able to do when using the English language” in the area of language production (see http://www.coe.int/t/dg4/linguistic/historique_en.asp and van Ek 1975). The unit/credit scheme was influenced by Schwartz’ system of learning and assessment units *unite capitalisables* (Schwartz 1974). The *unites capitalisables* contain that “[...] wherever possible, subjects should not be taught or examined globally, but broken down into constituent parts, which could be taken one by one as learners were ready to do so” (Trim 2007: 14). The Rüsclikon Symposium in 1971 then put together several work-parties, featuring (applied) linguists and curriculum designers René Richterich, David Wilkins and Jan van Ek, who should break “[...] down the global concept of language into units and sub-units based on an analysis of particular groups of adult learners in terms of the communicative situations in which they are characteristically involved” (Trim 2007: 15). The idea was that these situations would be able to account for language learning across national boundaries. Milanovic and Saville (2012: xiii) note that “[...] by the mid-1970s [...]”, it had become clear that “[...] it was not possible to divide up language learning into discrete modules [...]”, as this would be most arbitrary and imposing on European member states.

Trim (2007: 16) summarizes the subsequent lines of thought as follows:

The group therefore felt it to be more appropriate to support independent decision-making as close as possible to the point of learning by setting out general aims and principles, providing models which practitioners could adapt to their own circumstances and encouraging the exchange of ideas and experience amongst them. The first priority therefore attached to the serious consideration and formulation of the fundamental principles upon which a long-term European language policy could be based.

⁹ The approach is different from the situational approach to teaching that was popular at that time. Trim (2010: xxiv) argues that situations as the basis for spelling out learning objectives were based on contextualized dialogues that represent unique events. He and Wilkins suggested to rather focus on patterns of communicative interaction for determining learning goals. The idea was that the concepts and notions underlying those interactions give rise to a classification of language functions which, in turn, can be used to identify learning objectives. Those functions could encompass specific situations as well as more general features. See Barnett (1980) for more information on the situational approach.

The *Language Learning and Teaching for Communication* projects, especially project number 12 (see Trim 2007), that were operational between 1981 and 1988 utilized the specifications of the prior projects, along with Recommendation No 82, to “reform curricula, methods and examinations throughout the 1980s” (Trim 2007: 16) and closely cooperated with teacher trainers to spread notions of how “to implement a more communication-oriented language-teaching approach relying on a wide range of methodologies in order to cater for the various teaching situations” Trim (2007: 30).

The projects that followed, integrated, amongst others, the bilingual education sector and focused more extensively on the concept of plurilingualism (see Trim 2007: 31ff), which resulted in spelling out Recommendation No R (98) 6 by the Committee of Ministers. R (98) 6 “[...] emphasises intercultural communication and plurilingualism as key policy goals” (http://www.coe.int.-/t/dg4/linguistic/historique_en.asp). At another symposium at Rüslikon, that took place 20 years later than the former one, the aims for the development of Common Reference Levels were broadened to cover linguistic and cultural diversity. Furthermore, the groups set goals for developing reference levels. North (2007: 21) summarizes these goals as follows:

- To establish a meta-language common across educational sectors, national and linguistic boundaries that could be used to talk about objectives and language levels [...].
- To encourage practitioners in the language field to reflect on their current practice, particularly in relation to learners’ practical language learning needs, the setting of suitable objectives and the tracking of learner progress.
- To agree common reference points based on the work on objectives that had taken place in the Council of Europe’s Modern Languages projects since the 1970s.

The new agenda thus specified the contextualization of the earlier work of the Modern Language Projects. It is stated that this endeavor was soon abandoned, as it was considered impossible to develop a unit-credit scheme that sufficiently breaks down all aspects involved in language learning into a set that could be applied universally. Rather, the idea evolved to develop a document that most holistically captures ideas from (at that time) current research, which could serve

as a reference book that language professionals could consult when they wanted to align their work to a European consensus (see Schärer & North 1992: 1ff.).¹⁰

The outcome of North's work and the CoE agenda was mainly influenced by Trim, Coste, North and Schärer. It was a draft version of the CEFR published in 1996 (CoE 1996). After this version had been revised, the official document was commercially published in the *European Year of Languages* in 2001. Since then, the CEFR has undergone extensive (post-hoc) research and has been used in areas of curriculum development and language examinations. In 2007, the *Languages of schooling within a European Framework for Languages of Education: Learning, Teaching and Assessment* intergovernmental conference gave opportunities to discuss policy issues raised by the CEFR and its wide-spread use of proficiency levels

(CoE: http://www.coe.int/t/dg4/linguistic/conference_bis_en.asp#P40_1517).

Currently, a research team is conducting a number of studies on the extension and development of the CEFR that invite practitioners all over Europe to participate. This survey seeks to explore the use and usefulness of new descriptor additions that cover, amongst others, the validation of descriptor-items for mediation more closely (see Qiriazzi & North in prep. and North & Panthier 2016).

This brief, and by no means exhaustive description of the research tradition that preceded the development of the actual CEFR descriptors, depicts the multitude of ideas, concepts and policies that influenced the current descriptors of language competence. It also explains the origin of the idea of the CEFR to picture the language learner "as a social agent", whose development of their "whole personality and sense of identity in response to the enriching experience of otherness in language and culture" is supposed to be promoted by an intercultural approach (CoE 2001: 1). In this context, the question arises as to what exactly these notions are and where the notions come from.

¹⁰ At the same point in time, Schärer (1992: 3) reports the development of a European Language Portfolio (ELP) to "systematically report learner progress and achievement" within the context of a European Framework. The ELP still is a widely-used means for self-assessment in school, curriculum and testing contexts with recourse to the CEFR.

In order to understand the notions of language competence displayed in the CEFR, it might be helpful to briefly describe its predecessors and major sources of influence at this point. In this way, this section will contribute to understanding my assumptions about interfaces between the CEFR and Processability Theory made in chapter 3 and 4.

The official CEFR document was shaped by the research its working parties conducted and reviewed. Harsch (2006: 2), with recourse to the German translation of the CEFR, points out that the CEFR claims to summarize the state of the art of language (education) research, in order to introduce levels of language competence that describe aspects of knowledge that learners use in varying situations. It is quite interesting to see that the development of the CEFR seems to mirror the trends in applied linguistics for language teaching purposes. These seem to go hand in hand with the development of the notions in the CEFR. As far as I understand it, this connection is appreciated by the authors of the CEFR, in that they want to provide a most holistic but not exhaustive reference tool for language professionals. On the other hand, the multitude of concepts that can be found in the CEFR, reduce its readability and makes it hard to connect its qualitative and quantitative dimension so that at times, the document is characterized by internal contradictions (see chapter 2.1.2, 2.1.3, 3.3.3.1 for more details).

In the earlier CoE projects, the aim was to come up with a unit-credit scheme that allows “the fully participatory development of language learning systems appropriate to different learning situations at different times and places” (Trim 1978: 22). The major projects of influence within the Projects of Modern Languages, are summarized by Little (2007: 174) as follows:

- i) The analysis of learners’ needs (Richterich 1983, Richterich & Chancerel 1978, Porcher 1980)
- ii) The development of a notional-functional approach (Wilkins 1973, 1976)
- iii) Based on notational-functional approach, the discrimination of Threshold levels (van Ek 1975, Richterich 1983)
- iv) The elaboration and promotion of learner autonomy (Holec 1979)

For reasons of clarity and comprehensibility, I will refrain from going into detail about all four major projects above. These works are not the only ones that were taken into account in the development of the CEFR and due to the limited scope of this thesis, I will not discuss them in detail.¹¹ A brief summary of the views taken in points i, ii and iv will be given and only point iii will be discussed in more detail because point iii constitutes the most influential view that shaped the CEFR. In what follows, I will thus focus on briefly describing the most influential approach, namely the Threshold Level (van Ek 1975; van Ek & Trim 1998). This approach attempts to integrate the core aspects of the other research projects into their concept of a language learning level. I consider it helpful to briefly discuss the Threshold level that shaped the current version of the CEFR, as it will help to gain a deeper understanding of the notions that the CEFR adopts.

Van Ek (1975: 5) proposed the Threshold Level (t-level) that was developed to establish a European unit-credit scheme for foreign language learning for adults. It comprises operational language learning objectives that were formulated against the background of the English language and focus on oral language production. There are only a few instances in which van Ek makes recourse to writing and reading as a skill as he argues “[...] the learners’ need to use the foreign language orally will be much greater than their need to use its written forms”, because reception is seen as an integral skill to speaking (van Ek 1975: 17).

T-Level specifications are based on the analysis of learners’ needs (see Richterich 1973, Richterich & Chancerel 1978, Porcher 1980). Learners are conceptualized as temporary visitors to other countries and the Threshold Level gives reference to (1) the role that the learners as language users play, (2) the settings in which they play these roles, as well as (3) the topics that the learners deal with in communication (van Ek 1975: i). As to the target group of the t-level, van Ek (1975: ii) specifies that it “[...] is seen as people who want to be able to communicate socially on straight-forward every-day matters and lead a socially

¹¹ I will deliberately skip all aspects of innovative research on plurilingualism and language policy generated by the projects above. I do so, as this research is not related to the aim of my study. For an overview of the works from which the CEFR authors distilled their concepts, see Trim (2007: 17ff) and Little (2006: 174ff.).

normal life when they visit a foreign country.” The t-level is thus seen as a tool for the performance of communicative functions, and neither as a finite set of lexis and grammar, nor as a recommendation of a vocabulary list minimally required in communicative situations. The term *objective* is defined in the t-level in terms of behavior that enables the learner to do something that s/he was not able to do at the beginning of the learning process (van Ek 1975: 4).

The operational objectives are based on descriptions of:

- a) situations in which the foreign language is used by learners,
- b) language activities in which language learners engage,
- c) language functions that learners fulfil,
- d) topical specifications that the learner will use,
- e) general and specific notions that learners will handle,
- f) specific lexical and grammatical forms that learners will use,
- g) a few details how well the learner will perform all of the above.

These objectives are supposed to mirror the improvement in the ability to use the foreign language in various situations. The principles are spelled out on the basis of situations based on which specific details in language use are hypothesized. In my view, the use of situations as a basis for spelling out operational objectives seems quite reasonable at first glance. However, I consider it impossible to anticipate every situation in which language learners might use the language. Additionally, I presume that this way of spelling out learning objectives assumes a rather unproductive, uncommunicative way of language learning. It suggests to best prepare learners for specific situations in which they only have to use a certain number of predefined phrases; i.e. in which a limited number of communicative options are assumed. Subsequently, the learner would most likely be lost in unforeseen communicative situations. This is why currently, a more communicative approach to language teaching is adopted, which prepares learners to use the target language creatively. However, it should be noted that van Ek’s proposal is a milestone in standardizing language objectives that are applicable to various language learning environments.

I will now briefly exemplify the points above based on van Ek's (1975) document. Situations (a) in which the learner will find him/herself are described according to social roles (such as a friend/stranger) (van Ek 1975: 10), psychological roles (being neutral or equal to others, showing sympathy or antipathy) (van Ek 1975: 11), settings in which the learner uses the target-language (such as indoors or outdoors, in public or private life, etc.) (van Ek 1975: 12) and the topics communicative acts will cover (such as personal identification, house and home, trade, occupation, etc.) (van Ek 1975: 13). Language activities (b) are described in terms of the different language skills, e.g. speaking, reading, understanding and writing (van Ek 1975: 17f.). Regarding *understanding*, for example, activities are concerned with understanding "the texts of the commonest announcements via public address systems in airports, at railway-stations, etc." (van Ek 1975:17). Language functions (c) to refer to non-language-specific functions that are distinguished in 6 main categories of verbal communication (van Ek 1975: 19). These are: (1) imparting and seeking factual information, (2) expressing and finding out intellectual attitudes, (3) emotional or moral attitudes and (4) socializing and suasion. In order to seek factual information, the learner would need to be able to identify information and ask for something. The behavioral specifications for topics (d), encompass amongst many others, the need to "describe their own accommodation and seek familiar information from others" (van Ek 1975: 22). General notions (e) refer to concepts that "people use in verbal communication [...], which are heterogeneous in that they represent a wide variety of levels of abstraction." (van Ek 1975: 29). Here, notions of properties and qualities, such as existential (presence/absence of something) or spatial and temporal (such as location and dimension; size and length) are given. Specific notions (e) are topic-related and van Ek (1975: 33) argues that a "[...] method for selection of these notions is to a very large extent subjective; it is based on introspection, intuition and experience." Specific notions should be seen in relation to general notions. Examples would be *to call someone, to give an address, etc.* (van Ek 1975: 66). Language forms (f) are language-specific to English and marked for the grammatical categories that underlie their production/reception (van Ek 1975: 33). From the literature, it is

not entirely clear as to how these forms were selected. Van Ek remains rather vague on this aspect. For *identifying* under the function (c) of *imparting and seeking information*, for example, the author gives the following specifications: demonstrative pronouns (this, that, these, those) + be + NP. Point (g) relates to the degree of skill that a learner displays. Van Ek (1975: 112) explains that the degree of skill (g) is only briefly touched upon, as it was not the core objective in the development of the t-level. The author makes suggestions for testing that include, amongst others, reasonable speed, sufficient precision and reasonable correctness (van Ek 1975:114). Based on van Ek's proposal, it cannot be fully determined as to how the roles, activities, functions and forms came about. It is, however, a first proposal to systematize language learning instances and standardize learning objectives.

North (1992) takes up on the t-level in his PhD project, which sets the ground for the Common Reference Levels that are known today. North gives a detailed account on the ideas that had been discussed for underpinning the Common Reference Levels. He describes the dissociation of the Reference Levels to the first proficiency scale that was available, namely the Interagency Language Roundtable scale, developed in the USA in the 1950s¹². He also describes the many-faceted *Rasch item response model* (Rasch 1992) that was used to scale the descriptions of language ability provided by language teachers. The *Rasch item response model* is situated within *Item Response Theory* and is a probabilistic psychometric statistical means to, inter alia, order items of a test according to their difficulty in response to a person's ability (Bond & Fox 2015: 11).¹³ In the case of the CEFR, the teachers' interpretations of the level of a descriptor were measured with the *Rasch Model*. This methodology lets North (2014: 24) conclude that the levels show an "empirically proven interpretation of difficulty". A problem with this kind of methodology is that the scaling of

¹² The United States Foreign Service Institute put forward the Interagency Language Roundtable (ILR) scale. It was developed for determining whether someone had the language ability to engage in diplomatic and intelligence activities. North (1992: 10) points out that this 5-level scale is rather product-oriented, purely interested in results rather than a continuous learning process, and that it showed biases to high levels of proficiency and therefore not applicable for a Common Reference book.

¹³ Item Response Theory "is built around the central idea that the probability of a certain answer when a person is confronted with an item, ideally can be described as a simple function of the person's position on the latent trait plus one or more parameters characterizing the particular item" (Molenaar 1995: 4).

teachers' perceptions of progression in the scales is done in a post-hoc fashion, and not based on theoretical assumptions. Wisniewski (2017a: 2) highlights the main points of critique of the CEFR. She argues that the CEFR levels face considerable challenges, "[...] many of which are related to the scaling methodology." She summarizes that the CEFR levels have been criticized for a lack of consistency because the descriptors were chosen one by one, following the criterion of their statistical quality, so that some concepts only appear at single levels (Wisniewski 2017a: 2). She shows some more obvious problems of the CEFR, such that as some descriptors are vague,¹⁴ whereas others are self-referential¹⁵ or subjective¹⁶ (Wisniewski 2017a: 2). Wisniewski (2017a: 3) also establishes a discussion of more fundamental problems, such as that the relationship of the descriptors to SLA is unclear, so that it cannot be argued that they reflect language development. Another major drawback, she points out, is that "[...] the exclusively teacher-based scaling perspective [...]" found in the descriptors was never empirically validated, so it is not clear if the teachers' perceptions actually match authentic learner behavior (Wisniewski 2017a: 3). One of Wisniewski's major concerns is – and it ties in with the suggestions made in this thesis – that the descriptors were not derived from theory (Wisniewski 2017a: 3).

North summarizes further criticism of proficiency levels from various angles: Frawley & Lantolf (1985), Lantolf & Frawley (1988), for example, claim that it is impossible from a philosophical point of view, to capture the concept of language proficiency. SLA scholars, such Brindley (1991), Pienemann; Johnston and Brindley (1988), mainly criticize the use of proficiency scales for their circularity, unidimensionality¹⁷ and norm-referenced nature. The language testing community mainly criticized the lack of precision in determining concepts

¹⁴ See the vocabulary control scale for level C1 "[...] occasional minor slips [...]" (CoE 2001: 112).

¹⁵ See the fluency scale at C1 level "Can express him/herself fluently [...]" (CoE 2001: 129).

¹⁶ See the fluency scale at B2 level "[...] regular interaction with native speakers quite possible without strain for either party." (CoE 2001: 129).

¹⁷ Bachman (1990: 203) explains the problem with unidimensionality in tests that "[...] make the specific assumption that the items in a test measure a single, *unidimensional* ability or trait, and that the items form a *unidimensional* scale of measurement". Henning (1992) concludes that a distinction between psychometric and psychological dimensionality has to be made. He argues that psychometric unidimensionality is a rather inconsistent concept, since language measures cannot focus on a single trait.

such as “many”, “some” in the descriptors contained in the scales (see Alderson 1991), as well as the confusion of traits with their elicitation methods (see Bachman 1990). North counters the arguments above by stating that the Reference Scales are not supposed to be a theoretical model, but rather constitute an operational model to defining language proficiency (North 1992: 7). North describes the difference between both to be as follows: an operational model is much simpler than a theoretical one, as it reinterprets elements so that they can be used in particular contexts. He further argues that even theoretical accounts do not describe reality, but that they would make ideas about experience explicit. For the context of a universal European Reference tool, an operational account seems more appropriate to him. The problem with North’s argument is that an operational definition without a theoretical and an empirical basis lacks validity. The authors of the CEFR argue however, that validity is the quality criterion in language education that the CEFR is most concerned with. If this is the case, then an operational model to defining language proficiency, in my view, is not sufficient.

Initially, the CoE only focused on the adult language learning sector, see (CoE 1973). Later, connected to developments beyond the t-level (Van Ek 1976), this focus was expanded to beginner and more advanced learners as waystage and vantage levels were designed. With John Trim as a core researcher, the 1973 document is guided by the principles of analyzing learners’ needs, in order to determine what they have to learn to fulfil those needs.

To summarize, the t-level outlines incidences, situations and topics that a language learner needs to handle when engaging in communication during a temporary visit to another country. The t-level was the major source of influence of the CEFR. However, the use of situations as a basis for spelling out learning objectives bares a number of problems. Since these problems, too, have influenced the CEFR, many researchers expressed concerns as to the theoretical soundness and validity of the CEFR. The aim of this thesis is to find interfaces between SLA theory and the descriptive nature of the CEFR to add to its validity. After having given a brief sketch of the history of the CEFR and having summarized one of the preceding documents, I will now turn to the CEFR in its

current version. As stated above, the CEFR is currently edited for scales of mediation and pre-A1 levels. However, since those have not been validated up to the point of writing this thesis, I will use the 2001 version as the major source of my description.

2.1.2 The Structure of and Notions in the CEFR

The CEFR consists of 9 chapters and several appendices. Chapters one and two describe the political and educational context of the CEFR and specify the action-oriented approach adopted¹⁸ in the document. Chapters four and five give a taxonomy of language competence, knowledge, skills and characteristics; stating how, inter alia, competences, domains and strategies are defined. Green (2012: xxxvi) maintains that chapters six, seven, eight and nine of the CEFR are a “[...] survey of methods of learning, teaching and assessment [...]” that present language professionals with an “[...] open, non-dogmatic account of the various options open to them [the language professionals], to encourage reflection of their own current practice, to consider alternatives and communicate to others their opinions and their reasons for holding them” (additions by KH).

According to Harsch (2006: 3), the CEFR follows a holistic view in describing language competences with the help of a taxonomic descriptive scheme in the areas of reception, production, interaction and mediation. Based on that description, more than 57 scales in total (North 2007: 566) have been produced with ongoing work on scales for mediation, as well as the production of pre-A1 levels to account for early learners. These scales give a definition of communicative proficiency in various situations and commonly¹⁹ follow the structure of 6 levels arranged in three bands - A1 and A2 (basic user), B1 and B2 (independent user), C1 and C2 (proficient user). These levels contain descriptions

¹⁸ The action-oriented approach is discussed in chapter 2.1 in more detail.

¹⁹ I used the term commonly here because the extended version of illustrative descriptors that is being piloted upon writing this thesis includes scales that have been condensed to only 3 levels, such as those scales for using text, e.g. EXPRESSING A PERSONAL RESPONSE TO LITERATURE, which uses the labels of the 3 bands: basic, independent and proficient user (see CoE 2016: 61ff.). Also, some descriptors for levels of competence below level A1 have been included, see 2014-2016 projects (CoE 2016: 9ff.). Interestingly, the newer version also deleted all references to native speaker-like competence and substituted this term with *proficient* in response to ongoing criticism about the native speaker-term as an ideal for language learners.

of language competence utilizing a *can-do* approach. The *can-do approach* results from the functional-notational approach that was taken up by the t-level in response to the rather deficient but predominant approach to language assessment that used cannot-do statements. The scales have been most influential within and even across European boundaries. Hulstijn (2007: 663), with recourse to Little (2007), observes that its strong influence

[...] might well be caused by its combination of what is familiar (the traditional distinction between “beginner,” “intermediate,” and “advanced” levels) and what is new (an elaborate system of descriptors giving communicative content to the levels of beginner or basic, intermediate or independent, and advanced or proficient).

In its introductory chapter, the authors of the CEFR clarify that the document is intended to “overcome the barriers to communication among professionals working in the field of modern languages” (CoE 2001: 1). Professionals are “educational administrators, course designers, teachers, teacher trainers, examining bodies, etc.” (CoE 2001: 1). It is important to highlight here that the authors “have NOT set out to tell practitioners what to do or how to do it” (CoE 2001: xi; capitalization in original) but rather to provide, in North’s (2007: 21) words “[...] a concertina-like reference tool that provides categories and levels that educational professionals can expand or contract, elaborate or summarise, according to the needs of their context.” The CEFR thus makes aspects of language proficiency explicit but does not develop them. Rather, the CEFR proposes to give language professionals the opportunity to maneuver and exploit the document for their context. The Reference levels provide “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe” (CoE 2001: 1) in that they give a “descriptive scheme that can be used to analyze L2 learners’ needs, specify L2 learning goals, guide the development of L2 learning materials and activities, and provide orientation for the assessment of L2 learning outcomes” (Little 2006: 167). Four major points are important in this context: 1) The CEFR is to be seen as a reference book that is not fully developed, but to be used as a point of reference for the alignment of curricula, syllabuses and language tests. 2) It does not favor a particular theory but remains open and non-dogmatic. 3) It is not a testing tool,

but language professionals are free to exploit the document for their purposes. 4) It is based on informed teacher's perceptions of language proficiency that were statistically scaled and cover four skills; production, reception, interaction and mediation. The careful reader of the CEFR might find that in many aspects, the document shows some internal contradictions (see chapters 2.1.2, 2.1.3, 3.3.3.1 for more details). Let us now turn to the horizontal and vertical dimension of the CEFR that aim at modeling language proficiency holistically.

2.1.3 Dimensions of Language Proficiency and Competences

An aspect that is often neglected when examining the CEFR is that it encompasses 2 dimensions, a horizontal and a vertical axis. The vertical dimension presented in chapter 3 contains the Reference Levels, i.e. the scales that form probably the most well-known part of the CEFR. The authors of the CEFR maintain that Common Reference Levels need to meet descriptions and measurement criteria to be applicable and relatable across national boundaries (CoE 2001: 24). The question arises here whether the authors of the CEFR would equate a nation with a language and thus imply the notion of a Eurocentric, monolingual nation-state, despite promoting a multilingual Europe. Following this line of thought, the user of the CEFR might assume the following linguistic assumptions in the CEFR: there is a universal communicative basis that can be related to all human language, and the resulting levels can be compared across target languages. This is an argument that would be disputed by many language typologists.

On the descriptive basis, the authors maintain that the CEFR needs to be *context-free* but still *context-relevant* to remain applicable to various language education backgrounds. It also needs to be based on *theoretical work* on language competence but should still be *user-friendly* to encourage reflection on what competence means for practitioners in their context (CoE 2001: 24). Measurement issues relate to *objectivity* and the *number of levels* employed to show progression. The CEFR posits that measurement should not be based on intuition as it is subjective, but that a scale of progression should be based on an

ongoing process of validation and analysis (CoE 2001: 22). The structure suggested by the CEFR is a six-point scale presented in Figure 2 below:

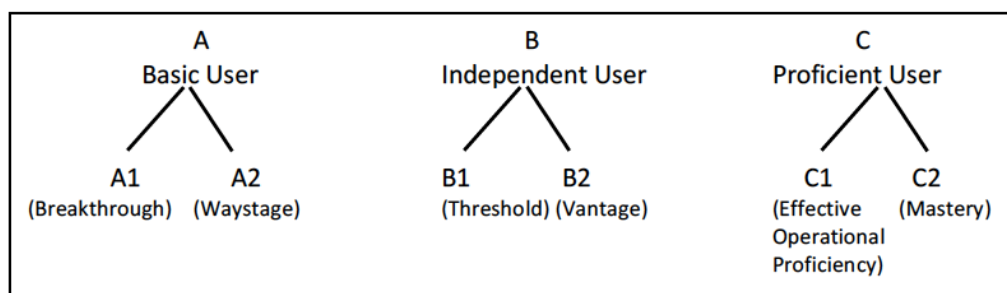


Figure 2: The three bands and six levels of the CEFR

The Figure presents the three basic bands, A – the basic user, B – the independent user, and C - the proficient user. These are subdivided into the levels A1 (breakthrough), A2 (waystage), B1 (threshold), B2 (vantage), C1 (effective operational proficiency) and C2 (mastery). Here, the relations of the CEFR to the preceding projects administered by the CoE become apparent. The authors of the CEFR (2001: 23) explain that “Breakthrough” relates to Wilkins’ (1978) proposal of formulaic proficiency and Trim’s (1978) publication of “Introductory”. The levels “Waystage” and “Threshold” relate to the content specifications given by Van Ek (1975). “Vantage” refers to the level of limited operational proficiency as described by Wilkins and Trim. Effective operational proficiency reflects an “[...] advanced level of competence suitable for more complex work and study tasks” (CoE 2001: 23). “Mastery” refers to the highest objective as spelled out by ALTE the Association of Language Testers in Europe (ALTE) and occurs at the top end of the scale (see CoE 2001: 23). For each of the levels, illustrative *can-do* descriptors are presented that were developed and validated based on results of a project conducted by the Swiss National Science Research Council (1993 - 1996) (CoE 2001: 217). The descriptors were written based on two surveys, completed by around 300 teachers and around 2800 learners who are supposed to represent about 500 different classes, ranging from lower secondary school to adult education in Switzerland (CoE 2001: 217). The CEFR does not further elaborate on the make-up of the questionnaires that rely mainly on the PhD thesis by North (see 1996). As far as the methodology is concerned, it is explained

as follows: After an intuitive phase,²⁰ in which an analysis of existing scales of proficiency as well as a deconstruction of those scales into descriptive categories took place, a qualitative phase followed. This phase included a “category analysis of recordings of teachers discussing and comparing the language proficiency demonstrated in video performances to check that the meta-language used by practitioners was adequately reflected” (CoE 2001: 217). This was followed by 32 workshops with teachers who sorted and judged the descriptors compiled in the intuitive phase. The follow-up quantitative phase used a Rasch rating scale model to statistically scale the selected descriptor items. In the interpretation phase, cut-off points for the final compilation of the Common Reference Levels were produced and specifications for the illustrative scales, presented in chapters 4 and 5 of the CEFR, were drawn up. It is to be noted that the skill of writing was not the focus within that project (CoE 2001: 220). Therefore, the descriptors for writing are still somewhat underdeveloped in the CEFR and not fully validated. Regarding the descriptor formulation, the CEFR (2001: 205ff.) specifies several criteria that descriptors should meet. They should be (1) worded positively, when levels of proficiency should serve as objectives, (2) definite, in that they describe a concrete task, or a concrete degree of skill involved in carrying out a task, (3) clear, meaning that descriptors should be transparent without use of jargon, (4) brief, since teachers tend to prefer short descriptors of approximately 25 words and (5) independent, so that they might be used as checklists or for questionnaires (CoE 2001: 207). Despite the careful and neat validation process, the question remains as to whether a) the teachers from Switzerland who were involved in the validation process, are representative of all the other language professionals who are supposed to use the CEFR, b) the target learner group is representative of the other learners for whom the descriptors were spelled out and c) a post-hoc item scaling is appropriate for producing a scale of progression. In my view, a sound theoretical basis from which linguistic progression can be deduced would be more valid than the operational model that the CEFR presents.

²⁰ Intuitive, qualitative, quantitative phases and their methodological steps are explained in the CEFR (2001) from page 208 onwards. I will refrain from going into detail on those for the sake of comprehensibility in this thesis but suggest the interested reader should read up on the methods in the CEFR appendix A and annotated bibliography.

What resulted from those considerations and scaling is, along with other more detailed scales, the global scale of Common Reference Levels presented in Figure 3 below:

Proficient User	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent User	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, leisure, etc. Can deal with most situations likely to arise whilst traveling 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.
Basic User	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate needs.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Figure 3: Global Scale of the CEFR, taken from CoE (2001: 24)

This scale is intended to serve as a point of orientation for language practitioners (CoE 2001: 24) and consists of illustrative ‘can-do’ statements that purport how a language learner might exploit strategies to act within certain communicative

activities, in which s/he draws upon (communicative language) competences.²¹ It can be seen that the communicative activities broaden from basic activities at A1 level to more abstract operations at C2 level. At the A1 level, the learner can mainly act in everyday situations and provide information of immediate approximation to themselves. At the C2 level more abstract situations such as using and reconstructing several sources are described. The specification of communicative activities, strategies and competences is provided in the horizontal dimension of the CEFR. The horizontal dimension outlined in chapter 2 and described in chapter 4 and 5 of the CEFR is about language use in general and the language user's competences in a taxonomic form. The CEFR suggests reading the taxonomy as intertwined with the action-oriented approach. It is posited that this should help the reader to gain a deeper understanding of why proficiency is determined in terms of performance in communicative activities with the help of strategies and competences. This action-oriented approach is described by the CoE (2001: 9) authors as follows:

Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of **competences**, both **general** and **communicative language competences**. They draw on the competences at their disposal in various contexts under various **conditions** and under various **constraints** to engage in **language activities** involving **language processes** to produce and/or receive **texts** in relation to **themes** in specific **domains**, activating those **strategies** which seem most appropriate for carrying out the **tasks** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences. (bold print in original)

The words in bold print are generic categories that form the core of the CEFR's taxonomy. They are conceptualized as interwoven with language use and learning, as well as teaching and assessment, but can be divided into several sub-categories relatable to specific needs of language professionals (CoE 2001: 10). The term *competence* itself is not explicitly defined in an operational manner in the CEFR. Morrow (2004: 15) argues that the CEFR treats *competences* from a

²¹ Weinert (2001: 2433) defines competences as referring to "combinations of cognitive, motivational, moral, and social skills available to (or potentially learnable by) a person that underlie the successful mastery through appropriate understanding and actions of a range of demands, tasks, problems, and goals". The term competence will be discussed later in this chapter.

global, plurilingual and pluricultural point of view in which the development of learners' individual competences can only be partial. Morrow further infers that this means that each speaker develops *unique individual competences* that cannot be compared to those competences of a native-speaker. This, in turn, would be an argument against the universality in the CEFR.

For this thesis, I consider the categories 'general competence, communicative languages competences, activities and processes' to be of major importance. In my view, these categories provide a basis for linking Processability Theory (Pienemann 1998, 2005) to the CEFR. In this endeavor, an empirically sound theoretical framework which predicts universal processing procedures that result in the production of morpho-syntactic features, can complement the CEFR. The term competence in general needs to be explained in more detail in order to understand its component parts. I will therefore describe only these categories in more detail, and revisit communicative competences in chapters 2.4.1 and 3.3.2.1. The Figure below, based on Steininger (2015: 67), provides an overview of the different competences described in the CEFR:

General Competences
<ul style="list-style-type: none"> ▪ Declarative knowledge (<i>savoir</i>) <ul style="list-style-type: none"> -world knowledge -sociocultural knowledge -intercultural awareness ▪ Skills and procedural knowledge (<i>savoir-faire</i>) <ul style="list-style-type: none"> -vocational know-how -intercultural skills ▪ Existential competences (<i>savoir-être</i>) <ul style="list-style-type: none"> -attitudes -motivation -values -beliefs -cognitive style -personality traits ▪ Learning (<i>savoir-apprendre</i>) <ul style="list-style-type: none"> -beliefs -language and communication awareness -general phonetic awareness and skills -learning strategies -heuristic skills

Figure 4: General competences in the CEFR, adapted and modified from

The term competence is defined as “[...] the sum of knowledge, skills and characteristics, which allow a person to perform actions” (CoE 2001: 9)²². Competences are divided into “general competences” and “communicative competences”. General competences refer to those competences, which are not language-specific, but can be employed for any kind of action a person wants to carry out. In this context, the CEFR makes recourse to Byram’s (1997) model of Intercultural Communicative Competence (ICC).²³ In their concept, the authors of the CEFR use *savoir* to refer to declarative knowledge on several levels, such as world knowledge, sociocultural and intercultural knowledge. *Savoir-faire* (CoE 2001: 104) refers to skills and know-how, such as social or vocational skills. Existential competence is linked to Byram’s *savior-être* and includes values and beliefs, attitudes and motivations of learners (CoE 2001: 105). *Savoir-apprendre* encompasses the ability to learn and includes, amongst others, language and communication awareness and general phonetic skills (CoE 2001: 106). The CEFR authors extend Byram’s categories by including *study and heuristic skills* that refer to making “effective use of the learning opportunities” (CoE 2001: 106) and the ability to “bring new competences to bear” (CoE 2001: 107). A subcategory of general competence comprises communicative competences.

The term *communicative competences* refers to competences which “empower a person to act using specifically linguistic means” (CoE 2001: 9). They are conceptualized as internal representations and mechanisms that manifest themselves in observable behavior of a social agent and that can be transformed and altered through learning processes (CoE 2001: 14). Communicative language competences are divided into linguistic, sociolinguistic and pragmatic competences, as can be seen in Figure 5:

²² The authors of the CEFR do not explicitly state the sources that contributed to their definition of competence.

²³ Byram (1997) put forward an influential multidimensional model of Intercultural Communicative Competence that consists of several types of knowledge. There seems to be another internal contradiction here, because although the authors of the CEFR claim that they do not favor any particular theory, they clearly put Byram’s model at the center of their conceptualization of competences.

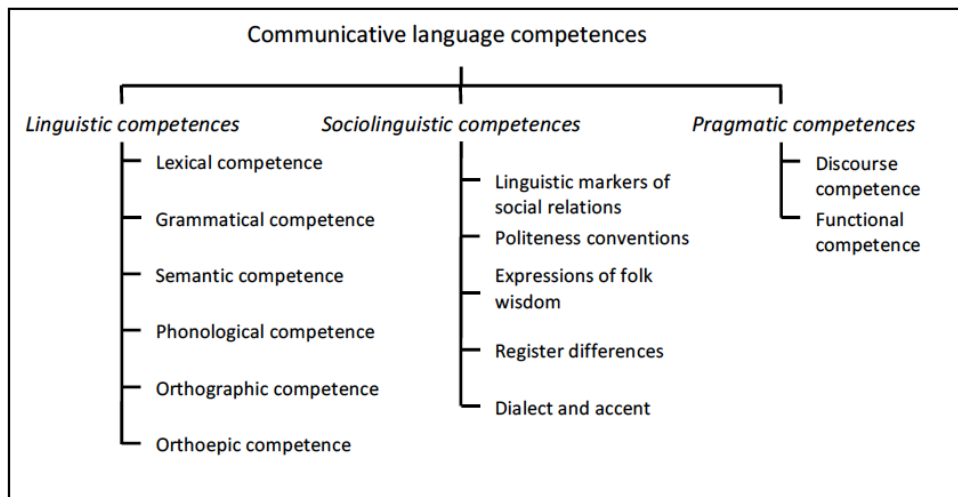


Figure 5: Communicative Language Competences in the CEFR, taken from Green (2012: 20)

Linguistic competences involve “[...] lexical, phonological, syntactical knowledge and skills and other dimensions of the language system [...]” (CoE 2001: 13), that include range and quality of knowledge, as well as the cognitive organization of knowledge storage, and its accessibility. The CEFR authors argue that linguistic competence covers both declarative and procedural knowledge and highlight the various dimensions of variability of this knowledge in different learners (CoE 2001: 13). Sociolinguistic competences are mainly conceptualized in terms of sociocultural conditions and conventions that operate when language users get in touch with each other. The authors state that even though participants in communicative situations might not be aware of conventions, such as rules of politeness, or norms that affect relations between generations, the sociolinguistic component affects all language use (CoE 2001: 13). Pragmatic competences become observable in the functional use of linguistic means in interactive exchanges, in which the language user draws upon discourse-pragmatic features such as cohesion and coherence or uses forms of irony and parody. These competences are again strongly influenced by the cultural environment of the respective user (CoE 2001: 13). According to North (2014: 17), the notion of pragmatic competence is based on Chomsky (1980: 224), and sociolinguistic competence is informed by Canale & Swain (1980).²⁴ North (2014: 17) further observes that the conceptualization of discourse and functional

²⁴ These concepts will be explained in more detail in chapter 3.3.2.

competence as subdivisions of pragmatic competence outlined in the CEFR bear traces of Bachmann's model (1990) of competence that uses textual and illocutionary competence.²⁵

Language activities are regarded as situations in which the users' competences described above are called upon. These activities involve the skills reception, production, interaction and mediation.²⁶ Productive and receptive activities are viewed as primary for engaging in conversation. This was determined by the needs analyses by Richterich (1973) and spelled out in the t-level (Van Ek 1975). For examples of receptive activities, the authors point to the understanding of course content or consulting of textbooks etc. (CoE 2001: 14).

In the context of this thesis, the CEFR's description of language processes is particularly interesting as it is here that relations to Pienemann's theoretical framework can be established. These interfaces will be described in more detail in chapter 3.3.3.3. In the CEFR, the notions of language processes "[...] refer to the chain of events, neurological and physiological, involved in the production and reception of speech and writing" (North 2014: xxxv). In chapter 4 of the Framework, the authors maintain that processes are viewed as communicative processes and specify the user's actions involved in those processes. In terms of production, the speaker is required to "*plan and organise a message (cognitive skills); formulate a linguistic utterance (linguistic skills); articulate the utterance (phonetic skills)*" (CoE 2001: 90, italics in original). I assume that the processes involved in oral production can be linked to those described in Levelt's (1989) Blueprint for the Speaker, which is a yardstick in PT (see chapter 2.2.1.1 for more details).

After having outlined the CEFR's action-oriented approach that views a language user as a social agent who operates within various communicative situations, I now turn to its concept of communicative competence with special

²⁵ Bachmann (1990) developed a model of competences which will be described in more detail in chapter 3.3.2.

²⁶ North (2014: 18) explains that the CEFR draws upon the 4 skills-model from Lado (1961) and extends it by including mediation as well as spelling out unique descriptors for each of these skills in several communicative activities. Lado (1961) argues that language use involves mainly 4 skills, namely listening, speaking, reading and writing.

focus on production and linguistic range. Those notions are useful to determine potential interfaces between the CEFR and PT.

2.1.4 Communicative Competence – Language Production and Processes

According to the CEFR, *communicative competences* entail the ability to exploit strategies for interaction in communicative activities, so that the user realizes his/her communicative intentions (CoE 2001: 57/108). The term *strategies* refers to ways in which language users activate skills to engage in communication in different contexts. These strategies might be seen as the application of meta-cognitive principles involved in message formulation (CoE 2001: 57). Meta-cognitive principles in turn, refer to the skills described above and encompass the *processes* outlined in section 2.1.6 and 3.3.3.3. As regards language production, the term *skill* would refer to the planning and organization of a message. For the oral production of language, which is part of productive activities, the CEFR presents the following set of descriptors:

	Overall Oral Production
C2	Can produce clear, smoothly flowing well-structured speech with an effective, logical structure which helps the recipient to notice and remember significant points.
C1	Can give clear, detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
B2	Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail.
	Can give clear, detailed descriptions and presentations on a wide range of subjects, related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples.
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points.
A2	Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes etc. as a short series of simple phrases and sentences linked into a list.
A1	Can produce simple mainly isolated phrased about people and places.

Figure 6: Overall Oral Production Grid, taken from CoE (2001: 58)

The Overall Oral Production grid is concerned with oral text production directed towards any audience of listeners. Activities in this context include, e.g. speaking spontaneously, seeking public address²⁷ to, for example, gain information or instructions and speaking from notes (CoE 2001: 58). The descriptors at the A1 level for overall oral production mainly describe the restricted repertoire of language users that might relate to the production of formulaic sequences, as indicated by the term 'isolated'. In this context, the term *formulae* denotes information about people or places. At the A2 level, the authors present a specification of 'places and people' in that the descriptors are concerned with living or working conditions and daily routines.

One could argue from the descriptors that although chunks may still play a role in the learner language, users should be able to produce a series of simple phrases and sentences. Fluency²⁸ comes into play at level B1, in which users are supposed to be able to sustain straight-forward descriptions of different subjects. The B2 level seems to be more concerned with the style of the output that a learner produces: the learner presents descriptions, which are "presented systematically" and mediation activities, such as highlighting important points, are "employed successfully". The same seems to be the case for the C1 level, in which the subjects that language users can deal with become more complex. The user is also expected to round those subjects off with a conclusion. At the C2 level, fluency is taken up again and combined with stylistic descriptors, as the user can produce "clear, smoothly-flowing well-structured speech" (CoE 2001: 58). Here, it becomes apparent that not all language features are equally distributed across each level. This might reflect the cumulative nature of the language learning process. Pienemann; Johnston and Brindley (1988) raise this issue in their criticism of proficiency rating scales that aim to assess language proficiency. Hulstijn (2007: 663) observes that the notion of language proficiency adopted in the CEFR rests on two loosely intertwined pillars, namely quality and quantity, which, simply put, translates into *what* and *how well* a learner can use

²⁷ Note the connection to the Threshold level (Van Ek 1975) here that specified communicative situations. See chapter 2.1.1 for more information.

²⁸ One could argue that if a learner has the ability to describe different subjects in a straight-forward manner – other than to rely on chunks - this is the reason for being perceived as more fluent in the language.

the target language at a given point in time. These two dimensions seem to be mixed as is apparent in the descriptors above.

The CEFR presents more illustrative scales in the area of oral production. In the descriptors, these include: sustained monologue; describing experience and the putting of a case (for example in a debate), public announcements and addressing audiences (CoE 2001: 59ff.). The speaking skill is further represented in scales for spoken interaction. The scale presents communicative situations such as understanding a native speaker interlocutor, informal discussion and conversation (see CoE 2001: 73ff.). These scales are more elaborate than the scale for overall production (see CoE 2001: 73).

I will not go into detail about the spoken interaction scale and the spoken production scale, because it is not relevant for the purpose of relating PT and the CEFR. PT focuses on the mental processes involved in the production of language and does not aim to explain situational dependencies of language processes. However, Nicholas & Wigglesworth (in prep.) show that the modular approach in PT has the power to also be aligned to pragmatic language use.

2.1.5 Linguistic Competences in the CEFR

The authors of the CEFR view linguistic competences as part of communicative language competences, which are used “to realise communicative intentions” (CoE 2001: 108). Linguistic competences are subdivided into lexical, grammatical, semantic, phonological, orthographic and orthoepic competence (CoE 2001: 109). The main components of linguistic competence are defined as “[...] the knowledge of, and the ability to use, the formal resources from which well-formed, meaningful messages may be assembled and formulated” (CoE 2001: 109). The authors argue that their definition lies outside the approaches adopted by traditional models to describe linguistic competences. In my view, the approach taken in the CEFR to be brief in order to remain user-friendly (see CoE 2001: 24), takes its toll here, as 1) traditional models of description are not defined in the document and 2) they are not further specified. Rather, the authors refer to section 4.2 of the CEFR as their adopted approach to linguistic

competences. Section 4.2 describes communication themes that comprise the topics of conversation. In particular, these include communication acts in the sense of the Threshold level (CoE 1990). Communication themes may relate to personal identification, house and home, environment, travel, etc. (CoE 2001: 52). Thus, linguistic competences should be seen in relation to the communicative situations outlined above. The scale that is supposed to capture these communicative themes is termed *linguistic range*. Linguistic range can be seen as the umbrella scale for the grids on lexical, grammatical, semantic, etc. competence. The scale for linguistic range is presented in Figure 7 below.

General Linguistic Range	
C2	Can exploit a comprehensive and reliable mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity...No signs of having to restrict what he/she wants to say.
C1	Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.
B2	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.
	Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so.
B1	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.
	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies, and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times.
A2	Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words.
	Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information. Can use basic sentence patterns and communicate with memorized phrases, groups of a few words and formulae about themselves and other people, what they do, places, possessions etc. Has a limited repertoire of short memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations.
A1	Has a very basic range of simple expressions about personal details and needs of a concrete type.

Figure 7: CEFR Scale for General Linguistic Range, taken from CoE (2001: 110)

As a logical consequence of its alignment to the themes described in chapter 4.2 of the CEFR, the descriptors for general *linguistic range* encompass concrete situations and topics that the language user is hypothesized to encounter. This can be seen at A2.1 level, which includes ‘personal details, daily routines, wants and needs’ or level B1.1, which contains ‘topics such as family, hobbies and interests, work, travel and current events’ (CoE 2001: 110). It is noticeable that the themes at C-levels are not as distinguished as those at lower levels. This might be due to the unpredictability of situations that language users can find themselves in at those higher levels. Additionally, one can find rather qualitative descriptors that give hints as to the linguistic ability needed to perform in those situations. Those are, for example, of lexical nature. At the A2.1 level, this applies to: ‘everyday expressions’ or at the B1 level ‘sufficient vocabulary’. Only some descriptors can be found that are of grammatical nature, such as the descriptors at the A1 level ‘basic range of simple expressions’, at the A2.1 level ‘memorised phrases and formulae’ or at level B2.1 ‘complex sentence forms’. To me, it is difficult to single out more descriptors that might be informed by a grammatical component. The fact that the descriptors mix general, holistic statements about proficiency with bold behavioral objectives constitutes Green’s (2012) main point of criticism. It seems that the CEFR descriptors encompass a ‘constrained-based’ view on language performance. An overview of the constrained-based descriptors is provided below:

A2.1	‘frequent breakdowns and misunderstandings occur in non-routine situations’
A2.2	‘he/she will generally have to compromise the message and search for words’
B1.1	‘lexical limitations cause repetition and even difficulty with formulation at times’

Interestingly, the authors of the CEFR do not mention constraints on the B1.2 level. From level B2 onwards, there are descriptors for signs of struggle with the target language that the language user does not show:

B2.1	‘without much conspicuous searching for words’
------	------------------------------------------------

B2.2	'without much sign of having to restrict what he/she wants to say'
------	--------------------------------------------------------------------

Whereas the B2 level shows signs of restriction, C1 level does not cover restrictions anymore, but rather:

C1	'without having to restrict what he/she wants to say'
C2	No signs of having to restrict what he/she wants to say'

The reader might find it hard to distinguish nuances of linguistic ability based on those constraint-based can-do statements. North (2014: 26) argues that the CEFR deliberately uses a normative style of descriptor formulation,²⁹ which “[...] assumes assessors have internalized a clear understanding of the standard for the level concerned, around which they just norm-reference” and explains this choice by being informed by the Cambridge ESOL scales of the 1980s for assessing speaking and writing skills. However, based on the descriptors presented above, a clear distinction between the levels and/or descriptors is hard to find.

Additionally, the subcomponents (lexical, grammatical, semantic, etc.) are not equally distributed at each level. If they were, it would suggest an implicational relationship or a clear progression across the levels. It is also not possible to link each of the sub-scales to the broader linguistic range scale. However, North (2014: 101) argues that the CEFR provides a descriptive apparatus of scales mirroring that users “can generally do more things at higher levels, since, because progress can be lateral as well as vertical, competences learned in one context can be applied to another”, because progress in language learning is not linear (North 2014: 101). North also (2014: 102) maintains that being B1 in one context does not mean that a user can be considered being B1 in all other contexts. Rather, the levels are supposed to describe that someone at level B2 is better than someone at level B1, but not yet a level C1 (North 2014: 103). In my view, this is all that a reference tool might be able to aim at, although

²⁹ North (2014: 26) describes that another way of descriptor formulation would assume a systematic approach. He discards a systematic approach by arguing that it was too repetitive, that it heavily relied on alternating qualifiers (as no, some or many) and that it could not be used for mathematically scaling descriptors as it would reduce “[...] differences to mere semantic variation”.

I reckon that when used as a reference for assessment, this fact poses a severe problem to the alignment, comparability and administration of language tests.

Further, the CEFR authors argue that any language system is highly complex and dynamic; that it is under continuous evolution, so that it can never be fully mastered by language users (CoE 2001: 109). This line of thought originates from the holistic, action-oriented view of language proficiency, that conceptualizes language users to operate dependent on cultural conventions. In this context, North (2014: 23) argues that “the possibility of one universal model of description for all languages has been denied. Recent work on linguistic universals has not yet produced results which can be used directly to facilitate language learning, teaching and assessment.” North (2014: 23) further claims that the reason that insights from SLA research have not been incorporated in the CEFR descriptors is, because in-depth, large-scale longitudinal studies of SLA were “[...] not available in the mid-1990s. SLA researchers have had great difficulties in establishing even the simplest fixed orders of acquisition of grammatical structures.” Although North (1997) discusses the state of the art of SLA research on linguistic universals, he seems to be unaware of Ellis’ (1994: 21) claim that „there is now general acceptance in the SLA community that the acquisition of an L2 grammar [...] occurs in stages.“ This research tradition goes back to the 1970s as can be seen, inter alia, in the studies by Felix (1984), Wode (1976), Clahsen (1980), Meisel et al. (1981). Also, Pienemann’s psychologically and typologically plausible account to explaining developmental schedules, i.e. Processability Theory, was not considered. It is reasonable that the CEFR does not favor a particular linguistic framework, considering its overall holistic, action-oriented approach. In line with this, Lantolf & Frawley (1988) reflect that probably no theoretical framework will ever be able to capture the complexity of the language system with one single account. Hulstijn (1985: 277) therefore argues that instead of waiting for the ‘possibly impossible’ development of a comprehensive theory of language proficiency, language professionals need to “work with taxonomies that seem to make sense even if they cannot be fully supported by a theoretical description”. North & Schneider (1998: 242) follow this line of reasoning in arguing that a common reference framework needs to

work with the taxonomies, even though they might cause tensions between theoretical models and operational models developed by practitioners.

Considering that the CEFR aims at providing most complete although not exhaustive (see CoE 2001: 1) reference points based on a literature review of current language research (see Harsch 2006), for language professionals to reflect on their practice, I argue that the strong theoretical and empirical tradition within the Processability Framework might add to the descriptive and empirical basis of the document and its taxonomies. Although, Pienemann opts for theoretical parsimony in the development of PT (Pienemann 2005b: 66), he Processability account takes a modular approach to explain developmental schedules in L2 acquisition. Therefore, it might inform the CEFR in terms of grammatical development. Grammatical competence in the CEFR will be explained in the following chapter.

2.1.6 Grammatical Competence in the CEFR

Grammatical Competence forms a subcomponent of linguistic competences in the CEFR. Linguistic competences are broken down into, inter alia, grammatical competence. It is this subcomponent that Processability Theory might contribute to, both to the theoretical basis, as well as the empirical validation of the CEFR's descriptive machinery.

As outlined above, grammatical competence is defined as “the knowledge of, and ability to use, the grammatical resources of a language” (CoE 2001: 112). Grammar in this regard is explained as “the set of principles governing the assembly of elements into meaningful labeled and bracketed strings (sentences)” (CoE 2001: 113). The CEFR authors seem to ascribe a prime value to grammatical accuracy as they maintain that grammatical competence is the ability to “produce and recognize well-formed phrases and sentences in accordance with these [the assembly into sentences] principles (as opposed to memorizing and reproducing them as fixed formulae)” (CoE 2001: 113, addition by KH). What can be inferred from this, is that grammatical competence should not equal formulaic language use. The CEFR provides a description of formal grammatical elements

(morphs, affixes, etc.), categories (number, case, gender, etc.), classes (conjugations, declensions, etc.), structures (phrases, clauses, sentences, etc.), descriptive processes (affixation, suppletion, gradation, etc.), and relations (government, valency, etc.) (CoE 2001: 113). Additionally, the CEFR makes a clear distinction between morphology and syntax. Morphology is regarded as the internal organization of words into morphemes, with roots and stems as well as affixes (CoE 2001: 114). The document also includes brief comments on word formation processes and morphophonology. Syntax is defined as the “[...] organization of words into sentences in terms of categories, elements, classes, structures, processes and relations involved, often represented in the form of a set of rules” (CoE 2001: 115). Here, it is stated that mature language users mainly rely on the unconscious organization of words into sentences, which is characterized by a certain amount of complexity. The organization of sentences is regarded as central to communicative competence, as it is a means to convey meaning (CoE 2001: 115). The authors provide a scale for grammatical accuracy that they suggest should be read in connection with the one provided for general linguistic range.

Grammatical Accuracy	
C2	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged, (e.g. in forward planning, in monitoring others’ reactions).
C1	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.
B2	Good grammatical control; occasional ‘slips’ or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.
	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.
B1	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is usually clear what he/she is trying to express.
	Uses reasonably accurately a repertoire of frequently used ‘routines’ and patterns associated with more predictable situations.
A2	Uses some simple structures correctly, but still systematically makes basic mistakes – for example, tends to mix up tenses and forgets to mark agreement; nevertheless it is usually clear what he/she is trying to say.
A1	Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.

Figure 8: Grammatical Accuracy Scale in CEFR, taken from (CoE 2001: 114)

In a similar fashion to the scale for linguistic range, the lowest level (i.e. A1) of the scale for grammatical accuracy is characterized by learned repertoire and simple grammatical structures. The constraint-based fashion of descriptors (as discussed above) is reflected in level A2, in that a user is said to systematically “make basic mistakes”. These mistakes are specified as errors of tenses and absent agreement marking. At the A2 level however, the user is able to bring his/her communicative intention across. The B1.1 level is said to be characterized by a reasonable accuracy, which is restricted to routines and linguistic patterns of predictable situations. Those predictable situations are extended to familiar contexts at the B1.1 level, in which errors are said to occur based on mother-tongue influence. However, here, as at level A2, it should be clear what the user wants to express. At level B2.1, the user has as much control over his/her grammar so that mistakes would not lead to misunderstandings. A progression from ‘relatively high degree of control’ at B2.1 to ‘good control’ of grammar at B2.2 level is visible. At this level, slips and non-systematic errors are mentioned, which as stated in the descriptors, can often be corrected in retrospect. I assume that the authors are referring to the distinction between learner errors and mistakes. According to Corder (1967), errors are systematic in nature and provide the teacher with insights into the language learning process. James (1998) argues that learner errors cannot be self-corrected as they reflect a ceiling point in the language acquisition process. Mistakes, however, can be self-corrected, as they rather constitute “slips in performance”. This is why mistakes are sometimes referred to as performance errors (Corder 1967). Clahsen, Meisel & Pienemann (1983) take this up and extend this distinction, by integrating a processing perspective on errors. They show that some errors can be attributed to a variational dimension whereas others occur because of developmental readiness. Keßler & Plesser (2011: 112) describe developmental errors as those that the learner produces because s/he is not yet able to perform the underlying psycholinguistic operations required for producing the target-structure. Thus, the learner has to find different solutions to producing the target structure, which may lead to inaccurate grammatical forms. Variational errors are those that the learner produces although he/she should, in principle, have acquired the

underlying mechanisms. This might be due to backsliding effects, or lack of paying attention (Keßler 2006; Keßler, Liebner, Mansouri 2011). For this reason, one could argue that a lot of slips can be considered variational errors.

C1 descriptors state a high degree of grammatical control, with errors being difficult to spot. A qualitative dimension is added at the C2 level. At this level, the user is supposed to have consistent control over grammar of complex language. Control is supposed to remain steady, even when attentional resources are allocated to other factors, such as forward planning or monitoring of others' reactions. If this scale serves as a basis for producing assessment grids, then the user might find it difficult to operationalize descriptors such as "consistent control over complex language". In this context, the question arises whether the authors would assume that 'monitoring of others' reaction' was absent at earlier levels. This reflects the critique by Pienemann et al. (1988), that the levels are not implicationally related. The reason for this lack might be because the authors decided to use a normative approach to spell out the descriptors (see North 2014: 26). To recapitulate, the normative approach was chosen, as the overall aim was to provide objectives as reference points for language professionals (see North 2014: 26). Steininger (2014: 47) criticizes that the CEFR instructs its users to explicitly specify as to which theoretical framework they follow. Yet itself fails to do so, as it seems impossible to exactly trace back which notion/framework influenced which descriptors.

I propose that a view of grammatical competence in language learners – which is the target group described in the CEFR – might integrate a more learner-based approach to grammatical competence. This might be beneficial, in that it does not exclusively rely on accuracy as a point of reference. Considering that a language user should be regarded holistically and as a social agent in the CEFR, who operates under cultural conditions and constraints, a less accuracy-based view is appropriate. Although the authors of the CEFR recognize the multitude of competing theories on the organization of words into sentences (CoE 2001: 113), they argue that it is not the function of a framework to advocate any of those competing theories. Moreover, the authors claim that "[i]t is not considered possible to produce a scale for progression in respect of grammatical structure

which would be applicable across all languages” (CoE 2001: 113, lower case by KH). What the Framework rather aims at, is to encourage language professionals to reflect upon different accounts and the respective consequences of a particular choice. In my view, the description of the user’s competences does not ultimately aim at integrating language learning principles (which are described in chapter 6 of the CEFR). However, the CEFR is aimed at informing language professionals who work with language learners and I thus consider it of primary importance to gear it towards language learners as its audience in all respects. In its current form, however, it is problematic to find well-established language learning principles in the conceptualization of grammatical competence. Consequently, I argue that Processability Theory has the power to inform the CEFR’s view of grammatical competence through a language learner lens. This could be done by integrating grammatical progression, as underpinned by the universal processes that Processability Theory specifies. To determine whether this is possible, is what this thesis sets out to do. A more detailed discussion of interfaces between the CEFR and PT will be given in the theoretical account to bridging scales and stages (chapter 3), as well as the discussion in the empirical part (chapter 4.4.4). I argue that if a combined approach was possible, the scale for grammatical accuracy would need to be relabeled to *Grammatical Range*. The most pressing problem in the scale for grammatical accuracy in the CEFR, is that it covers only grammatical accuracy. For this reason, the relationship between the qualitative and the quantitative dimension in the CEFR is elusive. The following chapter will look at language learning issues raised in the CEFR.

2.1.7 Language Learning in the CEFR

Chapter 6 of the CEFR describes language learning and teaching. It is argued that, whereas “[...] chapters 4 and 5 attempt to set out to what a fully competent user of a language is able to do and what knowledge, skills and attributes make these activities possible” (CoE 2001: 131), chapter 6 elaborates on what learners need to acquire, in order to be able to fully participate in society and exploit what was described in chapters 4 and 5. These chapters are part of the qualitative

dimension. However, although chapter 4, for example, is labeled: “Language use and the language user/learner” (see page 43), based on the quote above, it seems as if that very chapter was rather based on fully competent users of a language. The question that arises is whether the scales that were presented there, i.e. inter alia, the scale for grammatical competence, are then conceptualized for fully competent language users, or for language learners. I reason that it can be maintained that the CEFR is intended as a reference tool for language professionals to use, mostly, for language learning contexts or assessment of language learners. The needs analysis by, e.g., Richterich (1973) was used as a means to develop descriptors for the Threshold level (Van Ek 1975). These two go into the development of the CEFR. Thus, the CEFR should be based on the needs of learners and not fully competent language users. To me, there seem to be a few inconsistencies in how the language user/language learner terms are used throughout the CEFR. Those inconsistencies might lead to the narrow concept of grammatical accuracy as the only scale that represents grammatical competence that I do not consider geared towards language learners (see chapter 3.3.1 for more details).

However, chapter 6 postulates that the steps learners need to learn in order to participate in communicative events, are to acquire the competences laid out in chapter 5. The ability to use these competences in activities and the ability to put the strategies to use that are necessary to exploit the competences are described in chapter 4 (CoE 2001: 131). As the CEFR’s aim is to neither highlight any specific theory,³⁰ nor to advocate a specific route in terms of learning, teaching or assessment, the concept of language learning itself remains rather vague. The authors hint at a definition by describing that the development and improvement of strategies enables “[...] an individual to mobilise his or her own competences in order to implement and possibly improve or extend them [...]” (CoE 2001: 137). This definition seems to be somehow informed by Krashen’s (1981) acquisition-learning distinction³¹. Following Krashen, they

³⁰ However, it is to be noted here that Byram’s model of ICC seems to be highlighted in the conceptualization of competences.

³¹ Krashen (1981: 1ff.) argues that L2 acquisition is similar to children acquiring a first language, whereas learning takes place in formal settings that are aided by explicit rule teaching and error correction.

describe that formal language *learning* might be seen as a process “[...] whereby ability is gained as a result of a planned process, especially by formal study in an institutional setting” (CoE 2001: 139). In their concept of learning, the authors of the CEFR also integrate a Chomskian approach in stating that *learning* furthermore encompasses interpretations of the language of non-native speakers in terms of a universal grammar, and whereas *acquisition* was rather natural, informal language acquisition (see CoE 2001: 139). In their chapter on how learners learn a language, the authors of the CEFR discuss several scenarios that seem to be informed by Krashen’s (1985) input hypothesis,³² as they describe that for some “[...] the most important thing a teacher can do is provide the richest possible linguistic environment in which learning can take place without formal teaching.” (CoE 2001: 139). When further describing different approaches to language learning, the authors seem to draw upon Pienemann’s (1985) Teachability Hypothesis.³³ They discuss that mainstream education providers might want to follow an eclectic approach to designing language scenarios. In that context, they use the phrase “[...] recognizing that learners do not necessarily learn what teachers teach and that they require substantial contextualised and intelligible input [...]” (CoE 2001: 140). Pienemann (1985) uses this phrase to illustrate that, in terms of morpho-syntax, learners only acquire what they are able to process despite the teachers’ input and objective. However, again the sources are not stated explicitly.

What the document also describes, is variation amongst learners in terms of age, learning types and backgrounds, which should be considered with regard to aims of course designs. The following chapter focuses on the learning and teaching of linguistic competences.

³² Krashen (1981) hypothesized that all it takes for language acquisition is rich input in an i+1 manner. I+1 means that the input should be one level above the learners’ current competences.

³³ The Teachability Hypothesis will be explained in chapter 2.2.6 in more detail.

2.1.7.1 The Learning and Teaching of Linguistic Competences

On page 149, the authors of the CEFR again highlight that linguistic competences are central and indispensable to language learning. For the learning of grammatical competence, the authors seem to advocate a step-by-step presentation of linguistic material in terms of inherent complexity; i.e. from single clauses “[...] with its constituent phrases represented by single words [...]”, to more complex mult clause sentences (CoE 2011: 151). The use of formulae as a means for complex material at early stages of learning is also suggested. Materials could include fixed frames for lexical insertions, or as learnt words of a song (CoE 2011: 151). It is stated that the general domain for grammatical description should take place at the sentence level, so that inter-sentential relations (e.g. anaphora, pro-verb use and sentence adverbs) can be regarded as belonging to linguistic rather than to pragmatic competence. As further ordering principles for grammatical instruction, the following aspects are given:

- a) The communicative field of grammatical categories and their role as exponents of general notions
- b) Contrastive factors, for e.g. word order problems
- c) Authentic discourse with regard to grammatical difficulty for providing learning opportunities
- d) The natural order of first language acquisition (CoE 2001: 151).

For formal instruction, the authors list a number of techniques that cover the aspects above, ranging from inductive exposure to authentic texts, to more explicit explanations and formal exercises. Learner errors are also considered in the qualitative dimension of the CEFR. They will be discussed in the following chapter.

2.1.7.2 The Role of Learner Errors in the CEFR

When discussing learner errors, the CEFR takes up on Selinker’s (1972) concept of *interlanguage*. Interlanguage, from the viewpoint of the authors of the CEFR, is a simplified version of the target competence. This competence concurs with

the learner's performance when he/she produces errors (CoE 2001: 155). Mistakes, in contrast, are described as instances in which the learner "does not bring his/her competences properly into action" (CoE 2001: 155). The authors argue that mistakes happen in all language use, even in that of native speakers. Errors, according to the Reference Framework, are either evidence of a failure to learn or inefficient teaching or, on a more positive note, evidence of the learner's willingness to take risks in communicative situations and a result of the developing interlanguage (CoE 2001: 155).

The descriptive nature of the CEFR yet again becomes apparent in its elaboration of learner errors. The intention of the CEFR authors is to most holistically describe aspects entangled with language proficiency, informed by learners' needs in relation to, *inter alia*, communicative competences. Thus, a clear opinion about how to deal with errors or which approach to favor is not the target of the CEFR.

2.1.7.3 Assessment in the CEFR

Martyniuk (2010: viii) states that the CEFR is most influential in the domain of standardized high-stakes and large-scale assessment as there is "[...] growing interest world-wide in establishing comparability between assessment tools and external standards [...]". The CEFR is neither intended as an assessment tool, nor as a standard to describing language proficiency. It is rather a framework of reference. However, in the notes for the user it says "[...] a set of reference levels as a calibrating instrument is particularly welcomed by practitioners [...] who find it advantageous to work with stable, accepted standards of measurement and format" (CoE 2001: 7). Many assessment providers have thus set out to use the CEFR as fixed standards and to adopt the CEFR levels for their particular use; i.e. for the specification of assessment grids.

Chapter 9 of the CEFR is concerned with assessment issues. The authors state that the term assessment is used "[...] in the sense of the assessment of the proficiency of the language user" (CoE 2011: 177). They maintain that their idea of assessment is distinct from all broader concepts, such as evaluation, but that

all assessment is part of evaluation (CoE 2001: 177). In the first part of the chapter, a number of assessment quality criteria are mentioned: validity, reliability and feasibility. According to the authors, “[v]alidity is the concept with which the CEFR is mainly concerned” (CoE 2001: 177). In stating that “[a] test or assessment instrument can be said to have validity to the degree that it can be demonstrated that what is actually assessed (the construct) is what, in the context concerned, should be assessed [...]” (CoE 2001: 177), special focus seems to be attributed to construct validity. When engaging with the term reliability, the authors argue that the accuracy of decisions made in relation to a standard is more important than the actual reliability of tests, because, in their view, the accuracy of decisions depends on the validity of the standard with which the CEFR is mainly concerned (CoE 2001: 177).

Furthermore, the authors suggest three main ways in which the CEFR might be used for assessment (CoE 2001: 178):

1. *For the specification of the content of tests and examinations: what is assessed*
2. *For stating the criteria to determine the attainment of a learning objective: how performance is interpreted*
3. *For describing the levels of proficiency in existing tests and examinations thus enabling comparisons to be made across different systems of qualifications: how comparisons can be made*

For content specifications (1), the user is directed to consult the chapter ‘communicative language activities’ as a source for task-specifications in assessment (CoE 2001: 178). As an example of spoken language production, an activity might be to have a learner describe his/her own academic field (see CoE 2001: 179). The CEFR highlights that the user who seeks to gain information about assessment, needs to be aware of the distinction between descriptors for aspects of competences (such as given in the CEFR chapter 5) and descriptors for language activities (such as those presented in chapter 4). The authors prefer the former descriptors because “[...] assessment should not be primarily concerned with any one particular performance, but rather seek to judge the generalisable competences evidenced by that performance” (CoE 2001: 180).

Further details for content specifications given in chapter 9 comprise a list of *assessment* options for *assessing* activities and competences. Options for

activities include the use of checklists for self-assessment, or the use of grids for continuous or summative assessment (see CoE 2001: 180). For *competence assessment*, the CEFR distinguishes between *self-/teacher-assessment* and *performance assessment*. As regards self-/teacher assessment, the authors highlight that the CEFR descriptors are phrased in a positive way, which is in contrast to many existing scales. In this context, they point out that existing scales “[...] are often negatively worded at lower levels and norm-referenced around the middle of the scale” (CoE 2001: 181).

Performance assessment is to be carried out by stakeholders.³⁴ The user is again directed to the aspects of competences, as described in the CEFR’s chapter 5. For this type of assessment, the CEFR suggests using scales, checklists or grids (CoE 2001: 181). Scales can be subdivided into proficiency scales and examination rating scales. Proficiency scales allow for a more fine-grained distinction between categories, such as the distinction between B+ and B-. An examination rating scale is designed as a cut-off scale that helps to assess whether the learner performance represents a pass or a fail for a specific category (see CoE 2011: 182).

The second aspect focuses on *how performance is interpreted*. The CEFR provides a description of different types of assessment, such as achievement versus proficiency assessment, formative versus summative assessment or performance versus knowledge assessment. It is important to note that the CEFR itself is not a testing tool. Instead, its focus is to describe different scenarios for assessment that a stakeholder can consult. Also, the CEFR does not favor one type of assessment over the other but suggests situations in which one type of assessment might be preferable to another one.

In addressing *how comparisons can be made*, the authors again state that chapter 4 and 5 might be consulted. The authors highlight the aspect of feasibility and state that any practical assessment system should reduce the number of possible categories to a feasible one and that the CEFR might be consulted as a

³⁴ In my view, some test providers seem to confuse this distinction between self-/teacher assessment and performance assessment, because they often seem to use the positive wording of their assessment instruments as a quality criterion for their tests.

reference tool for this purpose. They exemplify the feasibility aspect by presenting criteria used in the *Cambridge Certificate in Advanced English* (1991) as an example for feasible assessments.

This chapter briefly described the specifications that the CoE presents for assessment in the CEFR. Following the undogmatic notion of the CEFR, assessment issues are simply addressed descriptively, and the document does not recommend one particular type of assessment based on the CEFR. This descriptive basis forms the connection to the following chapter, in which I introduce PT, a psycholinguistic theory of second language acquisition. I assume in this study that the combination of the CEFR and PT is especially valuable for adding to the specifications of grammatical competence, as well as the scale for grammatical accuracy presented in the CEFR. My assumption is based on the idea that PT proposes a learner-friendly developmental path in second language acquisition. I consider it learner-friendly because PT does not use grammatical accuracy as a point of describing and measuring linguistic progression. The CEFR, however, only provides a scale for grammatical accuracy for grammatical competence (see chapter 2.1.6 for more information). In order to substantiate this argument, the next chapter introduces PT in more detail.

2.2 Processability Theory

Processability Theory by Pienemann (e.g. 1998, 2005a) is a psycholinguistic theory to SLA that explains the development of morpho-syntactic structures in second language learners, based on a universal predictable developmental path. It provides “a systematic perspective on some central mechanisms underlying the spontaneous production of interlanguage (IL) speech” (Pienemann 1998: xv) by taking a processing perspective to SLA development. PT was initially designed to explain the developmental problem (see Pienemann 1998) but, more recently, also discusses issues concerned with the logical problem (see Pienemann et al. 2005; Lenzing 2013). The developmental problem focuses on the question as to why learners follow a predictable sequence, in terms of morpho-syntactic development, in the acquisition of their L2. These sequences have been found

by, inter alia, e.g. Wode (1976); Clahsen (1980); Meisel et al. (1981); Pienemann (1981) and are, as Ellis (1994: 21) states, one of the most important findings in SLA research. Pienemann & Lenzen (2015: 161) illustrate the logical problem by asking the following question: “How do learners come to know what they know if their knowledge is not represented in the input?”. The logical problem thus makes recourse to the source of linguistic knowledge.

These two questions are addressed in PT. PT predicts a hierarchy of grammatical structures that language learners acquire cumulatively and successively (Pienemann 2005b: 2). With the Processability account of second language acquisition, Pienemann (1998, 2005a) predicts the acquisition of morpho-syntactic features on the basis of processing procedures. The development of the processing procedures accounts for a universal developmental path that proceeds in stages. The logic behind PT is that second language acquisition can be explained by the architecture of the human language processor (Pienemann 2005b: 3). Pienemann (1998: 4f.) thus explains that for the learner to be able to produce a certain linguistic structure, the necessary processing prerequisites need to be in place. These processing prerequisites and, consequently the structural options available to the learner, are constrained by the language processor. The view of the language processor in PT is largely adopted from Levelt (1989). The formal theory of grammar, Lexical Functional Grammar (Bresnan 2001), is integrated into PT as its second yardstick. Both yardsticks in PT, the Blueprint for the Speaker by Levelt (1989) and Lexical Functional Grammar by Bresnan (2001) will be sketched out in the following chapter.

2.2.1 Yardsticks in Processability Theory

As stated earlier, PT (Pienemann 1998, 2005a, 2015) is a psycholinguistic account of the acquisition of second languages. It explains and predicts a developmental path for specific morpho-syntactic structures, based on the acquisition of processing procedures that are operative in the learner’s mind. The assumption is that the developmental path depends on the architecture of the human

language processor. Therefore, a language learner is able to produce only those linguistic structures that he/she is able to process. The processing procedures develop gradually and successively. The view of sentence production is adopted by Levelt's (1989) Blueprint for the Speaker. The structural correlates to Levelt's language processor are captured by the mapping principles which were formulated in Bresnan's (2001) Lexical Functional Grammar. Levelt's model and Bresnan's formal theory of grammar thus form two important yardsticks in Pienemann's conceptualization of PT. For this reason, both will be outlined in the following parts of this chapter.

2.2.1.1 A Brief Outline of Levelt's Blueprint for the Speaker

Pienemann (1998) integrates Levelt's (1989) Model of Sentence Generation into his theory of second language acquisition in order to achieve psychological plausibility. Levelt adopted notions of Incremental Procedural Grammar as put forward by Kempen & Hoenkamp (1987).³⁵ The core idea of PT is that the developmental path can be explained by the make-up of the human language processor. The Blueprint for the Speaker thus forms the psycholinguistic basis of PT.

Levelt (1989: 1) views speaking as "[...] one of man's most complex skills." He argues that the examination of this complex cognitive skill, as with any cognitive skill, "[...] requires a reasoned dissection of the system into subsystems, or processing components", as well as a "[...] characterization of the representations that are computed by these processors [...]" (Levelt 1989: 1). He developed a model that can account for message generation from the speaker's intention to realize a communicative act to the final version of output. Figure 9 displays Levelt's account of the components and processes involved in message generation.

³⁵ Incremental Procedural Grammar is mainly concerned with sentence assembly during spontaneous speech production (Kempen & Hoenkamp 1987: 202). The core idea of Incremental Procedural Grammar lies in its incremental, left-to-right mode of sentence production that is characterized by constraints on the shape of possible syntactic building procedures and appointment rules (see Kempen & Hoenkamp 1987: 204).

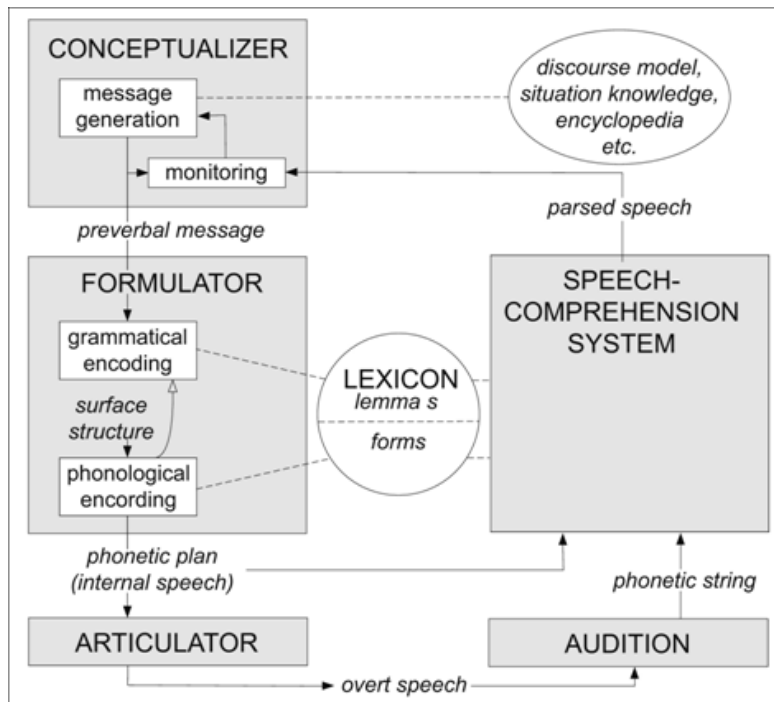


Figure 9: A blueprint for the Speaker, taken from Levelt (1989: 9)

In Figure 9, the boxes represent the different processing components, whereas the circle and ellipse display knowledge stores (see Levelt 1989: 9). For Processability Theory, two components and one knowledge store are of major importance. These are the Conceptualizer, the Formulator and the Lexicon.

In the Conceptualizer, the intention of the speaker's message is generated. Conceptualizing therefore involves, inter alia, conceiving of an intention, selecting relevant information, ordering information for expression, and monitoring the utterance (Levelt 1989: 9). In this module, the pre-verbal message is generated. It does not yet have a linguistic shape but contains the propositional content of the intended message.³⁶ The pre-verbal message is fed into the Formulator which "[...] translates a conceptual structure into a linguistic structure" (Levelt 1989: 11). This process involves two steps, namely grammatical encoding and phonological encoding.

Grammatical encoding is the process of accessing lemma and syntactic building procedures. Lemma³⁷ are located in the knowledge store "lexicon".

³⁶ The process of message generation, according to Levelt, requires more aspects than stated here, such as micro-and macro-planning. Levelt also maintains that declarative and procedural knowledge are needed to conceptualize a message and makes recourse to Baddeley's (1986) concept of Working Memory. For more information see Levelt (1989: 10ff.).

³⁷ Levelt argues that lemma information is of declarative nature (see Levelt 1989: 11).

Lemma contain the meaning of a lexical item, its diacritic features and its syntactic distribution. When the meaning of the lemma and the pre-verbal message match, the syntactic building procedures are activated and noun phrases, propositional phrases, clauses, etc. are built (see Levelt 1989: 11). When the process of grammatical encoding is completed, i.e. all lemmas have been accessed and syntactic building is completed, a surface structure is produced. The surface structure is stored in a syntactic buffer. This is when step two, phonological encoding, takes place. In phonological encoding, the surface structure is translated into an articulatory plan by activating the lexical form contained in the lexicon. Levelt (1989: 12) argues that the lexical form contains information about morphology and its phonology. The lexicon holds, for example, that the word “[...] *dangerous* consists of the root (*danger*) and the suffix (*ous*), that it contains three syllables of which the first one has the accent, and that its first segment is /d/” (Levelt 1989: 12). This articulatory plan still has the shape of an internal representation and is fed into the Articulator. The Articulator translates the phonetic plan into overt speech by activating the physical properties needed to produce overt speech. To do so, it utilizes an Articulatory Buffer for the temporal storage of information.

Apart from these major steps, the Audition module and the Speech Comprehension System serve to monitor and comprehend language use. Since those modules are not of immediate necessity to understand the notions behind PT, they will not be discussed in this chapter.³⁸ Levelt developed his model for monolingual, mature language users (see Levelt 1989: 1). Some attempts have been made to apply this model to bilingual speakers (see DeBot 1992). DeBot (1992) hypothesizes two language formulators for each of the speaker’s languages. The systems are closely related and influence each other. He further assumes that learners, at least in part, draw upon the same procedural and lexical knowledge when speaking one of their languages. Pienemann (1998: 73) argues that Levelt’s model can account for second language learners.

³⁸ Refer to Lenzing (2017) for a detailed and comprehensible account of Levelt’s model of message generation and recent research on its use for conceptualizing the interface between comprehension and production from a Processability perspective.

Pienemann & Keßler (2011: 28) summarize the key features of language production based on Levelt's model that is informed by Kempen & Hoenkamps (1987) IPG as follows:

1. Processing components (such as the Formulator, the Grammatical Encoder and the lexicon) are relatively autonomous specialists which operate largely automatically,
2. Processing is incremental,
3. The output of the processor is linear, while it may not be mapped onto the underlying meaning in a linear way,
4. Grammatical processing has access to a grammatical memory store.

These features in language processing “[...] characterise the processing environment within which the learning of language takes place” (Pienemann 2005: 3).

The first claim can explain the processing speed that underlies language production, because the processing components are restricted to receive and pass on only highly task-specific information. This leads to a gain in processing speed, as the task-specificity allows for unattended information to be processed (Pienemann 2005: 4).³⁹

The notion of incrementality in the second claim describes the ability of the processing components to work on their input without having received the complete set of information. Levelt (1989: 24) adopts the term incrementality from Kempen and Hoenhamp (1982) and explains that the benefit of assuming incremental processing is that “[a]ll components can work in parallel, but they all work on different bits and pieces.” In this way, one component can start working on the incomplete output of another processor without much look-ahead (Pienemann 2005: 5). Incremental processing requires memory stores that are able to process non-linear sentences. This leads to both claims three and four.

Although the processing components produce only linear output, human beings are able to produce sentences that are not in line with the natural order

³⁹ The underlying idea is that the recalling of declarative, attended, information, such as meta-linguistic information, requires more time accessing procedural information (see e.g. Garmann 1990).

of events. As an example of one such sentence, Pienemann (2005: 5), based on Levelt (1989: 138), uses “Before the man rode off, he mounted his horse.” The first event that must have happened is the mounting of the horse, and only then did the man ride off. However, humans are able to produce a sentence in which the second proposition is produced first.⁴⁰ In order to be able to do so, one needs to store propositional information in a memory store, so that the events can be expressed in a non-linear way. Another example of information that needs to be stored in a memory store concerns subject-verb agreement, as displayed in the tree diagram in Figure 10 below:

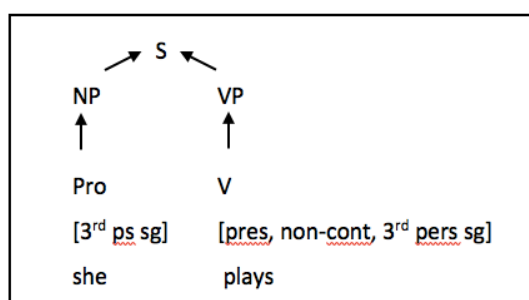


Figure 10: Subject-verb Agreement, example adapted and modified from Pienemann (1998)

In order to unify the information of the pronoun (third person, singular) and the verb (present, non-continuous, third person, singular) in the top sentence node (s-node), information about the diacritic features of the noun and the verb need to be stored in a grammatical memory store; i.e. information on person and number. One such temporal disposal is necessary because of the automatic nature of the processing components that was described in claim one. To recapitulate, automatic processing is inattentive and therefore faster than attentive processing (Pienemann 2005: 4/6).⁴¹

After having laid out the psycholinguistic basis of PT and the key psychological factors in language processing, I will continue with a brief sketch of the second yardstick of Processability Theory; i.e. Lexical Functional Grammar.

⁴⁰ Levelt (1983) calls this the linearization problem.

⁴¹ Please note that due to the limited scope of this thesis, the text above is a very condensed summary of Levelt’s Blueprint for the Speaker and Pienemann’s hypotheses about integrating his model into PT. For more detailed information, see e.g. Pienemann (1998, 2005) and Lenzing (2017).

2.2.1.2 A Brief Sketch of Bresnan’s Lexical Functional Grammar

Lexical Functional Grammar⁴² (Bresnan 2001) was adopted for PT because it can account for feature unification in typologically diverse languages (Pienemann, DiBiase, Kawaguchi 2005: 205). Bresnan (2001: vii) states that “LFG is a theory of grammar which has a powerful, flexible, and mathematically well-defined grammar formalism designed for typologically diverse languages.” Thus, LFG represents the typologically plausible component of PT. Lenzing (2016: 4) summarizes the central idea of LFG as follows: “A central component of LFG is its projection architecture with three independent levels of linguistic representation that exist in parallel and are related to each other by specific linking or mapping principles.” The three levels comprise argument structure (a-structure), functional structure (f-structure) and constituent structure (c-structure). Pienemann points out that LFG is compatible with the key features in language processing outlined above (see Pienemann 2005: 15). This also applies to the procedural nature of language generation, put forward by Kempen & Hoenkamp (1987) and adopted by Levelt (1989). An example for this procedural nature is the storage of diacritic information in the grammatical memory store when generating S-V-agreement as depicted in Figure 10 above. Feature unification⁴³ is one of the key mechanisms in PT. Pienemann, DiBiase & Kawaguchi (2005: 200) illustrate feature unification with the help of the phrase ‘Peter sees a dog’ in the following way:

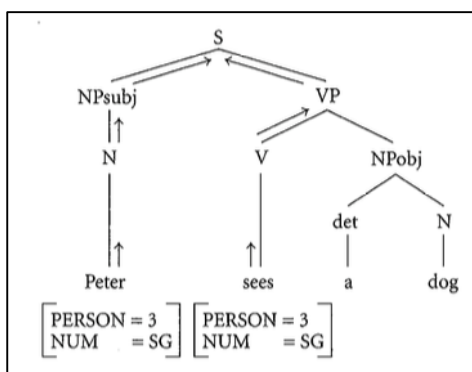


Figure 11: Three parallel structures in LFG, taken from Pienemann, DiBiase & Kawaguchi (2005: 200)

⁴² For a thorough and comprehensible account of LFG, see Lenzing (2013).

⁴³ The way in which representations of thematic roles are mapped onto grammatical functions is modelled in LFG by Lexical Mapping Theory (see Pienemann, DiBiase, Kawaguchi 2005: 212 for more detail).

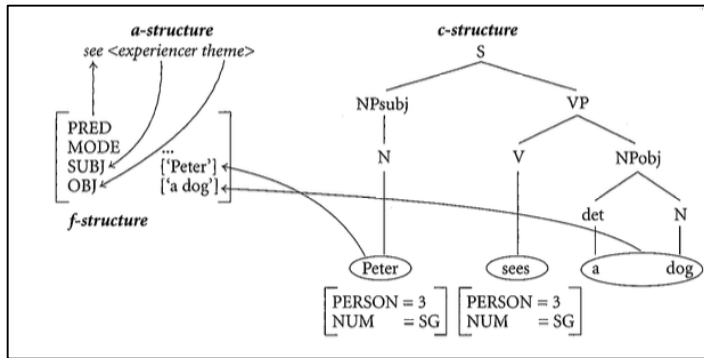


Figure 12: Feature unification in the s-procedure, taken from Pienemann, DiBiase & Kawaguchi (2005: 200)

Figures 11 and 12 illustrate how the production of the third-person affix '-s' relies on the features PERSON and NUMBER and their values PERSON=3 and NUMBER=SG contained in the subject noun phrase, so that S-V-agreement can be produced. Pienemann, DiBiase & Kawaguchi (2005: 200) explain that during the generation of this sentence, grammatical information on PERSON and NUMBER need to be stored in a grammatical memory store and need to be exchanged between the N and the V (see Levelt 1989).

Pienemann (2005: 15) argues that in LFG, this process is modelled by feature unification between three levels of linguistic representation. These levels are: (1) a-structure representing the semantic and syntactic side⁴⁴ of an utterance. In a-structure, the verb/predicator and its corresponding arguments are represented. (2) *F-structure* represents the linking element between the two levels named above and contains universal aspects of grammar, i.e. grammatical functions. The term (3) *c-structure* refers to the surface structure realization (Dalrymple 2001: 45) generated by phrase structure rules. Argument structure is the level of representation at which the core participants in events are represented (Bresnan 2001: 304). It "consists of a predicator with its argument roles, an ordering that represents the relative prominence of the roles and a syntactic classification of each role [...]." Lenzing (2016: 5), based on Bresnan (2001: 307), summarizes and explains the core aspects represented at a-structure as follows:

- the predicator and its corresponding argument roles
- the hierarchical ordering of the thematic roles according to their prominence

⁴⁴ Lenzing (2016: 5) points out that in the LFG tradition, there are different views on the nature of the semantic side contained in a-structure and refers to Falk (2001) and Fabri (2008) for different conceptions of a-structure.

- the syntactic features which are necessary to map arguments onto grammatical functions

The prominence of thematic roles is reflected in a thematic hierarchy that Bresnan (2001: 307) presents in a left-to-right order:

agent>beneficiary>experiencer/goal>instrument>patient/theme>locative

Lenzing (2013: 46) explains that

[t]his hierarchy descends from agent to locative and is responsible for structuring the thematic roles of verbs. According to Bresnan and Kanerva (1989: 23), the hierarchy of thematic roles is based on the assumption that there is a theoretical order of the relation of arguments to a predicator. This means that the arguments of a predicator are ordered in a specific way in the mental lexicon and that this order depends on the relative prominence of a thematic role that a particular argument takes.

(2) *Functional structure* contains grammatical functions, such as SUBJECT (SUBJ), OBJECT or OBLIQUE. These functions represent universal syntactic features that relate a-structure and c-structure (see Lenzing 2013: 23, based on Bresnan 2001: 47). Bresnan (2001: 95) maintains that grammatical functions can be realized in different forms in typologically different languages because the “[...] SUBJ function has no single universal structural form.” This is why f-structure is sometimes referred to as the glue in language (see the glue approach on the interface between syntax and semantics in Dalrymple 1999, 2001).

(3) *Constituent structure* is generated by phrase structure rules (Pienemann 2005: 16) and represents the surface syntactic organization of phrases (Lenzing 2013: 34). C-structure can be modelled by phrase structure trees, which are specific to any one language. Bresnan uses X-bar theory to formally model c-structure (see Lenzing 2013: 34). In this way, LFG cannot only account for endocentric languages such as English, but also for lexocentric languages that display more flexible word order and exhibit a case and agreement morphology (see Lenzing 2013: 34).⁴⁵ How are the levels of representation linked?

The linking element between a-structure and c-structure are mapping

⁴⁵ Bresnan (2001: 98) explains that “[e]ndocentric organization appears in highly hierarchical c-structures, such as we find in English. Lexocentric organization appears in flat c-structures with all arguments (including subjects) sisters of the verb, such as we find in [...] non-configurational languages of Australia.”

principles via functional structure (f-structure).⁴⁶ An illustration of the architecture and interaction between a-structure, c-structure and f-structure is depicted in Figure 13 below (taken from Lenzing (2013: 94), based on Pienemann et al. 2005):

Mapping process	Structures	Example
Linear default mapping	a-structure	<i>play</i> <agent patient/theme>
	f-structure	SUBJ OBJ
	c-structure	<i>John</i> played <i>the guitar</i> NP _{subj} NP _{obj}

Figure 13: Levels of representation in LFG, taken from Lenzing (2013: 94)

The most left column in Figure 13 labels the mapping process. In this case, the mapping process is linear.⁴⁷ The three structural levels are given in the second column. The third column shows an example of a linear mapping process between argument-, functional- and constituent structure using the phrase *John played the guitar*. Linear mapping depends on a one-to-one correspondence of thematic roles, grammatical functions and constituents. Linear mapping processes are considered to be easier to process (and therefore to be acquired earlier) than non-linear mapping operations. Mapping principles, such as those from c-structure to f-structure, ensure that one specific c-structure node can only be linked to one related f-structure (see Lenzing 2013: 39). Against the psycholinguistic background of Processability Theory, this entails that linguistic features have to be unified.

⁴⁶ Mapping principles use a number of well-formedness conditions, for more information see Bresnan (2001: 47f.)

⁴⁷ Linear mapping processes are assumed to be operable earlier in language development than non-linear mapping processes. Pienemann, DiBiase & Kawaguchi (2005: 201) explain that relationships between the levels of representation cannot only be linear because, if they were, “[...] semantic predicate-argument relationships could only be expressed by fixed surface word and phrase configurations.” Non-linear mapping processes underlie, for example, the production of non-canonical word order from c- to f-structure. This is realized, e.g., in the assignment of discourse functions (TOPIC and FOCUS). Another form of non-linearity is displayed in non-canonical word order from a- to f-structure as in the assignment of passive or causative constructions (Pienemann, Di Biase & Kawaguchi 2005: 223).

After having introduced the key mechanisms of the two yardsticks of PT, i.e. Levelt's model of sentence generation and Bresnan's formal theory of grammar, I will introduce the PT hierarchy in more detail.

2.2.2 The Hierarchy of Processing Procedures and Structural Options for English

PT is concerned with language processing in L2 acquisition, that explains universal developmental patterns in the form of a hierarchy of processing procedures. Pienemann (2005a: 8) hypothesizes the following hierarchy of universal processing procedures in L2 acquisition:

- i. lemma access,
- ii. category procedure (lexical category of the lemma),
- iii. phrasal procedure (instigated by the category of the head),
- iv. s-procedure and the target language word order rules,
- v. subordinate clause procedure – if applicable.

This hierarchy reflects a gradual, successive and cumulative development of processing procedures. The processing procedures are acquired in a step-wise fashion and the procedures are involved in the process of sentence generation in order to produce more and more complex linguistic structures. The hierarchy that results from this is implicational in nature, which means that “[...] the presence of a later structure implies the presence of an earlier structure.” (Pienemann 2011: 51).⁴⁸ The processing procedures are implicationally related (Pienemann 1998: 134). An implicational relationship (Pienemann 2005b: 21ff.) assumes that one processing procedure needs to be in place before the next processing procedure can develop. The implicational relationship that underlies the processing procedures leads to the prediction that no stage can be skipped by a learner, as each stage is a necessary pre-requisite for the next stage. The key assumption of PT that Pienemann (1998: 1) thus puts forward, is that a language learner can only produce those structures which are processable for him/her at any given point of time: “Structural options that may be formally possible, will be

⁴⁸ Meisel, Clahsen & Pienemann (1981) have shown that if an implicational relationship is assumed, cross-sectional study designs can be used for studying sequences of acquisition.

produced by the language learner only if the necessary processing procedures are available that are needed to carry out [...] those computations required for the processing of the structure in question” (Pienemann 1998: 1). The principle of information exchange enables the learner to unify grammatical information (Pienemann 1998: 97). The procedures will be described in more detail below.

In the first step, the lemma is accessed. Unanalyzed chunks might be retrieved from the lexicon. When the category procedure is in place (ii), it allows the learner to assign the grammatical category to a word, e.g. noun or verb. The phrasal procedure (iii) can be called after the grammatical category was assigned. The phrasal procedure allows the head of the phrase to be assigned. At this stage, information can be unified within phrase boundaries, so that for morphology, for example, determiner and noun agreement can be produced and appointment rules can be applied. Appointment rules determine the grammatical function of a phrase (e.g. subject) so that the s-procedure can be called. Now, information across phrase boundaries can be unified in the top s-node. If applicable, grammatical information can be exchanged between subordinate and main clause by the subordinate clause procedure (see Pienemann 2011: 36). Pienemann (1998: 68) illustrates the psycholinguistic background of the processing procedures by drawing on Levelt’s (1989) and Kempen & Hoenkamp’s (1987) notion of incremental language generation.

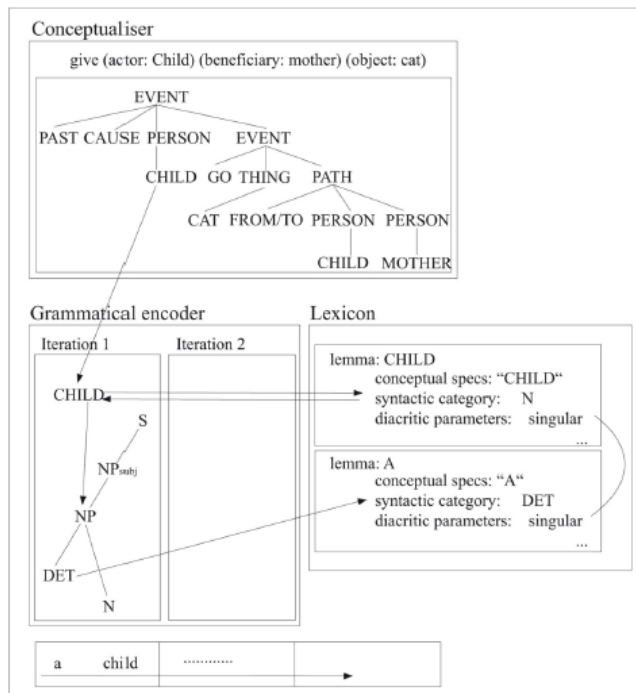


Figure 14: Incremental language generation, taken from Pienemann (1998: 68)

As envisaged by Levelt (1989), parts of the preverbal message that is generated in the Conceptualizer enters the Formulator. Thus, this piece of information is matched with the information stored in the lexicon and the lemma “child” is activated as it matches the pre-verbal message. As described in section 2.2.1.1, the lemma contains diacritic features and information about the lexical item’s syntactic category. In the case of the above example, the lemma contains the category information ‘noun’ for the word ‘child’. The information ‘noun’ calls the noun phrase procedure, so that the head of the phrase can be assigned. Here, the incremental nature of language processing becomes most apparent because at the same time that the head of the phrase is being processed, the conceptual material is inspected for possible complements and specifiers. When all this information is accessible, the lemma for ‘a’ is activated and the determiner is attached to the noun phrase, so the determiner ‘a’ can be inserted. ‘A’ contains the information ‘singular’ and this piece of information needs to be stored by the category procedure until it can be matched to its possible modifier (Pienemann 2005a: 7). In a next step, the grammatical functions need to be assigned by using appointment rules (in this case subject of S) (Pienemann 2005a: 8). In the example phrase ‘a child’, the attachment to a higher (s-) node is missing at this

point of the production process. If a sentence were to be produced, the noun phrase could be assigned the role of subject and the s-procedure could be called. For this, the diacritic features *person* and *number* would have to be stored in the s-procedure (Pienemann 2005a: 9). In order to illustrate the building procedures described above, it might be helpful at this point to refer to Pienemann's (2008: 16) figure displaying the locus of information exchange at different PT levels.

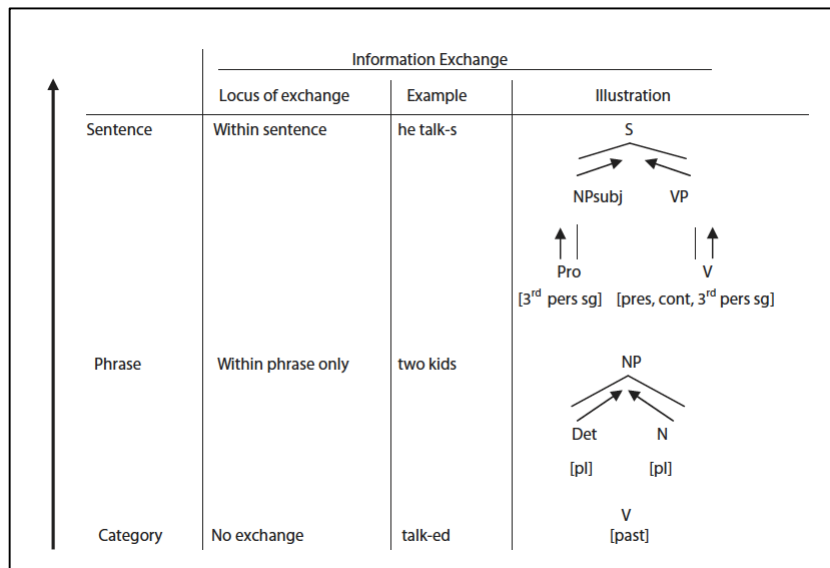


Figure 15: Locus of information exchange for morphology, taken from Pienemann (2008: 16)

Figure 15 shows that at the lemma level, no information exchange is assumed to take place. At the category level, lexical morphemes can be unified, such as in the example *talked* given above. At the phrasal level, information exchange within the phrase is possible, so that grammatical information can be exchanged between the determiner and noun. In the example above, this concerns the feature plural. At the sentence level, information exchange across phrasal boundaries, but within the sentence, is processable. Thus, subject-verb agreement can take place.

Pienemann (2005a: 13) summarizes the core claims of the implicational hierarchy of processing procedures as follows:

A word needs to be added to the L2 lexicon before its grammatical category can be assigned. The grammatical category of a lemma is needed before a category procedure can be called. Only if the grammatical category of the head of phrase is assigned can the phrasal procedure be called. Only if a phrasal procedure has been completed and its value is returned can Appointment Rules determine the function

of the phrase. Only if the function of the phrase has been determined can it be attached to the S-node and sentential information be stored in the S-procedure. And only if the latter has been stored can the target word order be arranged. In other words, it is hypothesized that processing devices will be acquired in their sequence of activation in the production process.

This perspective does not assume a target-language perspective on SLA, but a learner-centered one. Pienemann (2005a: 13) illustrates this by giving an example concerning developmental readiness. He argues that when learners are faced with developmental problems, i.e. when they are supposed to produce a structure which they are not yet able to process, the missing processing mechanism interrupts the generation of the structure in question. Instead of the regular operations taking place in a capable speaker, the conceptual material will be directly mapped onto the surface structure. Most often, this results in the production of canonical word order when information exchange cannot take place.

Pienemann (2005a: 14) illustrates the following hypothetical hierarchy of processing procedures:

	t1	t2	t3	t4	t5
S'-procedure (embedded)	-	-	-	-	+
S-procedure	-	simplified	simplified	Inter-phrasal information exchange	Inter-phrasal information exchange
Phrasal Procedure (head)	-	-	Phrasal information exchange	Phrasal information exchange	Phrasal information exchange
Category Procedure (lex. Category)	-	Lexical morphemes	Lexical morphemes	Lexical morphemes	Lexical morphemes
Word/lemma	+	+	+	+	+

Figure 16: Hypothetical hierarchy of processing procedures, taken from Pienemann (2005a: 14)

Figure 16 shows that since the S-procedure has not yet been developed at the first three stages of acquisition, phrases are generated using simplified procedures “[...] based on a direct mapping of argument structure onto functional structure” (Pienemann 2011: 37).

As discussed above, Pienemann (2015: 127) points out that PT was designed to address the developmental problem. Its roots go back to a more than 40-year-old tradition of second language research (see, e.g. Wode 1976, Clahsen 1980, Meisel et al. 1981). This tradition focused on, inter alia, German, with findings related to sequences of language acquisition and the question as to why learners seem to follow the same describable path in acquiring morpho-syntactic structures. Felix (1984) referred to the latter question as the ‘developmental problem’. Pienemann developed a theoretical framework that is powerful enough to address this problem, in that he claims that (2005b: 3) “[f]or linguistic hypotheses to transform into executable procedural knowledge (i.e. a certain processing skill), the processor needs to have the capacity of processing those hypotheses”. These predictions are an answer to the view of, e.g., Berwick & Weinberg (1984), who conceptualized the learnability of language in terms of a logo-mathematical problem (see Pienemann 1998: 1). Pienemann argues that humans are not simply equipped with a computing device, but that the human “[...] mind rather operates within psychological constraints” (Pienemann 1998: 1). The human mind needs to acquire processing routines which resemble procedural skills. The reason for arguing that processing routines, that underlie the developmental path are of procedural nature, rather than declarative nature (Pienemann 1998: 40f.), is the assumption that language production relies on non-conscious processes (Pienemann 1998: 5) because word retrieval happens very fast. This processing speed would not be maintained if declarative knowledge was used, because declarative knowledge is generally attributed to short-term memory with slower retrieval rates (see Pienemann 1998: 5).

In their later publications, Pienemann (2005c: 36) and Pienemann, Di Biase & Kawaguchi (2005) extended the explanatory scope of Processability Theory to address the logical problem. The logical problem is concerned with the source of linguistic knowledge. In other words, how do we know what we know about a language? To answer that question, the initial state of language acquisition has to be explored. Pienemann (2015: 134) explains PT’s hypotheses about the initial state to be based on minimal assumptions about innate linguistic resources. The assumption is that “[...] the basic notion of constituency and the

one-to-one mapping of semantic roles (such as agent, patient, etc.) is a given, and all other formal aspects of grammar follow from this"⁴⁹ Pienemann (2015: 134). In this regard, Lenzing (2013) was able to extend this discussion on explaining phenomena at the initial state in spelling out the Multiple Constraints Hypothesis (MCH):

[...] the L2 initial mental grammatical system is not fully developed in terms of mental representations. I [Lenzing] hypothesise that the initial L² mental grammatical system is highly constrained at the different levels of linguistic representations spelled out in LFG and that these restrictions also apply at the level of a-structure. The initial restrictions at a-structure level result in the learners' inability to map arguments onto grammatical functions. I [Lenzing] argue that beginning L2 learners rely on direct mapping processes from arguments onto surface form (Lenzing 2016: 3f.) [insertions by KH]

Lenzing (2013) provides striking evidence for her hypothesis that the initial mental system is highly restricted because the lexicon is not fully annotated and thus, the three linguistic levels of representation as assumed by LFG (Bresnan 2001) are not fully developed. This is why, very early learners at stage 1 of the PT hierarchy might produce an utterance like „it's a pink“. The adjective occurs in the wrong position. Lenzing assumes that in this kind of utterance, the arguments at a-structure level are directly mapped onto c-structure. This is displayed in Figure 17 below:

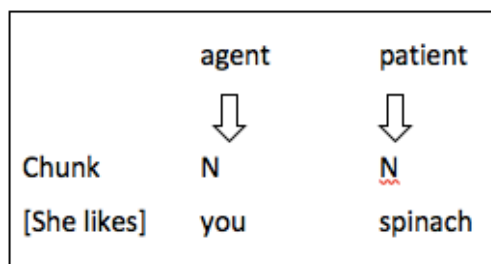


Figure 17: Direct mapping of argument onto surface form, taken from Lenzing (2013: 216)

A full representation of the multiple constraints on the initial mental grammatical system, as hypothesized by Lenzing (2013), is given below:

⁴⁹ This position is very different from strong nativist positions, such as Chomsky's Universal Grammar account.

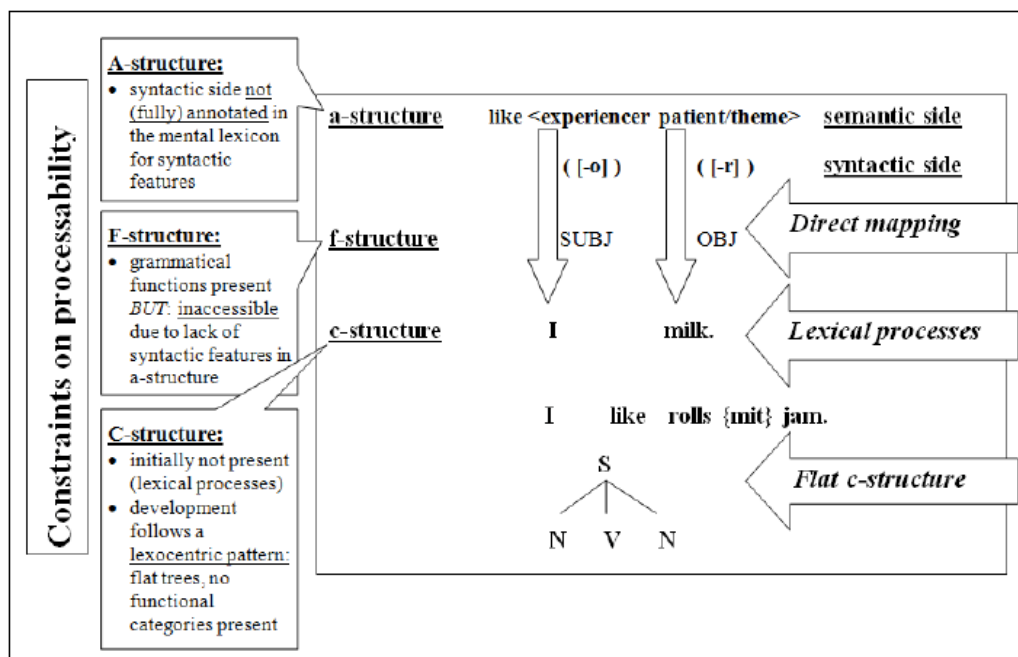


Figure 18: The Multiple Constraints Hypothesis, taken from Lenzing (2013: 8)

Figure 18 shows that the early mental grammatical system is constrained at all three levels of linguistic representation, because the respective mental representations are not fully developed. At a-structure level, the syntactic side of the mental lexicon is not fully annotated, which hinders the learner to map arguments onto grammatical functions. Grammatical functions at f-structure level are inaccessible because syntactic features at a-structure level are not present. This is why a direct mapping process from a- onto c-structure is performed, that results in a flat c-structure (see Lenzing 2016: 3f.). Therefore, learners at stage 1 of PT hierarchy are only able to produce holistically stored linguistic chunks or formulaic patterns (and single words). These patterns are assumed to appear at stage 1 of the acquisition process as modelled in PT.

In the following, the structural options processable for language learners at the different stages of acquisition are exemplified by the hierarchy for English as a second language. The hierarchy for English as an L2 is presented as follows:

Stage	Processing procedures	Phenomena	Examples
6	Subordinate clause - procedure	Cancel Aux-2 nd	I wonder what he wants .
5	S-procedure	Neg/Aux-2 nd -? Aux-2 nd -? 3sg-s	Why didn't you tell me? Why can't she come? Why did she eat that? What will you do? Peter likes bananas.
4	VP-procedure	Copula S (x) Wh-copula S (x) V-particle	Is she at home? Where is she? Turn it off !
3	Phrasal procedure	Do-SV(O)-? Aux SV(O)-? Wh-SV(O)-? Adverb-First Poss (Pronoun) Object (Pronoun)	Do he live here? Can I go home? Where she went? What you want? Today he stay here. I show you my garden. This is your pencil. Mary called him .
2	Category procedure	S neg V(O) SVO SVO-Question -ed -ing Plural -s (Noun) Poss -s (Noun)	Me no live here. / I don't live here. Me live here. You live here? John played. Jane going. I like cats. Pat's cat is fat.
1	Word / lemma access	Words Formulae	Hello, Five Dock, Central How are you? Where is X? What's your name?

Table 1: PT hierarchy for English as a L2, taken from Lenzing, Plesser, Hagenfeld & Pienemann (2013: 272), on the basis of Pienemann (2005a: 24)

At stage 1, the learner is hypothesized to produce mostly formulaic sequences and unanalyzed chunks. This means that early learners are not yet equipped with the necessary processing procedures that allow for syntactic operations. Thus, they rely on lexical processes for the production of a respective structure. A mere retrieval of words or chunks of words from the mental lexicon is therefore assumed. Lenzing (2013: 160) discusses and reviews the various terms and definitions that have been used in the literature to model formulaic sequences. Terms comprise, e.g., unanalyzed forms, prefabricated routines, formulae, etc. Her definition of formulaic sequences, in the context of PT, is the following:

[...] at the beginning of the L2 acquisition process formulaic structures occur as unanalysed forms in learner's speech. These unanalysed sequences are located at stage 1 of the PT hierarchy, as at this stage, the early L2 learner lacks the necessary processing procedures to (1) assign a lexical category to the lexical material and (2) exchange grammatical information within a constituent or across constituent boundaries (cf. Pienemann 2002). It is precisely for this reason that the learner is initially only able to produce single words and unanalysed units" (Lenzing 2013: 162)

In the context of her study, and on the basis of Krashen & Scarcella (1978), Lenzing (2013: 163) further unravels the term 'formulaic sequence' in a logical way. In her view, 'formulaic sequences' is an umbrella term that covers (1) formulae and (2) formulaic patterns. The term *formulae* refers to those structures that the learners encountered as fixed expressions in their textbook. Lenzing (2013: 163) hypothesizes that these expressions are stored holistically in the learner's mental lexicon. Formulaic patterns consist of an unanalyzed chunk along with an open slot. To fill this slot, the learner has to employ his/her own strategy, such as in 'how is X'. She argues that the identification of formulaic patterns requires a careful distributional analysis to identify unanalyzed language use from productive use (Lenzing 2013: 164). At stage 2 of the PT hierarchy, the category procedure is in place. At stage 2, there is still no unification of grammatical features, but diacritic features are present (such as number), so that lexical entries can be directly mapped onto conceptual structures, if the feature is constrained to one constituent (Pienemann 1998: 171). Structural operations are limited to, inter alia, producing plural and possessive forms with nouns or the past-'ed' in terms of morphology. Simple SVO structures and SVO interrogatives

are possible in terms of syntactic procedures. The latter are often indicated by rising intonation. At stage 3, the phrasal procedure can be called. It allows for the fronting of do/auxiliaries or Wh-question words in otherwise canonical SV(O)-questions. In terms of morphology, plural agreement is possible because determiner and noun with the diacritic feature for plural can be unified when the phrasal head is assigned (Pienemann 1998: 172). Interrogative structures of the English hierarchy, located at stage 4, are 'Copula S (x)', 'Wh-copula S (x)', such as 'Is she at home?', 'Where is she?' or the 'V-particle' 'Turn it off'. At stage 5, a new processing operation is possible, which enables learners to produce subject-verb agreement. Information can now be unified at sentence level (inter-phrasal morphemes), so that diacritic features for person and number that are required for the production of the third-person-s, can be held in the S-procedure (Pienemann 2011: 58). At the level of syntax, sentences like 'Peter likes bananas' or questions in which the auxiliary is placed in second position (Where did she come from?) can be produced. Information between subordinate clauses and main clauses can be unified at stage 6 in the hierarchy. A stage 6 phenomenon is that learners are able to cancel the auxiliary in second position, such as in 'I wonder what he wants.' Stage 6 is the last stage modelled in PT so far. This does not mean that it is the top of the acquisition process. Rather, many of the structures contained in the hierarchy are obligatory in nature and therefore more directly assessable from the learner's spontaneous speech than optional structures (as e.g. a passive construction). Optional structures also have the potential to be integrated into the hierarchy.

PT focuses on the processing of linguistic features that can account for the development of morphosyntactic features. PT does not aim to explain all issues connected to language acquisition. Pienemann (1998: 32ff.) explains that PT deliberately takes a modular approach to explaining second language development and that "[...] currently there is no one framework which can provide satisfactory answers to all [...] explananda in language acquisition" (Pienemann 1998: 32). Explananda encompass e.g. the origin of linguistic knowledge, how linguistic knowledge is generated, acquired and produced. Therefore, Pienemann (1998: 33) makes the case that "[...] it is a worthwhile

research strategy to reduce the task of explaining SLA to discrete subtasks, and to employ different theoretical modules for each of those tasks as long as the different modules are to communicate with each other and are theoretically consistent.” The module that PT develops is concerned with processability of linguistic structures. This specialist theory is assumed to be “[...] capable of unifying a whole range of domains and thereby solidifies its explanatory value” (Pienemann 1998: 34). However, Pienemann (2005b: 69) describes that this modular approach can be extended by other necessary modules at a later stage.

To recapitulate, the learner is able to produce those structures that are processable for him/her at a given point in time. If the learner is supposed to solve a developmental problem that s/he is not yet ready for, then there is a certain leeway of options for him/her to produce. This variation in learner language is captured by the concept of Hypothesis Space.

2.2.3 Hypothesis Space

To describe any current state of L2 development, Pienemann takes up on Selinker’s (1972) definition of *interlanguage (IL)*⁵⁰. Selinker argues that when acquiring a language, the learner develops a separate interim system that is neither the first language nor the target language but bears features of both systems. A number of early studies on interlanguage systems have shown that ILs are systematic and internally consistent in nature (Corder 1967; Selinker 1972). Others, such as Huebner (1979) or Tarone (1983) have argued that variability plays a prominent role in interlanguages, rendering them unsteady systems.⁵¹ Ellis (1985: 118), for instance, argues that “[t]o claim that interlanguage is on the one hand systematic and on the other hand variable is potentially contradictory.” Within the PT framework, Pienemann (1998: 231ff.)

⁵⁰ Corder (1967) introduced the notion of a separate linguistic system during the process of acquiring an additional language and termed this notion ‘transitional competence’.

⁵¹ Liebner and Pienemann (2011: 70) note that Huebner’s (1979) or Tarone’s (1983) argument about unsteady interlanguage systems arises from their view of language acquisition based on accuracy measures. However, Pienemann (1998: 132) was able to show that accuracy is not a valid measure of development. This discussion is particularly interesting for this thesis, as it seems that the CEFR takes a similar accuracy-based approach to describing grammatical ability in language learners (see chapter 2.1.6).

was able to show that IL variation is indeed systematic and can be captured by the concept of Hypothesis Space.⁵² Hypothesis Space determines the possible range of interlanguage variation in an a priori way (Pienemann 1998: 239), by assuming that variation is constrained by the level of processability. Pienemann & Lenzing (2015: 164) explain that Hypothesis Space “[...] is created by the interplay between the Processability hierarchy and the leeway it generates at every level” of processability. Hypothesis Space is illustrated in Figure 19 below.

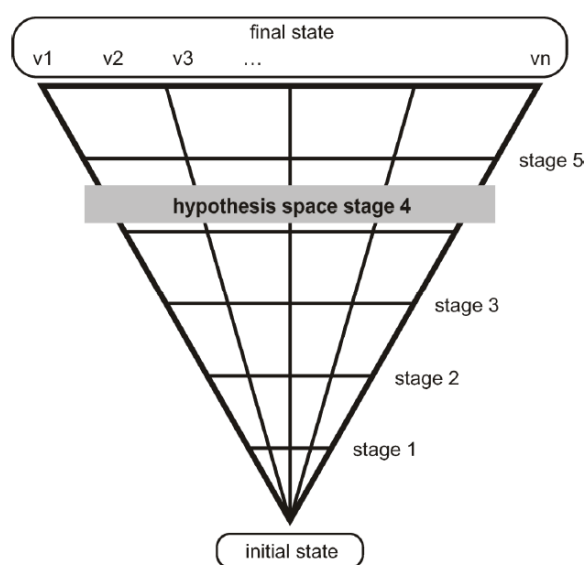


Figure 19: Hypothesis Space, taken from Pienemann (1998: 232)

Figure 19 shows that Hypothesis Space is two-dimensional. The horizontal dimension shows the developmental stages, starting at the initial state. The vertical dimension shows the increasing range of possible structures that a learner can potentially produce while progressing through the developmental stages. The possible range of interlanguage structures is constrained by the level of processability, but at the same time Hypothesis Space “[...] captures the dynamics of the interlanguage system, as it permits individual developmental trajectories (including the variants chosen by the learner) to be represented within one overall system” (Liebner & Pienemann 2011: 70).

⁵² Pienemann (1998: 232) maintains that the notion of Hypothesis Space is based on the Multidimensional Model by Meisel et al. (1981). As described in chapter 2.2.5, the MMM assumes two dimensions, a horizontal and a vertical dimension, that exhibit the SLA process. PT extends the horizontal dimension by formally modelling it in Hypothesis Space.

The idea of variation is that it arises from the choices a learner has at any stage of development, given the constraints on processing. If a learner wants to produce a stage 5 structure, like an auxiliary in second position in a Wh-question (Where have you lost it?), but the language processor prevents the structure from being processed, s/he needs to make recourse to different solutions. The learner might use canonical word order after the Wh-question word (Where you have lost it?) or omit the auxiliary in second position (Where \emptyset (have) lost it?) (examples taken from Liebner & Pienemann 2011: 65). The learner utilizes those processing resources that are available to him/her at their current stage of development, to solve the problem of not being able to insert the auxiliary in second position. Although the learner's strategy results in the production of ungrammatical/non-target-like utterances, these utterances can be explained by Hypothesis Space. What is implied in this, is that grammatical accuracy is not a valid measure of language development, because order of accuracy in the acquisition process does not necessarily reflect the order of development. In this regard, Pienemann (2015: 142) states "frequency and accuracy rates are invalid measures of development when development is understood as increased complexity of the overall system." Thus, Pienemann (1998), based on work by Meisel et. al 1981, proposes an alternative criterion to measure development: The Emergence Criterion.

2.2.4 The Emergence Criterion

The Emergence Criterion is an answer to the question as to how the term *acquisition* can be measured, and, more specifically, how the onset of acquisition can be defined. The Emergence Criterion (e.g. Meisel et.al. 1981, Pienemann 1998, Pienemann 2005, Pallotti 2007, Pienemann & Lenzing 2015) is an approach to define and operationalize an acquisition criterion, which in turn can be used for identifying the point in time when acquisition takes place. PT deliberately takes a modular, psycholinguistic view to explaining language acquisition and views acquisition in terms of universal processing mechanisms, that account for the unfolding of a developmental sequence of morpho-syntactic structures.

Acquisition is thus viewed as development rather than competence. In PT, the acquisition criterion can be formulated in a most straight-forward manner so that it might be used for research purposes. Pienemann (1998: 138) explains that

[...] emergence can be understood as the point in time at which certain skills have, in principle been attained or at which certain operations can, in principle, be carried out. From a descriptive viewpoint one can say that this is the beginning of an acquisition process and focusing on the start of this process will allow the researcher to reveal more about the rest of the process.

In other words, when a linguistic structure emerges in the interlanguage of a learner, this structure can, in principle, be viewed as acquired. In order to identify if a structure is used productively by a learner, several conditions of quantification of the emerged structure have to be met. Pienemann (1998: 133) exemplifies the EC with the phrase *he goes*. *He goes* requires subject-verb agreement and is placed at stage 5 of the Processability Hierarchy. Pienemann argues that to determine whether the third-person-s has been acquired, both subject and verb need to vary morphologically and lexically. This means that in the speech sample, the third-person-s needs to occur with different verbs, such as *sleeps* or *talks* and additionally, the verb needs to occur with morphological variation, such as *going* or **goed*. These conditions are used to rule out if the structure in question is produced by chance or was primed in a conversation, or simply is an unanalyzed chunk/formulaic language. In PT-based research, further distributional analyses are carried out in connection with the EC to unambiguously determine whether a structure can be assumed to be acquired (see e.g. Lenzing 2013; Lenzing 2017). Pienemann (1998: 135ff.) points to several advantages of the use of an emergence criterion over an accuracy criterion. Accuracy criteria were used in the Morpheme Order Studies that were popular in language acquisition research in the 1970s (see e.g. Dulay & Burt 1973). Morpheme Order Studies investigated the suppliance of morphemes in obligatory contexts and placed them in rank orders of acquisition. Based on this methodology, Krashen (1977) proposed a 'natural order of acquisition'. However, Pienemann (1998: 137) argues that "[t]his analysis does not have the potential of describing the dynamics of interlanguage development even though it

produces a neat rank order of accuracy of morpheme insertion". He illustrates this reasoning by using the following Figure on accuracy and development:

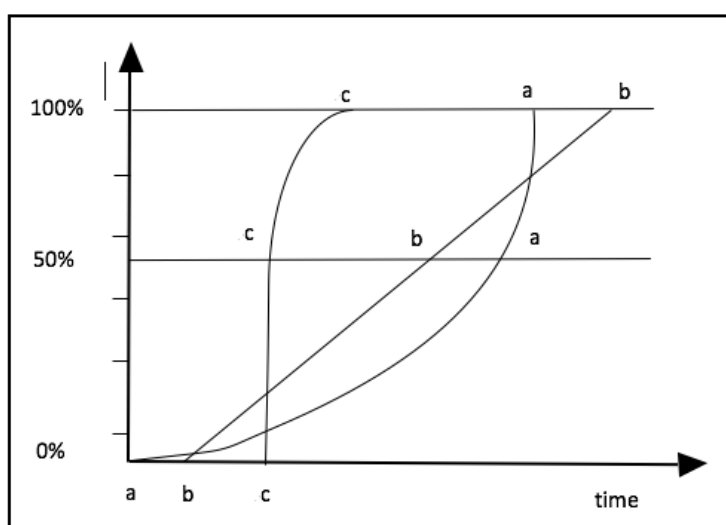


Figure 20: Accuracy and order of acquisition, taken from Pienemann (1998: 137)

The lines termed 'a, b and c' illustrate the development of different linguistic structures. The horizontal axis denotes the time and the vertical axis depicts the percentage of accurate rule applications. Figure 20 shows that the application of accurate structures in obligatory contexts show different patterns at different points in time. Structure b, for example, is depicted to increase in a linear way whereas structure a "[...] has a flat gradient" (Pienemann 1998: 137). So, if one took cross-sections at different suppliance rates, the Figure would exhibit different orders of accuracy: 1% a-b-c, 50% c-b-a, 90% c-a-b (Pienemann 1998: 137). For this reason, Pienemann proposes to use the cut-off point of 'emergence of the structure', because this is the only point that remains constant over time (see Pienemann 1998: 138). The use of an emergence criterion together with a careful distributional analysis allows to determine which contexts and which lexical features are related to which interlanguage rule (Pienemann 1998: 139). Pallotti (2007: 365) further highlights that the notion of the EC "[...] is theoretically well-founded and many of the methodological problems involved in its operationalization are convincingly worked through."

Pienemann's hypotheses about accuracy criteria are central to the discussion on interfaces between the CEFR and SLA, in terms of grammatical

ability. The discussion so far shows that grammatical competence in the quantitative dimension of the CEFR should comprise more than a scale for grammatical accuracy (see chapter 2.1.6 for more details), because language development cannot be conceptualized on the basis of learner displaying certain levels of grammatical accuracy at any level of development. Before laying out this argument in more detail in chapter 3.3.3, PT's predecessors will be described in the next chapter.

2.2.5 Historical Background to Processability Theory

The history of describing developmental schedules covers more than 40 years of research (Pienemann 2015: 123). Lenzing (2013: 99) describes that “[o]ne of the most ground-breaking discoveries in the field of SLA has been the insight that the process of acquiring a second language does not take place in a random, unpredictable way, but proceeds in a regular and systematic fashion.” The 1970s were an era in which the so-called Morpheme Order Studies (see e.g. Brown 1973)⁵³ based on first and second language acquisition, raised the question whether language acquisition of specific morphological structures occurs in a specific order. The Morpheme Order studies were heavily criticized for taking, inter-alia, an accuracy-based view on language acquisition (see e.g. Dulay et al. 1982, Hatch 1978). Pienemann (1998: 137) argues that development cannot be modelled through accuracy, because “[...] accuracy rates develop with highly variable gradients in relation to grammatical items and individual learners.” A number of studies focused on a more learner-centered perspective on sequences⁵⁴ in the acquisition of German (see e.g. Wode (1976), Bongaerts & Jordens (1985), Zobl (1986) or du Plessis et al. (1987)). The most well-known early research of developmental sequences is probably that of the ZISA⁵⁵ research group on the acquisition of German by Italian and Spanish children (see e.g.

⁵³ For a detailed summary of the history of acquisition order research, see Ellis (1994) or Lenzing (2013: 99f.).

⁵⁴ Lenzing (2013: 100), following Ellis (1994), argues that the term ‘order’ refers to whether some morphological features are required earlier than others, whereas the term ‘sequence’ refers to an interlanguage-based approach to what developmental patterns look like.

⁵⁵ ZISA is an acronym for „Zweitspracherwerb Italienischer und Deutscher Arbeiterkinder“.

Clahsen 1980; 1984, Meisel et al. 1981, Pienemann 1981). There is still ongoing interest in developmental stages in SLA (see e.g. Lenzing 2013; 2017) and their implications for teaching (e.g. Maier et. al. 2016, Roos 2016).

The most important points of reference for the development of PT are the Multidimensional Model (MM) (Meisel, Clahsen & Pienemann 1981), the Strategies Approach (Clahsen 1984), the Teachability Hypothesis (Pienemann 1984) and the Predictive Framework (Pienemann & Johnston 1987) (see Pienemann 1998, 2005). Pienemann highlights that PT is not merely another label for the Multidimensional Model, but a new theoretical framework aiming to overcome limitations of its predecessors (Pienemann 2005: 71). The Multidimensional Model was developed in the 1980s as one possible account to explain developmental stages found in SLA. Pienemann (2005c: 71) explains that the Multidimensional Model is a framework to describing interlanguage dynamics assuming SLA to comprise at least two dimensions, i.e. development and variation. Variational features include those linguistic features that cannot be attributed to the developmental dimension but occur in language development at various different stages individually by a learner (see Pienemann 2015: 130). Pienemann (1998: 143) argues that within the MMM, no such predictions on two dimensions in SLA were made and that instead, development was seen in an a-priori manner; i.e. the prediction in the MM are subject to theoretical deduction rather than based on observation. This leads to Larsen-Freeman & Long's (1991) criticism as to the MM that any random deviation from the predicted sequence might be attributed to a variational dimension. Such a broad concept of variation would, so they claim, render the model unfalsifiable.

The Strategies Approach by Clahsen (1984) aims to explain the acquisition of German L2 word order in terms of the acquisition of strategies that can overcome constraints of psychological complexity. According to Pienemann (2005), the psychological complexity of a structure depends on how much rearrangement of the surface linguistic structures, in relation to its mapping on the semantic side of the utterance, has to take place. Pienemann (2005c: 73) further states that the Strategies Approach is a complementation to the Multidimensional Model, although they are two separate approaches, and

argues that the Strategies Approach provides an explanation for the developmental pattern that was described in the MM.

The Predictive Framework (Pienemann & Johnston 1987) was an extension of the Strategies Approach to English as an L2 to focus on selected morphological structures. Pienemann (2005c: 73) points out that the approach was soon discarded after its proposal, because of the limitations of the Strategies Approach and the conceptualization of Processability Theory.

With the development of PT, some of the issues present in the MMM and the Strategies Approach were overcome. Pienemann (2005c: 71ff) argues that there is a fundamental difference between PT and the ideas that had been developed before, such as the ones outlined in the Multidimensional Model. The difference lies in the more precise modeling of the developmental route and the explanation of a broader range of phenomena through a typologically and psychologically plausible framework in PT. Also, PT is able to address the developmental (Pienemann 1998) and the logical problem (Pienemann et. al 2005, Lenzing 2013; 2016).⁵⁶ This is due to the inclusion of Lexical-Functional Grammar into PT, a theory of generative grammar that “has a high degree of psychological and typological plausibility and that allows one to model several key aspects of language generation using feature unification” (Pienemann 2005c: 74). Further, the lack of falsifiability of the Multidimensional Model is overcome by the integration of the concept of Hypothesis Space (see Lenzing 2013: 122). This allows PT to model the two dimensions of language acquisition, i.e. development and variation, as described by the Multidimensional Model, in a clear and falsifiable way (Pienemann 2005c: 74). With the development of PT, criticism as to Clahsen’s strategies about its limited scope that only focuses on word order is also overcome. Furthermore, by integrating LFG, the lack of a relationship between processing strategies to representations of grammar are overcome (see Pienemann 2005c: 73).

⁵⁶ These issues are discussed in chapter 2.2 where the core ideas of Processability Theory were presented. In a brief and simplified manner, the logical problem is concerned with the nature and source of linguistic knowledge of an additional language.

Logically, it is easier to backtrack the predecessors of PT's modular approach to SLA than to find the roots for each of the multitude of concepts that the CEFR's holistic approach to competences in language users describes. Pienemann appreciates and makes explicit which ideas he used to conceptualize his theoretical framework. With the CEFR, in most of the cases (chapter 2.1.7), the informed reader needs to infer as to where the concepts of competence, acquisition, assessment, etc. originate from.

In the next chapter, the Teachability Hypothesis and the concept of Developmental Readiness will be presented. These notions show how the hypotheses, formalized in PT, can inform language instruction as they are crucial for the perspective on learner errors underlying this thesis.

2.2.6 Teachability, Developmental Readiness and Learner Errors

The Teachability Hypothesis (TH) was put forward by Pienemann (1984) and ties in with the notion of constraints on the learnability of linguistic structures that derive from an underdeveloped language processor. The formulation of the TH is spelled out in a way that "it is testable for the whole range of second language grammar" (Mackey et. al. 1991: 65). The assumptions made by the TH have therefore been empirically tested in a number of studies (see e.g. Pienemann 1984, Ellis 1989, Roos 2007, Spada & Lightbown 2008). In his early research on the effect of formal instruction on developmental sequences, Pienemann compared the natural order of the acquisition of morpho-syntactic structures by 10 Italian learners of German as a L2 before and after instruction (Pienemann 1984). He found that instruction did not alter the route of acquisition and concluded that "[...] the relevant acquisitional stages are interrelated in such a way that *at each stage the processing prerequisites for the following stage are developed*" (Pienemann 1984: 37, italics in original). This finding is addressed by the TH (Pienemann 1984, 1989).

The core claim of the TH is that "stages of acquisition cannot be skipped (through teaching intervention) because of the cumulative nature of the processing strategies. It also predicts that variational features are not subject to

the same constraints on teachability” (Pienemann 2005c: 73). This idea derives from the assumption discussed above that SLA, like natural language acquisition, follows a universal pattern and that this pattern is implicationaly related. The implicational nature of the hierarchy of processing procedures prevents the learner from skipping stages despite formal instruction (see Pienemann 2015: 137). Therefore, “[...] instruction can only promote language acquisition if the interlanguage is close to the point when the structure to be taught is acquired in the natural setting (so sufficient processing resources are developed.” (Pienemann 1984: 37). What follows from this, is that language instruction “[...] should build on the learning process occurring outside the classroom and incorporate them [internal syllabi] into [...] [formal] acquisition” (Pienemann 1989: 53).

The TH puts forward that the same constraints found in the acquisition of a second language apply to the teaching of this language. It entails that language teaching is successful only if the structures to be taught are manageable for the current state of the language processor (cf. Pienemann 1984, 1987). The TH’s claim is that learners are not able to skip stages through formal instruction, but that instruction may be beneficial, if it focuses on the current developmental stage or slightly above it, that means if the learner is developmentally ready for acquiring the respective structure (Keßler et al. 2011: 150). The concept of Developmental Readiness is illustrated by Keßler (2006: 96) in the following way:

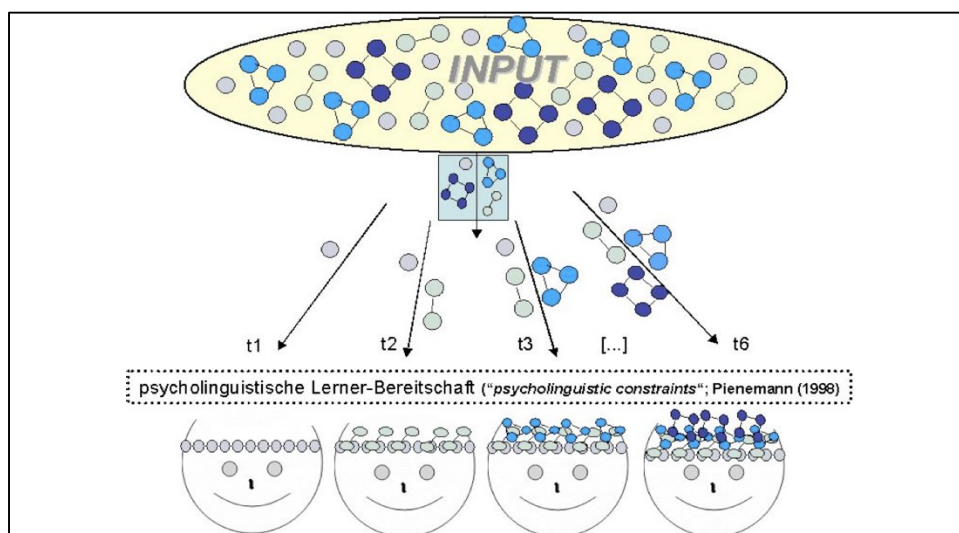


Figure 21: Developmental Readiness, taken from Keßler (2006: 96).

Figure 21 shows that learners integrate those linguistic structures into their IL, which they are able to process or which they are close to the point of processing. In other words, they acquire those structures which they are developmentally ready for. It is important to note that the TH defines constraints on teaching but does not promote a deficient approach to the teaching of grammatical structures. Rather, it should be acknowledged by teachers that if a learner is developmentally ready to acquire a structure, i.e. if the structure to be taught is in accordance with the current developmental stage or slightly above it, “instruction can improve acquisition with respect to (a) the speed of acquisition, (b) the frequency of rule application and (c), the different linguistic contexts in which the rule has to be applied” (Pienemann 1985: 37). In this context, Roos (2014: 3) argues that “[...] it “pays” to take the learner’s developmental readiness into account in the teaching process, it also adds a new and beneficial dimension with regard to the timing of instruction.” This is in contrast to many traditional approaches to formal language instruction. Roos (2014: 2) summarizes the notion of traditional formal instruction to be as follows:

With regard to the timing of instruction, and the question what to teach and when, traditional approaches to foreign language teaching are based on the idea that language learning is a linear process. The basic principle is that structures that are perceived to be simple are taught before complex or difficult ones. This principle is accompanied by an assumption that items are learned in the order in which they are taught.

However, researchers argue that SLA is a gradual, cumulative and dynamic process (see e.g. Ellis 2009: 237), and that the teaching of isolated grammatical structures might not lead to their acquisition (see Long & Robinson 1998). Amongst others, Di Biase (2002), Keßler & Plessner (2011) and Roos (2014) have argued for combining developmentally moderated approaches to language teaching, in that they highlight the potential of task-based approaches for catering for learners' internal syllabi.⁵⁷

The concept of developmental readiness also has repercussions for the treatment of learner errors⁵⁸ in class. With knowledge about developmental readiness, learner errors can be seen as positive indicators of language development that give rise to creative language use (Larsen-Freeman & Long 1991: 57). This view is contrary to errors being viewed as indicators of lack of competence (which accuracy-based approaches might assume). However, a distinction should be made between developmental and variational errors. Keßler (2006) argues that developmental errors occur because the learner is supposed to produce a structure that the current state of his/her processor is not yet able to handle. That is why a stage two learner might not produce correct S-V-agreement that is processable at stage 5. Variational errors, on the other hand, arise from the choices that learners make when they have already acquired the necessary processing resources. Keßler et al. (2011: 153) argue that not all errors should be treated in the same way. Based on the slogan *message before accuracy*, they claim that for learners who are not developmentally ready, it will not make sense to correct developmental errors, because the learners are simply not able to produce the target-like form. However, they maintain that corrective feedback⁵⁹ on a learner's developmental error

⁵⁷ Task-based instruction puts tasks at the center of lessons. The tasks are manageable by learners at all stages of acquisition and exhibit a number of communicative advantages with a primary focus on meaning (Ellis 2009, Mackey 1999, Roos 2007, Spada & Lightbown 2008, Lenzing & Roos 2012).

⁵⁸ The treatment of errors during the language acquisition process has been subject to extensive studies. See e.g. Dulay & Burt (1974) for the study of syntactic errors or Corder (1967) for early thoughts on interlanguage errors.

⁵⁹ Corrective feedback can be provided in several different ways, ranging on a continuum from more explicit to more implicit approaches. See, e.g. Mackey (2006) or Ding (2012) for a discussion of the effectiveness of different types of corrective feedback.

might be beneficial to others in the classroom (see Keßler et al. 2011: 153). Variational errors, on the other hand, should be corrected so that the learner does not overuse a specific problem-solving strategy. They argue that “[...] simplified choices may accumulate and result in simplified interlanguage variation” (Keßler et al. 2011: 153).

In order to determine the learners’ levels of development and to identify the type of errors produced by them, it is beneficial to assess their current state of interlanguage. Keßler (2008) argues that in order to diagnose a learner, a diagnostic task cycle can be employed that allows for informed formal intervention (see explanations in chapter 2.2.7). He suggests that the use of the semi-automatic diagnostic software *Rapid Profile* (Pienemann 1990; 1992) is beneficial for obtaining a full picture of learner development. This software is introduced in the next chapter, as I argue that a combined proficiency assessment with Rapid Profile paints a more reliable picture of a learner’s language proficiency than standard proficiency ratings alone.

2.2.7 Linguistic Profiling and Rapid Profile

Linguistic Profiling is based on early work by Crystal et al. (1976), Crystal & Fletcher (1979), Crystal et. al. (1989) in the field of speech pathology. This work resulted in the construction of the Language Assessment Remediation and Screening Procedure (LARSP). LARSP is a diagnostic procedure that helps to allocate learner language in terms of grammatical disability (Crystal et. al. 1989) and focuses on the acquisition of English as a first language in monolingual settings (Pienemann et. al. 1988: 231). Pienemann (1992: 2) reports that in the original versions of linguistic profiling, the assessor fulfills basically the same function as a researcher in conducting very long and thorough analyses that can take up to 20-40 hours of assessment for only one individual. He thus argues that such a time-consuming procedure, especially in the transcription and data elicitation phases, may be viable in speech impairment contexts, “[b]ut in the [present] situation of SL teaching, such a procedure has to be judged as

impractical (for reasons of time, expertise, training and costs)” (Pienemann et al. 1988: 231). Pienemann et al. (1988) put forward a simplified procedure for assessing the language development of ESL learners. This procedure contains an observation scheme which does not necessitate a full linguistic analysis but consists of an online-observation in which the test administrator notes down whether the linguistic structures in question are evident in the learner’s speech sample. In their study from 1988, Pienemann et. al. explored as to whether their revised version of profile analyses, based on universal language acquisition patterns, is a reliable and feasible way of determining a learner’s language development. Their results displayed a number of useful factors that helped to devise a more sophisticated version of an acquisition-based assessment tool. With technological advancement, it was possible for Pienemann & Jansen (1991), Pienemann (1992) and Mackey, Pienemann & Thornton (1991) to extend the feasibility of profile analyses. By using a computer as an assistance to the assessor, many of the contexts in which biases related to the person who administers the test are likely to arise, are minimized. In 1990, the COALA (Computer-assisted Linguistic Analysis) software (see Pienemann & Jansen 1991; Pienemann 1992)⁶⁰ was devised, a computational system for the linguistic analysis of language acquisition data. COALA constitutes the predecessor to Rapid Profile, the software that is currently used for profile analyses within the PT framework. Like COALA, RP is also based on general profile analyses, into which the semi-automatic make-up as well as internal algorithms were transferred.

Rapid Profile was developed by Pienemann (1990, 1992) at Sydney University as part of in the National Language Institute of Australia and the Language Acquisition Research Centre (LARC). Since its development, RP has been subject to a number of empirical studies and was also used as an instrument in a number of empirical studies (Mackey et. al. 1991, Pienemann & Mackey 1993, Pienemann et al. 2006, Pienemann & Keßler 2007, Keßler 2006; 2007; 2008, Keßler & Keating 2009, Michalska 2010, Lenzing & Plesser 2010, Hagenfeld

⁶⁰ Pienemann (1992) integrates a reporting function into the COALA software for more detailed feedback of the interlanguage systems as assessed during the interviews.

2017). It was devised as an acquisition-based diagnostic language assessment tool that compares learner language to standard patterns of language acquisition. The standard patterns can be used as fixed reference points for the assessment of language development (see Pienemann et. al 1991: 61). In this regard, Pienemann (1992) highlights that Rapid Profile is a criterion-referenced screening procedure that is able to condense the standard profile analyses to a 20-minute interview.⁶¹ Its criterion-reference allows the program to consider both what the learner is able to do with the language, as well as what the learner cannot do (see Keßler & Plesser 2011: 230). This refined version of profile analyses is assumed to be more compatible with the needs of practitioners with regard to formal instructional settings.

The major rationale behind the development of Rapid Profile was to apply it to classroom contexts, in order for the teacher to 1) make more informed claims about the current state of the learner's language development, 2) monitor the learner's actual development and subsequently 3) gear syllabi and formal instruction towards their learnability (see Mackey et. al 1991: 62ff). This diagnostic assessment tool is a logical response to the Teachability Hypothesis (see section 2.2.6 for a more detailed account of the TH).

To elicit interlanguage structures, so-called 'semi-communicative tasks' are used. The notion of communicative tasks will be briefly introduced at this point, as they form part of the assessment setting. Communicative tasks are commonly known from being used in formal language as for instance in Task-based Language Teaching Contexts (see e.g. Van den Branden 2006, Eckerth 2008). Communicative tasks employ a number of features that are thought to be beneficial for communication in language teaching settings (see Ellis 2009: 4f and Van den Branden 2006: 7f. for an overview of task features). Eckerth (2008) describes the core idea of tasks to encourage meaning-oriented language use and target language communication used, in order to achieve a communicative goal. Van den Branden (2006) adds that the tasks invite the language learner to

⁶¹ It is to be noted here that the term 'interview' is in no way meant to represent a question-answer-pattern in the data elicitation phase, but rather a setting in which communicative partners aim to achieve a communicative goal together. This setting has been proven to be beneficial in eliciting linguistic structures that are to be assessed.

primarily act as a language user. These ideas are transferred to language assessment with Rapid Profile, as the language learner primarily tries to achieve a communicative goal. The communication that happens in order to achieve this goal, indirectly triggers the production of linguistic features that are aimed to be assessed. Thus, as opposed to former versions of profile analyses, the data are not elicited through open conversations, but by guided tasks that are aimed towards the production of those linguistic structures in question (cf. Mackey et. al. 1991: 62), i.e. the ones that are hypothesized to occur at different stages of the universal developmental path (Pienemann 1998, 2005). In this scenario, however, the formal objective is not to test grammatical knowledge, but to assess spontaneous spoken language data that contains a rich sample of interlanguage structures (see Pienemann & Mackey 1992: 17). In this way, a most realistic, natural and uninhibited production of speech, i.e. reflection of the current state of learners' interlanguages, is to be expected.

In this context, it might be helpful to introduce the diagnostic task cycle as put forward by Keßler (2008), which takes a closer look at the concepts of tasks for teaching and assessment purposes. On the basis of Willis' (1996) task cycle, Keßler (2008: 301) embedded a diagnostic task cycle which utilizes Rapid Profile for individual feedback and treatment in the L2 classroom.

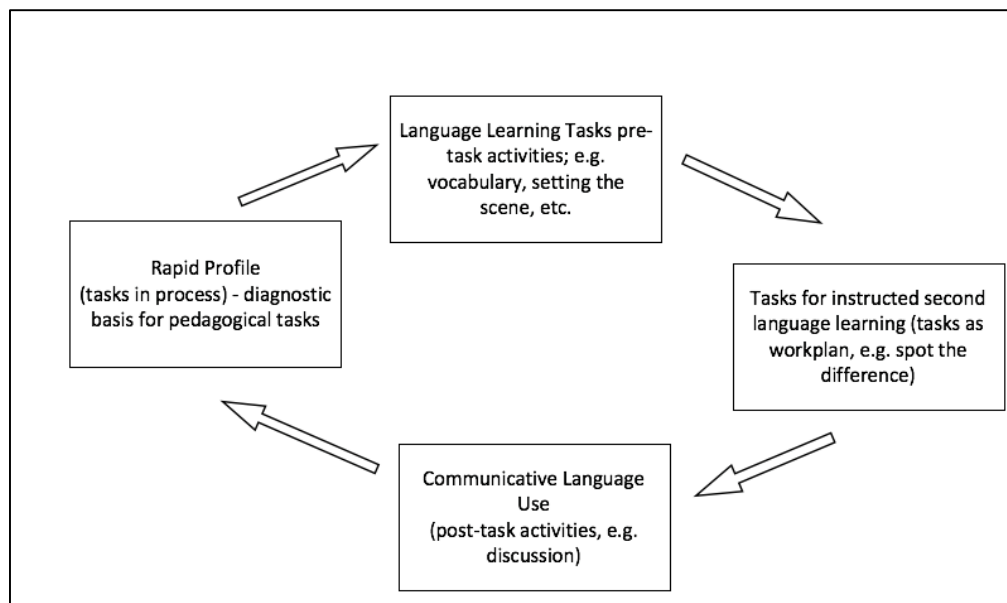


Figure 22: Diagnostic Task Cycle, taken from Keßler (2008: 301)

Keßler argues that in order to be able to support learners according to their needs, it is vital to use the feedback yielded by Rapid Profile (cf. Keßler 2008, Keßler & Plesser 2011: 234). Rapid Profile feedback usually comprises the overall developmental stage and percentages of the acquisition of (noun, pronoun, etc.) morphology as well as syntax. Ideally, this task cycle starts with the box on the left that uses Rapid Profile during tasks in process in order to assess the developmental stage of the learner. This feedback can then be used for establishing a developmentally moderated syllabus (Keßler 2007, 2008) which contains language learning tasks that are found in the box at the top of Figure 22. These tasks may include the introduction of vocabulary that is needed for completing subsequent tasks or introducing content that will be dealt with in concurrent steps. The following set of tasks refers to those that are crucial for instructed second language learning (box to the very right). These tasks may be focused tasks (see Ellis 2003: 16; Keßler & Plesser 2011: 166) with a particular focus on specific language structures, such as interrogatives in spot the difference tasks. Subsequent communicative language use in post-task activities that may also provide an explicit focus on a linguistic item⁶² to be learned can then again be used to base the selection of tasks for the Rapid Profile Assessment on.

During the learner's speech production phase, the profiler uses the computer interface to click the buttons that relate to the structure produced. Figure 23 shows the RP interface of version 4 with the boxes for those structures in the Processability Hierarchy that are indicative of each stage of development.

⁶² Explicit focus on a linguistic structure, in terms of feedback, can be given in various ways. Keßler & Plesser (2011:153ff.) as well as VanPatten (2004) review a number of different forms of implicit and explicit feedback according the Interaction Hypothesis by Long (1983, 1985, 1996). They discuss the role of input enhancement (Sharwood Smith 1993: 176) through consciousness raising (Long & Robinson 1998) with an emphasis on noticing. Their perspective is taken through a task-based language teaching lens.

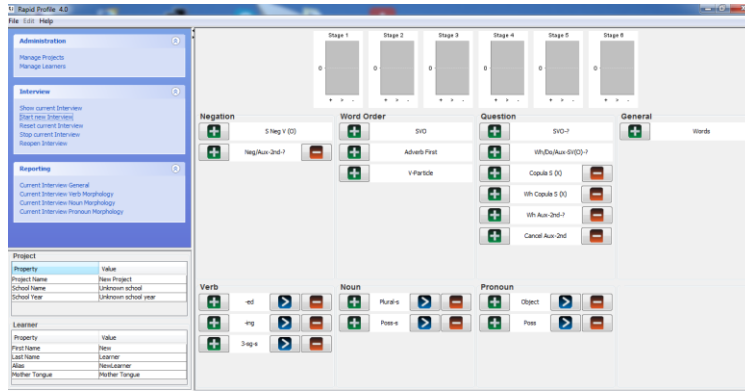


Figure 23: Rapid Profile 4.0 user interface.

The structures in Figure 23 are divided into syntactic phenomena on the top and morphological phenomena on the bottom of the interface. Both types of features are further subdivided into broad grammatical categories, such as negation, word order, or verb and pronoun. When the learner produces a verb in an obligatory context along with a morphological feature, such as the past-ed, the profiler clicks on the '+' button in the category 'verb'. The green '+' indicates the suppliance of a structure in a valid context (Mackey et. al. 1991: 72). Should the learner fail to attach a past-ed morpheme in an obligatory context, the profiler checks the red minus-box. The '-' indicates that the structure is missing, although an obligatory context was present. The blue button ">" indicates the production of overgeneralized forms of morphological features. Mackey et. al. (1991) give an example of overgeneralization with the following sentence "They walks to the park". In principle, the learner can produce the -s with the verb in present tense. However, the learner is not able to distinguish between the fact that it only has to be attached when the number for person is singular and those cases in which it is plural. Instances of over-suppliance of particular morphemes in the learner data provide insights as to the general acquisition of the rule without an informed distinction between the different contexts.

Provided a sufficient amount of data is entered, the program computes the developmental stage by checking it against standard learner language according to the *emergence criterion* (see section 2.2.4 for further information). During the elicitation, the program provides feedback on the amount of structures fed in by the analyst by means of colors. As soon as a sufficient amount

of structures are typed in for the program to assume that a particular stage is acquired, this stage turns black. This way, it is indicated that the assessor does not need to focus on those structures anymore. This kind of feedback can be consulted in order to give information about the still insufficiently supplied number of structures. This might be indicated by stage gaps.⁶³ Keßler (2008: 133) refers to gaps in the Rapid Profile Feedback that occur due to insufficient amount of data density⁶⁴ as diagnostic gaps. These do not violate the implicational hierarchy of processing procedures but are due to missed occasions of the analyst to elicit the structure at question and result from a strict application of the emergence criterion (see Pienemann's discussion on what counts as evidence for more information in section 2.2.4). Rapid Profile gives detailed feedback, not only on the learner's developmental stage, but morphological and syntactical features produced by the learner as well.

A major advantage of RP, as compared to other assessment instruments, is the computer-assisted nature of the program that compares standard patterns of development with a learner's interlanguage sample (Keßler 2006). Trained profilers are able to create a learner profile with high inter-rater-reliability (Keßler 2006: 241). Hence, the use of RP allows for accommodating reliable and valid results in only up to fifteen minutes (Keßler 2006). In his study, Keßler (2006) tested as to whether fifteen minutes were sufficient in order to elicit a dense data set. His results showed that "[...] the data elicitation took an average of 12.5 minutes and ranged between seven and 17 minutes" (Keßler & Plesser 2011: 214) with sufficient data density, and b) determining the PT stage by the author. Keßler (2006: 267) also showed that the semi-automatic nature of the computer program yields a high amount of reliability at 85,7% when the user is sufficiently trained.

Currently, Pienemann & Lanze are working on a dialogue-based automatic version that makes use of some of the principles inherent in Rapid profile that uses an artificial intelligence environment (Pienemann & Lanze

⁶³ To recapitulate, because of the implicational nature of the developmental path, a stage gap cannot occur. The implicational hierarchy assumes that every stage is a necessary prerequisite for the concurrent stage.

⁶⁴ Data density refers to a high amount of linguistic structures present in the language production by a learner as evoked by using communicative tasks (cf. Pienemann 1998, Keßler 2006).

2017). This automatic version overcomes the feasibility issues in Rapid Profile and is applicable in large-scale assessment environments.

In the context of this study, it is argued that the use of PT-based assessment instruments might add to eliciting learner language in terms of grammatical ability when interfaces between the CEFR and PT can be assumed. The next chapter is a theoretical account to finding interfaces between PT and CEFR.

3. Bridging Scales and Stages

It is the aim of this thesis to add to the descriptive, empirical and theoretical basis of the CEFR by proposing to integrate the universal aspects of the PT hierarchy into the scale for Grammatical Accuracy presented in the CEFR. Therefore, I argue that the Scale for Grammatical Accuracy needs to be relabeled into a Scale for *Grammatical Range*. I consider a scale for *Grammatical Range* to be more appropriate than a scale for Grammatical Accuracy because grammatical accuracy does not mirror grammatical development in language learners. Since the CEFR is intended to describe issues relating to language pedagogy, its main concern are language learners. I therefore assume that a scale for *Grammatical Range* that integrates aspects of universal language development as proposed by PT into the CEFR, paints a more learner-centered picture of grammatical development than currently presented in the CEFR.

In this part of the thesis, I will examine similarities and differences in the concepts of language (acquisition) as put forward by the CEFR and PT. I will investigate studies on interfaces between SLA and the CEFR as well as studies that investigate relations between PT and the CEFR more specifically. The remainder of this chapter will constitute a theoretical account to finding interfaces between the CEFR and PT in terms of grammatical competence. As stated above, my proposal is to relabel the scale for grammatical accuracy into a scale for *Grammatical Range*. Therefore, descriptors for accuracy need to be revised and substantiated by levels of processability. I argue that this should be done in a way that the action-oriented approach, as manifested in communicative themes that inform the scale for linguistic range in the CEFR (see chapter 2.1.5), is still

compatible with the new proposed assumptions. The point of departure for my claims is that I view grammatical competence as insufficiently depicted in the CEFR. In the CEFR (2001: 169), it is stated that “[a]ll knowledge of language is partial, however much of a ‘mother tongue’ or ‘native language’ it seems to be. Knowledge is always incomplete, never as developed or perfect in an ordinary individual as it would be for the utopian ideal native speaker.” This quote suggests that an error-free learner, native-speaker-like language learner is utopian. However, when describing grammatical ability in the CEFR, it seems to be conceptualized only on the basis of grammatical accuracy because it is stated in the CEFR that a) the production and recognition of well-formed phrases and sentences (CoE 2001:113) is essential and b) special emphasis is put on accuracy because this is the only scale that the authors of CEFR present for grammatical competence (see CoE 2001: 114). For other qualitative language features, such as vocabulary knowledge, more scales are given. My proposal also adds to the discussion of empirical and theoretical validity of CEFR scales.

I assume that PT has the power to inform the CEFR’s concept of grammatical competence because 1) of its universal hierarchy of processing procedures and 2) modular nature (see chapter 2.2.2). The CEFR is language-independent, therefore any claims about second language development that are supposed to be integrated into the CEFR also need to be language-universal. Ad 1) Currently, there is no theory of second language development that explains the universal developmental schedule in a more consistent, empirically-grounded manner than PT does. Therefore, I assume that PT is able to inform the CEFR in the area of grammatical development. Ad 2) the modular approach taken in PT that puts the processing of linguistic features at the center, is able to be integrated into the CEFR because it focuses on only one discrete subtask of SLA and the CEFR is structured in a way that it describes several subtasks (see chapter 2.1.2). While I do not argue that all of the ideas in the CEFR are compatible with PT, I assume that the CEFR is open enough to embrace features of language processing as proposed by PT (see chapter 2.2.2). Moreover, I assume that more sophisticated grammatical assessment based on the CEFR can be employed when

taking PT's implications for assessment into account and when CEFR scales and PT stages are aligned.

Before I set out for the analysis of interfaces between PT and the CEFR, the following should be made clear: I am well aware of the fact the CEFR is a framework for describing language competences holistically whereas PT is a psycholinguistic SLA theory that takes a parsimonious approach to explaining morphosyntactic development. I do not intend to treat either framework as being equal to the other. Rather, I assume that PT might be seen as complementary to the CEFR.

In order to investigate whether a scale for *Grammatical Range* that combines features of PT and the CEFR is theoretically acceptable, I first review some studies from an SLA viewpoint that seek to find interfaces between SLA and the CEFR. These studies have varying foci and are all language-specific. In chapter 3.2, I will report on studies that have been conducted within PT framework and that explore interfaces between PT and the CEFR. After having reviewed these studies, I will deal with some general issues in aligning the CEFR and PT from a theoretical perspective. These mainly encompass language universality in the CEFR and PT, differences in approaches to emergence and accuracy, and second language development. It needs to be stressed here that it is not the aim of this study to present a full account of a PT-informed version of the CEFR's *grammatical competence*. I do not seek to propose a PT-based version of *competence* or *proficiency* here. Rather, exploratory assumptions are made. All of these assumptions would need to be investigated in more detail in theoretical accounts and empirical studies.

3.1 Selected Studies on SLA and the CEFR

There is quite a large number of studies that investigate interfaces between the CEFR and second language acquisition in general (see e.g. Carlsen 2010; Hawkins & Buttery 2010; Hulstijn et al. 2010; Bartning et al. 2010; de Jong et al. 2012; Crossley & McNamara 2012; Thewissen 2013; Abel et al. 2014; Gyllstad et al. 2014; Díez-Bedmar 2015; Chen & Baker 2016, Wisniewski 2017b). These studies

take different viewpoints to investigating interfaces between the CEFR and SLA. A lot of criticism of the CEFR has been pronounced from an SLA viewpoint. This criticism addresses mostly the following issues 1) the lack of theoretical and empirical validity in the CEFR scales, 2) that the scales do not reflect language development and 3) that the descriptors are often inconsistent (see Wisniewski 2017a and sections 2.1.2, 2.1.3, 3.3.3.1 for more detail). Most of the studies that aim to validate the CEFR scales make use of learner corpora. Some of the studies from the SLA community will subsequently be reviewed.

Hulstijn et al. (2010) published a paper with the title “Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them?” Despite the promising title, the paper only describes the CEFR and states the research questions and goals by the SLATE (Second Language Acquisition and Testing in Europe) group. No reference to any research results is given. Interestingly, they do not even mention the developmental stages, as explained by PT, but rather focus on the description of proficiency testing scales. This paper does not give any conclusions for the relationship between SLA research and proficiency testing relevant for this thesis.

Prodeau, Lopez & Veronique (2012) present a study on an L2 developmental sequence for French that is based on a review of research by Bartning & Schlyter (2004), Veronique et al. (2009), Prodeau (2009) and Granget (2009). Their aim is to investigate the role of grammatical knowledge in the CEFR based on L2 French morphosyntactical features. Prodeau et al. (2012: 52) describe the general sequence for French as an L2 as follows:

Auxiliaries and modals are first to mark agreement with subjects. Subject-Verb (S-V) agreement takes place at the time when all conjoint pronouns are used instead of disjoint ones. The preverbal position is no longer used for the topic but from then on for the subject. S-V agreement is the first step towards full inflectional verbs. Another key moment is when complementizers are no longer left implicit.

Prodeau et al. (2012) investigate a corpus of 40 learners of French whose proficiency levels were elicited by the Test de Connaissance du Français. The test is aligned to the CEFR. To what extent the alignment is done, is not stated. Only

the written production⁶⁵ of the test was considered. It is, however, not made explicit who analyzed the written production in the methodology. Prodeau et al. (2012: 52) argue that the CEFR descriptors for grammatical accuracy propose a grammaticalization process to take place at levels B1 and B2. They further describe that at levels A1 and A2, simple structures in a learned repertoire are mainly used by language learners (see scale for grammatical accuracy, chapter 2.1.6). This is why they decide to only focus on the grammatical accuracy of CEFR levels B1 and B2 for their analysis (Prodeau et al. 2012: 53). Prodeau et al. (2012) use the automatic analyzer *Direkt Profil* (Granfeldt et al. 2005) as a means to assess the level of written French from an SLA perspective; i.e. by investigating a developmental sequence for French. The results of their study show that no distinction between the accuracy levels B1 and B2 can be made based on the morphosyntactic development of French L2 learners because all morphosyntactic features found at B2 level are also present at B1 level. They infer that the blurred line between B1 and B2 is related to the gradual nature of grammatical development because each level seems to result from a coalescence of mastered features and errors (Prodeau et al 2012: 63). The only difference that they found between B1 and B2 was the length of texts produced by the learners.

It may have been beneficial to investigate a broader range of CEFR levels than only B1 and B2 to link the developmental sequence for grammatical development of French to the CEFR. From this study, it is problematic to infer links between morphosyntactic development and CEFR levels because it cannot be ruled out that the linguistic features investigated might have already been present in learners at the A2 level. However, what I assume can be inferred from their study, is to question the progression of language learners assumed in the CEFR descriptors as investigated using Rasch Item Response Scaling (see chapter 2.1.1 for details).

In a longitudinal study, Gyllstad et al. (2014) examine 120 written texts produced by Swedish learners ranging from A1 to B2 levels learning English,

⁶⁵ The authors state that six different questions were answered by the learners, each of which defines a task that is linked to a level in the CEFR (Prodeau et al. 2012: 53). However, they do not further specify the questions, tasks or learner answers.

French and Italian. Their aim is to add to the empirical basis of the CEFR in finding linguistic correlates for the CEFR scale for overall written interaction. Learners were asked to write a letter and a short narrative. These were collected on three occasions over a period of several school years (early, intermediate and late in their school career). Eight CEFR raters (seven experienced raters and one teacher who is familiar with the CEFR) rated the texts. Gyllstadt et al. (2014) measure syntactic complexity based on length of t-units, subclause ratio, and mean length of clause as suggested by Norris & Ortega (2009). Gyllstad et al. report medium to strong correlations between the CEFR levels and the measures for syntactic complexity (2014: 16) for all learners in general. In fact, they report a “linear positive significant correlation between all three measures of syntactic complexity and rated CEFR levels” (Gyllstad 2014: 22). However, they found that syntactic complexity does not vary much at the A levels whilst the data of the learners at the B1 level showed an increase in syntactic complexity for mean length of T-unit and mean number of subordinate clauses per T-unit. They further observe (2014: 23) that B1 is the level where the CEFR authors start referring to complexity instead of using terms such as *simple* and *basic* at lower level. This shows that concepts occurring at higher CEFR levels do necessarily appear at lower CEFR levels. Gyllstad et al. (2014) provide a carefully designed study and make all of their sources and scales available to the reader. However, it seems to be the case that the scale that they use is again a descriptive scale of the CEFR but no rating grid. The authors of the CEFR make explicit that the descriptive scales are no rating grids. Rating grids need to be based on the descriptive grids but have to be produced especially for the purpose of language assessment (see North 2014).

Thewissen (2013) investigates 223 English learner essays taken from the International Corpus of Learner English in a quasi-longitudinal study. She reports that her aim is to show how SL accuracy developmental trajectories can be captured via an error-tagged version of an EFL learner corpus. Her corpus was manually error coded and the learner samples were rated by two to three experienced raters involved in the assessment of writing at the University of Cambridge Local Examinations Syndicate (UCLES). The raters assigned scores for

the domains of Vocabulary and Orthographic Control, Grammatical Accuracy, Vocabulary Range and Coherence and Cohesion (Thewissen 2013: 79). Thewissen uses the potential occasion analysis⁶⁶ to quantify 40 different error types in the learner data that were rated on the CEFR levels B1 to C2. She states that some of the errors seem to discriminate between adjacent CEFR levels whereas others are produced throughout the learners' progression through CEFR levels. She (2013: 95) describes that

a total of 33 error types (i.e., 72% of the total errors) did not display any marked change between B2 and C2. The results suggest that, within the B1 to C2 range, development in accuracy is most marked between the lower and upper intermediate levels, hence pointing to a possible accuracy threshold at B1, that is, a level after which accuracy can generally be said to remain stable. Conversely, accuracy is a less strongly discriminating feature at the higher B2 to C2 levels.

Thewissen (2013) provides an interesting account of English learner errors that help to distinguish between adjacent CEFR levels. The most interesting finding in this context, is that there seems to be an accuracy threshold at level B1. To me, it remains unclear to what extent the assessment grid used by the raters is aligned to the CEFR descriptors for linguistic range because no analyses as regards the level of fit between the grid and the CEFR levels is stated in her study. That is to say, it is not stated in her study, in how far and in which way the grid that was used by the raters, was aligned to the CEFR levels.

Díez-Bedmar (2015) uses a combination of frequency and accuracy measures to investigate errors in 26 written texts by Spanish learners of English, taken from a local academic corpus. In her study, she explores the article use of Spanish EFL learners in relation to how the learners were assessed on the basis of the CEFR. In the written texts, answers to the question "Where, outside Spain, would you like to go on a short pleasure trip?" (Díez-Bedmar 2015: 172) were given as part the University Admission Examination (Díez-Bedmar 2015: 166). These texts were subsequently rated by two independent raters according to the CEFR and only the texts with a 100% inter-rater-reliability were chosen for

⁶⁶ Thewissen (2013: 81) explains that the potential occasion analysis counts "[...] errors in relation to the number of times a learner could potentially have committed such an error; for example, modal auxiliary verb errors are best counted out of the total number of modal auxiliaries used, as these are potential occasions for error."

analysis (Díez-Bedmar 2015: 172). The texts are located at CEFR levels A2-B2. Díez-Bedmar explores the uses of definite, indefinite and zero-articles in the corpus and applies Bickerton's (1981) semantic wheel and Huebner's (1983) taxonomy to the article system of the learners for the analysis of her corpus. Bickerton and Huebner propose that the semantic and discourse-pragmatic features *specific reference* and *hearer knowledge* give rise to study the article system.⁶⁷ Therefore, they identify contexts of article use, namely generics (context 1), referential definites (context 2), referential indefinites (context 3) and non-referentials (context 4) (Díez-Bedmar 2015: 164). In Díez-Bedmar's study, the learners do not produce definite and indefinite articles in generic contexts as they seem to favor generic contexts in which the zero article is used (Díez-Bedmar 2015: 186). However, in her careful analysis, she finds a general decrease of errors regarding article use with higher CEFR levels: error percentages at A2 level amount to 18.42%, at B1 the percentage level is 13,48% and at B2 level she finds only 5,48% (Díez-Bedmar 2015: 178). More specifically, she reports a significant decrease in the use of the definite article in obligatory contexts and an increase in the accuracy of use of the zero article in non-referential contexts at CEFR level B2. These results seem to somewhat converge with Thewissen (2013), who also found that level B1 seems to be an accuracy threshold. Therefore, she argues that CEFR level B2 comprises a number of criterial features: a) more NPs with articles are present at this level in comparison to the other levels, b) the zero article is significantly more often used than at lower levels and c) the zero article is effectively selected in non-referential contexts with plural nouns; i.e. in contexts where the NP has no specific referent (Díez-Bedmar 2015: 178). She concludes that the accurate use of the correct article is a criterial feature of the B2 level. Unfortunately, Díez-Bedmar does not

⁶⁷ Geng (2010: 180) explains that Bickerton (1980) proposes universal features of referentiality "[...] namely, whether or not the noun phrase has a specific referent and whether or not it is assumed known to the hearer. Hence, noun phrases are classified as plus or minus the feature of specific referent ([±SR]) and plus or minus the feature of assumed known to the hearer ([±HK]). The four combinations of the two binary features constitute what Huebner calls semantic types. In Huebner's model, the use of English articles is determined by the semantic function of the NP in discourse. Each NP belongs to one of the four types/categories, permitting us to assign a semantic function to each NP. To determine with what accuracy articles are used, one considers what is used in Standard English."

state which CEFR scale was used in her study. Since the CEFR scales are no rating scales but descriptive scales, it would have been important to know which grids were used. Especially, since North (2014) reports that being a certain level on one scale does not mean that learners are on the same level on other scales, it is problematic to assume that certain features are criterial for CEFR levels in every scale. Apart from the rather outdated use of measures connected to morpheme order studies (e.g. Dulay & Burt 1974), Díez-Bedmar (2015: 185f.) points to further limitations of her study herself. They are mainly concerned with the corpus approach that she uses: the limited number of texts per CEFR level that were restricted to only one topic may compromise generalizability. She proposes that a combined approach of elicitation and a learner corpus might broaden the contexts of article use.

Williams' (2007)⁶⁸ unpublished study aims at defining grammatical criterial features that distinguish between the different CEFR levels. To identify the criterial features, she uses the Cambridge Learner Corpus (CLC) that consists of approximately 39 million words of written learner language. The written data derive from Cambridge ESOL Examinations (KET to CPE) that assess language proficiency and are that are aligned to CEFR levels A2 to C2 (Williams 2007, as cited in Salamoura & Saville 2010: 104f.). Williams (2007: 4) points out that the CLC has been partly manually error-tagged and corrected so that researchers can investigate "[...] what learners wrote and what they should have written". Also, the CLC is annotated with a pass and fail grade for the CEFR level that the test aims at (Williams 2007: 4). Williams (2007) investigates verb co-occurrences using the Briscoe-Korhonen subcategorization frame.⁶⁹ In computational linguistics, natural language processing tasks are often quantified in subcategorization frames. These frames are defined as syntactic frames that

⁶⁸ Nick Saville kindly provided the manuscript of this study. In a personal conversation, he pointed out that Williams's manuscript is still a draft version because she was not able to finish her research entirely. Therefore, the manuscript contains some issues that would have been resolved if she had been able to finish it. These issues mainly concern the unusual way of presenting raw scores of her data as well as misunderstandings about the technology that was used.

⁶⁹ Buttery & Caines (2012) describe that subcategorization frames are large-scale verb lexica that specify verb usages as probability distributions. These computational models are often used in psycholinguistic and neurolinguistic research.

consist of the number and type of arguments of predicates (Buttery & Caines 2012). Williams (2007: 1) explains that subcategorization frames describe “[...] a particular set of restrictions on the number, order and type of syntactic feature required by a particular head, in this case a verb” so that the subcategorization is a generalization over different syntactic contexts which a verb might take. In her study, Williams investigates 10 different verbs (arrive, buy, interest, like meet, need, see, think, visit, write) because they occur at least 100 times at each level. This is considered a minimum need for analysis (Williams 2007: 4). The following new verb co-occurrence frames are found at the B2 level by Williams (2007), as reported by Salamoura & Saville (2010: 116):

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-Vinfin (Subj control)	He set out to win

Table 2: New Verb Co-occurrence Frames at B2 level, taken from Salamoura & Saville (2010: 116)

These new verb co-occurrences at the B2 level serve as an example, see Salamoura & Saville (2010: 166ff.) for specifications of verb co-occurrences at CEFR level B1 and B2. Williams reports that most new verb co-occurrence frames are learned at the B2 level. Moreover, she states that the range of verbs used in each frame increases until level C1. However, there is not necessarily an increase in the use of frames in general when the CEFR level increases (Williams 2007: 25). Williams (2007: 19) also discovers that not all structures that appear at B2 level continue to be used after their first occurrence in the data. While Williams explains that this might be due to the data set itself, it would imply that some linguistic structures on lower CEFR levels are not necessarily present at later

levels. This finding would contradict the idea of progression in the CEFR scale as validated in the study by North (1998) by means of the Rasch Item Response Model. Williams (2007: 26) herself points to a number of limitations: a) the study relies on first appearances of the verb frames in the learner data and it is unclear if one appearance resembles the acquisition of that verb frame, b) verb frames that only occur with certain verbs for which there were no obligatory contexts, were not present in the data, c) the comparison of her data to a native speaker corpus might not mirror learner language use. Another issue in her study is that the subcategorization frames have been manually corrected by the analyst so that the classifier could be applied to the data (Williams 2007: 4). Williams gives the following example: *he said me that he enjoyed it* would be corrected to *he said to me that he enjoyed it* (Williams 2007: 4). So, the analysis of the data actually relies on what is assumed that the learners might have produced. Also, Williams (2007: 4) points out that the data stem from a variety of registers, genres and topics which highly influence the preference of subcategorization frames in all of these. Williams (2007: 18) reports that the data are not individualized and cannot be reconstructed entirely. Both of these issues would be central for a study from an SLA viewpoint. If the aim is to identify features at each CEFR level, it might be more suitable to look at each set of learner data individually and have it rated according to the CEFR afterwards. This is why I assume that Salamoura & Saville (2010: 125) draw a dangerous conclusion in stating that “[t]he emerging performance patterns per CEFR levels 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 [...]”, when they report on Williams’s study. Linguistic patterns have been ascribed and aligned to the CEFR levels in post-hoc fashion. In my view, this post-hoc fashion cannot give rise to the emergence of linguistic features in language learners or compete with a theoretical framework that proposes SLA developmental schedules based on empirical research. Since Williams’ (2007) aim is to identify the subcategorization frames in the different CEFR levels to validate them, it might be more promising to examine the learner data first and then see which CEFR levels result from them. It might have been promising to investigate the learners individually and

not to group their data together. Furthermore, the linguistic properties of the subcategorizations were not stated explicitly, which impedes the transparency of the results. Also, her study is not theoretically motivated but remains purely descriptive.

Wisniewski (2017a: 4) points to some issues relating to the corpora approach that was taken in the studies presented above. She maintains that their classification methodology is potentially imprecise and that many learner corpora are dependent on external criteria, such as the school year or type. She also argues that the testing situations learners find themselves in might have repercussions for their performance (Wisniewski 2017a: 4). If validation studies use these corpora, such as the CLC that is an accumulation of Cambridge ESOL examinations which were aligned to the CEFR, parts of the validation of the levels might be guarded by the testing situation. Even more so, Williams (2007: 26), who also used the CLC corpus in her study, states that it is expected that about 80% percent of the learners in the corpus have undergone a special training course for the ESOL Examinations and that it is likely that “[...] candidates will have been drilled repeatedly in structures which are necessary to answer various questions, or which are perceived as “advanced” by the examiners.” If Williams suspicion was true, and some kind of ‘teaching to the test’ had happened, one might wonder to what extent the data actually resembles natural, productive learner language. Despite the great potential of learner corpora for validating CEFR levels, Wisniewski (2017a: 4) identifies three major constraints on the generalizability of their results: a) the corpus size, b) the range of texts and c) the accessibility of the corpus for replication studies. It also has to be borne in mind that most of the corpora available at present consist mainly of written texts.

Wisniewski (2017b) provides a most interesting account of the validation of CEFR scales on vocabulary and fluency scales (A2-B2). She argues that the examination of ratings is not a valid means to empirically investigate the empirical robustness of CEFR scales. She states that the use of rating procedures test the behavior of the raters instead of the scale itself, and points to the flaws

that have been found in the reliability of rating procedures⁷⁰. Rather, she proposes (2017b: 4) to investigate the scales in their own right and look at validity through operationalized CEFR scales so as to avoid human ratings. She proposes that only observable behavior should be present in the descriptors and therefore deletes all descriptors that do not match this criterion. In her study, she uses the following methodology: 19 oral productions of South Tyrolean language learners of English who performed a dialogue and judgement task were selected. Their oral data were rated by two independent raters in terms of the CEFR vocabulary and fluency scale. The productions were transcribed. Additionally, the CEFR descriptors were operationalized. Wisniewski (2017b: 7) exemplifies the operationalization process by stating that if “[...] the scale claimed that it was typical for a learner not to show breakdowns in communication, the scale variable would count those breakdowns (normalized, i.e. per utterance and word token).” Scale variable in this case stands for the operationalized CEFR scale.⁷¹ In addition, the transcription of the audio-sample was annotated for, e.g. mean length of runs or phonation-time ratio by two independent coders. Following this, statistical measures were run to investigate a) the observability of the scales (by means of relative frequencies of AS⁷² units), b) the consistency of level descriptors (by means of Pearson correlations) and c) the link between scale variables to constructs in the scale (by t-tests) (Wisniewski 2017b: 7f.). Her results show many shortcomings in the CEFR scales for vocabulary and fluency in that many of the emphasized descriptor items were not or hardly measurable in the learner productions (e.g. the pauses in the learner data as an indication of fluency (Wisniewski 2017b: 8). She concludes that the suitability of CEFR descriptors to describe L2 competence is often overestimated and that this is often dangerous when learners’ life decisions depend on the CEFR scales (Wisniewski 2017b: 19);

⁷⁰ Wisniewski (2017b: 5) summarizes that flaws in the rating approach to scale validation comprise, for example, that raters do not necessarily refer to scale for their rating or that they are often intuitively used.

⁷¹ Wisniewski (2017b: 7) admits that the operationalization processes was not possible without a degree of interpretation but that subjective descriptors, such as “regular interaction with native-speakers quite possible” in the B2 fluency scale and self-referential descriptors, such as “interact with a degree of fluency” in the B2 fluency scale were deleted in this process.

⁷² An Analysis of Speech unit (AS unit) is referred to as a main clause and any attached subordinate clauses or sub-clausal units.

such as granting visa or access to educational classes. Wisniewski's study does not focus on the scale for grammatical accuracy, but I consider her conclusions about the suitability of the CEFR descriptors important and, most likely, generalizable to other scales as well. This is why her study is presented in this context.

Concerns about SLA-based studies on the CEFR have been expressed by Salamoura & Saville (2010: 107). They identify three major caveats: 1) the reliable identification of the proficiency level of the study participants, 2) the varied use of terminology for level description (which is not always adequately defined) and 3) the degree of generalizability and comparability of findings across different SLA studies. It is important to bear in mind that this fact makes the comparison of the results of the present study to other studies quite difficult.

Research on defining *criterial features* for English to exemplify the CEFR levels and to better distinguish between the CEFR levels has become quite popular recently. CoE (2011: 6) define criterial features as those “[...] language features concerned serve as a basis for distinguishing one proficiency level from another.” Hawkins & Filipović (2012: 11) further explain that criterial features are “[...] properties of learner English that are characteristic and indicative of L2 proficiency at each of the levels and that distinguish higher levels from lower levels.” These studies are helpful in determining the range of grammatical structures that learners usually produce at the different CEFR levels. In contrast to the assumptions put forward in this study, criterial features are obviously language-specific and are investigated mainly to aid language testing. What, to my knowledge, has been elusive is to find interfaces between language-independent; universal patterns of language acquisition and the CEFR. This is, I assume, the biggest advantage of shedding light on interfaces between PT and the CEFR in the present study over the language-dependent studies presented above. However, there is only a small body of research available that examines the CEFR and PT. Also, the studies that investigate the CEFR and PT are concerned with either testing/inter-rater reliability issues (see e.g. Michalska 2010, Hagenfeld 2013) or the scope-precision dilemma (Lenzing & Plesser 2010) (see chapter 2.4) rather than specifically focusing on aligning CEFR levels and PT

stages.

This section discussed SLA studies that focus on the CEFR with varying foci. The studies do not present any conclusive results with regard to interfaces between SLA and the CEFR because each studies subsections of language-specific linguistic features. The studies are purely descriptive, language-specific and not theoretically motivated. Despite the different foci, a finding that Thewissen (2013) and Díez-Bedmar (2015) seem to concur with is that the B1 levels seems to be some sort of threshold for accuracy development. What the studies presented above fail to provide is a theoretically motivated, language-universal account to SLA and its interface to the CEFR. The following chapter focusses on studies within the language-universal PT framework and its relationship to the CEFR.

3.2 Prior Studies on the CEFR and PT

Lenzing & Plesser (2010) explore correspondences between CEFR levels and Rapid Profile stages in order to challenge the scope-precision dilemma (see chapter 2.4). They investigate a total of 40 learners of English, 20 early and 20 advanced learners. Their oral speech data were rated according to the CEFR by one rater and the PT stages of the learners were determined using Rapid Profile.⁷³ Lenzing & Plesser's results on the relationship between PT and the CEFR can be depicted as follows:

⁷³ Lenzing & Plesser (2010) also examine written data and compare written PT stages and CEFR levels for written performance. Generally, they state that with written language, the results are more diverse than with oral language. I will not report on those results in detail at this point because my study only focuses on oral learner language and thus their results on oral PT-CEFR relations are more important in this context.

PT stage	CEFR level
1	Below A1
2	Below A1 A1
3	A1
4	A1 B1
5	B1 B2 C1
6	C1

Table 3: Relationship between PT stage and CEFR level found by Lenzing & Plesser (2010)

Lenzing and Plesser (2010) state that the relationships for PT stages 1, 4 and 6 need to be treated with caution because they are only based on one or two samples each. They suggest that learners who have reached a B1 level in the CEFR, are assessed PT stage 5 or higher. I assume that this implies that the communicative ability at early PT stages is too narrow to be captured by the descriptor items in the CEFR scales. This is one claim that will be investigated more closely in the present study. It should be noted that only one rater participated in Lenzing & Plesser's study and it is not fully clear as to which assessment grid was used.

In Hagenfeld (2017), I investigated the feasibility of Rapid Profile and Autoprofiling (Lin 2012) for language assessment in a small-scale study. As part of this study, I examined 8 learners of English whose CEFR level was certified by the ZfS at Paderborn University. The learners had completed language classes that aim at different CEFR levels at the time of the study. When the students pass this class, they are certified at the respective CEFR level. Two students each at CEFR level: B1, B2, C1 and C2 participated. Their oral performance was assessed with Rapid Profile and Autoprofiling. Each of the students reached PT stage 5 in the assessment with Rapid Profile.⁷⁴ It is to be noted that the elicitation did not

⁷⁴ The results yielded by the Autoprofiling analysis are a little more diverse. I argue that this difference is mainly due to lack of experience in keyboard type-writing apparent with some learners.

aim at PT stage 6, so some students might have reached a higher stage if the contexts for the stage 6 structure were present. The results partly converge with Lenzing & Plesser (2010) in that learners at PT stage 5 are generally rated on CEFR level B1 or higher.

Keßler and Plesser (2011: 236) report on an unpublished study by Michalska (2010). Michalska’s (2010) aim is to compare the inter-rater reliabilities of Rapid Profile analyses and ratings using CEFR grids. Keßler & Plesser (2011: 236) report that Rapid Profile “[...] scores higher in terms of inter-rater reliability as only one learner was rated differently by two assessors who used Rapid Profile as compared to 12 varying ratings conducted by the raters when using the CEFR.” Keßler & Plesser (2011: 236) present a table to display the raters, CEFR levels and PT stages, adapted from Michalska’s study. From their table, I summarized information on the overall PT stage of the study participants and CEFR below:

Learner	CEFR level	PT stage
1	B2	5
2	B2; C1	6
3	B1; B2	5
4	B2; C1	5
5	A2; B1	3
6	B1; B2	5
7	B1; C1	5
8	A2; B1; B2	5
9	B1; B2	4;5
10	B1; B2	3
11	A1; A2	3
12	A2; B1	4

Table 4: CEFR levels and PT stages by Michalska 2010, presented by Keßler & Plesser (2011:236), adapted and modified

In the study, Michalska had 12 learners rated by four different CEFR raters and two Rapid Profile analysts. Although it was not the aim of Michalska’s study to

investigate correlations between PT and the CEFR, and although it is not clear as to which assessment grid was used, there seem to be some relations between the PT stages of the learner languages and their assigned CEFR levels. The results show that participants on PT stage 3 were assigned CEFR levels ranging from A1-B2. The learner at PT stage 4 was assigned CEFR levels A2 and B1. Learners at PT stage 5 were rated on CEFR levels ranging from level A2 to C1, whereas in 11 of the 13 cases, the B1 and B2 levels were assigned (5 times B1 and 6 times B2). Participant 2 at PT stage 6 was rated on CEFR levels B2 and C1. One can see from the results that there are relations between the CEFR levels and Rapid Profile stages. Keßler and Plessner (2011: 236) infer from Michalska's results that "even when rating more communicative skills of the learners as aimed at by the CEFR, raters are not completely free of the underlying grammatical structures produced by the learners". This observation is taken as point of departure for the methodology, described in section 4.3 in the present study, because it will make explicit that grammar is an underestimated component part in rating oral language production. My study explores this phenomenon further (see chapter 4).

What the studies presented above have in common, is that they report that for PT stage 5, generally the following three CEFR levels are assessed: B1, B2 and C1. The other results of the studies investigating PT and the CEFR vary strongly and are somewhat incomparable. This might be due to the different study designs that partly aim at testing hypotheses other than finding CEFR-PT interfaces. Also, only sparse information on which assessment grids were used in the studies are presented. Moreover, Michalska (2010) reports on compromises in inter-rater reliability that might lead to the inconclusive results. What becomes evident from the studies is that there seems to be an interface between PT and the CEFR which should be investigated in more detail.

The next chapter will discuss an integrative account to the CEFR and PT in more detail that will support the proposal of a scale for *Grammatical Range* in chapter 4.4.6. The chapter will discuss chances and challenges in combining PT and the CEFR from a theoretical perspective.

3.3 Grammatical Range – An Integrative Approach to the CEFR and PT

Green (2012: xl) admits that the CEFR is rather unspecified in terms of grammar and lexicon. He states that the reason for this is the different language-specific shapes of the L2s and the L1s:

[...] the CEFR may appear to be underspecified in respect of grammar and lexicon. No-one would deny that, for any particular language, grammatical and lexical progression is of central importance and not merely a secondary consequence of notational-functional progression. However, the two are intimately related, so that exclusive attention on the one may seriously distort the other, as in the case above of refusing to speak of the past until the past tense was introduced after more than two years of the study. The necessary reconciliation has to be made and the optimal progression has to be established separately for each target language (L2) in turn and in principle for each source language.

I argue that by using the universal processing procedures as put forward by PT, Green's statement about establishing optimal progression for each target language can be circumvented, as the processing procedures are language-independent. However, for a combined version of PT and the CEFR, the scale for Grammatical Accuracy needs to be revised because PT is not concerned with accuracy but language development. Therefore, I propose to call the combined scale based on PT and the CEFR *Grammatical Range*.

In order to propose a combined scale for *Grammatical Range*, it is necessary to reconsider the specifications for grammatical competence in the CEFR in this context. This chapter will briefly revisit the concept of grammatical competence taken in the CEFR. It highlights the potentials of bringing PT and the CEFR together and raises some challenges in combining both approaches. I argue that grammatical ability, apart from accuracy, might also be conceptualized in terms of universal processing operations that explain interlanguage development in a more learner-centered way.

I want to stress that it is not the aim of this thesis to provide a full theoretical or empirical account to specifying a PT-based version of grammatical competence for the CEFR, because this would have to undergo extensive further quantitative and qualitative studies. Rather, this thesis wants to add to the discussion of combining SLA research and the CEFR in proposing a combined

theoretical and empirical perspective to grammatical ability while being aware of the different theoretical perspectives inherent in the competence debate itself (see chapter 1). Thus, this work attempts to find a practical solution for the empirical validation of the CEFR. I do not aim to hypothesize a PT-informed version of competence, because I do not consider a definition of competence situated within the PT framework possible because of PT's declared modular nature.

In section 2.1.6, I have argued that the CEFR paints a problematic picture in only proposing a scale for Grammatical Accuracy for grammatical competence in the quantitative dimension of the CEFR when all other linguistic competences are presented in a more detailed way. Although the qualitative part of the CEFR specifies grammatical competence in a more detailed way, Little (2014) states that the scales are probably the most well-known part of the CEFR. Therefore, it might be the case that language professionals tend to the scales before considering the elaborations of the notions behind the scales in the qualitative part. This is why I argue that the scale for grammatical accuracy needs to be revisited first and foremost. I therefore consider it necessary to clarify as to why I suggest relabeling the scale for *Grammatical Accuracy* into *Grammatical Range*.

Since the CEFR is suggested to be used by language professionals and to promote life-long learning (see Morrow 2004), laying out only a scale for Grammatical Accuracy might lead the reader to assume that accuracy plays a primary role in the acquisition of grammar by language learners. As stated above, numerous studies have shown that learners, during the acquisition process, develop an interlanguage. This is a system independent of the first language and the target language. Parts of these systems are necessarily grammatically inaccurate during their development (see chapter 3.3). Pienemann (1998) was able to show that interlanguage shapes, i.e. learner development and variation, are steady and can be captured by Hypothesis Space within PT. A combined CEFR-PT account to grammatical ability then calls for a term other than accuracy because of these necessarily inaccurate features during the language acquisition process. In my view, *range* is a suitable term as the Oxford Dictionary provides (amongst others) the definition: "The area of variation between upper and lower

limits on a particular scale” and “The scope of a person's knowledge or abilities.”⁷⁵ The term *range* also does not imply anything about the relationship between linguistic knowledge and its performance (see the discussion on competence and performance later in this chapter for more detail). In this way, I assume that developmental and variational dimensions (as spelled out by PT), as well as the scope of possible themes (as indicated in the CEFR), can be captured sufficiently. The following chapter focuses on the term *competence* in the CEFR and briefly introduces the claims that PT makes which can be related to competence. Again, PT does not focus competence but on processing and development and therefore, the chapter on competence in the CEFR is necessarily longer than the one on competence in PT.

3.3.1 Differences in the Frameworks: Universality, Emergence, Accuracy

One reason for proposing to combine PT and the CEFR is that PT assumes universal processing procedures that underlie the acquisition of second languages. Research within the PT framework has studied more than 12 typologically diverse languages and found striking evidence for the existence of universal processing procedures for typologically diverse languages (see e.g. Johnston 1985; Pienemann & Mackey 1993; Mansouri & Duffy 2005; Pienemann et al. 2006; Keßler 2006; Roos 2007, Pienemann & Håkansson 1999; Håkansson 2005; Håkansson & Norrby 2007; Pienemann 1998; Di Biase & Kawaguchi 2002; Kawaguchi 2005; Zhang 2005; 2007; Ghassan 2008; Di Biase 2008; Özdemir 2004). Because of the universal patterns, I assume PT to be compatible with the aim of the CEFR to be applicable to all (European) countries, and languages respectively. Despite the universality aspect in both frameworks, PT and the CEFR

⁷⁵ I consider *scope* to be another appropriate term for a combined CEFR-PT approach to grammar. However, *scope* might point towards the scope-precision dilemma that Pienemann & Keßler (2007) proposed. The scope-precision dilemma describes the problem of large-scale proficiency ratings in that they often lack precision in assessment, whereas acquisition-based measures often fail to be applicable to large-scale assessment because they are very precise and thus time-consuming. I, however, argue that given appropriate tasks, a combined assessment procedure is possible that overcomes scope-precision issues. Because of this discussion, I prefer the term *range* to *scope*.

make very different philosophical assumptions. These, of course, are grounded in the fact that PT is a theory that aims to make predictions about second language acquisition, whereas the CEFR is a descriptive, operational framework of language (see chapter 2.1.1). With the Multidimensional Model Meisel, Clahsen & Pienemann (1981, as cited in Pienemann 1998: 141) assume social variables to interact with interlanguage variation and not with L2 development itself. This forms a highly different basis to conceptualize language competences as compared to the CEFR, because the CEFR puts the learner as a social agent in the center of the discussion, and all assumptions about language (and language acquisition) would follow from this concept of a social agent and the action-oriented approach. Thus, the CEFR does not make a distinction between the two dimensions of development and variation as found in PT (see chapter 2.2.3). Rather, these two dimensions remain somewhat blurred in the CEFR and its conceptualization of competence. I assume, however, that this fact would not lead to an incompatibility of PT and the CEFR because of PT's modular approach. As a modular approach, PT has the potential to be extended to more pragmatic language production (see Nicholas & Wigglesworth 2003: 135 and Nicholas & Starks forthc.).

Another factor that needs to be considered when aligning PT and the CEFR is the difference in acquisition criteria. As laid out in chapter 2.2.4, PT uses the emergence criterion to define the onset acquisition of grammatical features, whereas the CEFR seems to use (grammatical) accuracy as a criterion to describe language progression. However, the descriptive scales for grammatical accuracy describe specific grammatical behavior at each level. With regard to the EC, Nicholas & Wigglesworth (2003: 142) maintain that

[t]his line of thinking traces its roots to arguments originally made in work of the ZISA group (Meisel, Clahsen and Pienemann 1981) and continuing in a modified form in the work of Nicholas (1985) and Pienemann (1998). The core argument of this position is that accuracy is an incomplete and, therefore, inadequate measure of learner progress. While accuracy is an important dimension of second language use, it is only one of many, and overuse of it as a construct disguises important ways in which learners make progress in their ability to make use of a new language.

Nicholas & Wigglesworth (2003) show that the overuse of an accuracy criterion would lead to false assumptions about the creativity of developing learner language. I argue that a more complete picture of learner ability can be painted when integrating an emergence criterion into the concept of grammatical competence adopted in the CEFR. The authors of CEFR do not use an acquisition criterion in their scales, because the CEFR is not concerned with measuring something that has been acquired by a language user. Rather, the description of language use is at focus in the CEFR. However, the quantitative scales of the CEFR specify features that distinguish one CEFR level from the other level.⁷⁶ The scale for grammatical accuracy, for example, suggests that the ability to self-correct errors distinguishes level B2 from level B1 (see CoE 2001: 114). There thus seems to be some sort of assumption about development in this CEFR scale, but it is not specified as to how to measure this development (because of the descriptive nature of the CEFR).

One of the most important differences between the CEFR and PT is that the CEFR is a descriptive framework that originated from the appreciation and promotion of a plurilingual Europe. It seeks to describe language competences and to inform language professionals about language education issues. PT aims at describing and explaining one component part of language acquisition, i.e. morphosyntactic development in language production and, more recently, comprehension (see Lenzing 2017) in a universal manner. PT does not use the term competence but development. Development reflects the gradual and accumulative acquisition of processing procedures. Also, the CEFR is not a theory but a reference tool that can be consulted by language practitioners, whereas PT is a psycholinguistic theory of SLA that seeks to explain and predict language L2 development. Moreover, the CEFR deliberately refrains from adopting a theory⁷⁷ of language but summarizes research on general and holistic issues connected to language use. Pienemann, on the other hand, takes a modular approach to explaining grammatical development that can be extended with necessary

⁷⁶ See also the criterial features specified by the CoE (2011).

⁷⁷ This is despite some internal contradictions in the document. For example, the authors of the CEFR clearly advocate Byram's model of Intercultural Communicative Competence.

modules (Pienemann 2005c: 69) (e.g. modules on the source of linguistic knowledge). When discussing the reason as to why a coherent, holistic and exhaustive theory to SLA has not yet been produced, Pienemann (2005: 69) argues that this “[...] is due mainly to the enormous complexity of the task at hand.” He also describes that “[...] in theory construction one should aim for theoretical parsimony” (Pienemann 2005b: 66). This is the reason why PT focuses only on a subtask of SLA for explaining the existence of developmental patterns.

I consider it valuable to find interfaces between the Common Framework and SLA in order to add to the descriptive machinery behind the CEFR so that a more coherent account to grammar in the CEFR might be possible. Even though it not the aim of the CEFR to favor one particular theory, I deem it useful to integrate SLA theory into the CEFR as it will make the document theoretically sounder and subsequently the work of language professionals, who use the CEFR, more informed. The following chapter lays out some chances and challenges for combining PT and the CEFR in more detail. The aim of this chapter is to substantiate the proposed combined scale for *Grammatical Range* in chapter 4.4.4.2.

3.3.2 Competence – the CEFR and PT

It might be useful to briefly revisit what constitutes grammatical competence according to the CEFR in the qualitative dimension at this point, since I claimed that this part of the CEFR is often neglected by language professionals (see chapter 2.1.3). Reconsidering the qualitative part of grammatical competence in the CEFR will help to determine the chances and challenges of combining PT and the CEFR more clearly. This is done in the following chapters.

3.3.2.1 Competence in the CEFR

Figure 24 displays a summary of grammatical competence in the CEFR in context. The summary is by no means exhaustive and solely focused on aspects concerned

with grammatical competence. Therefore, all other competences subsumed under linguistic competences are left out.

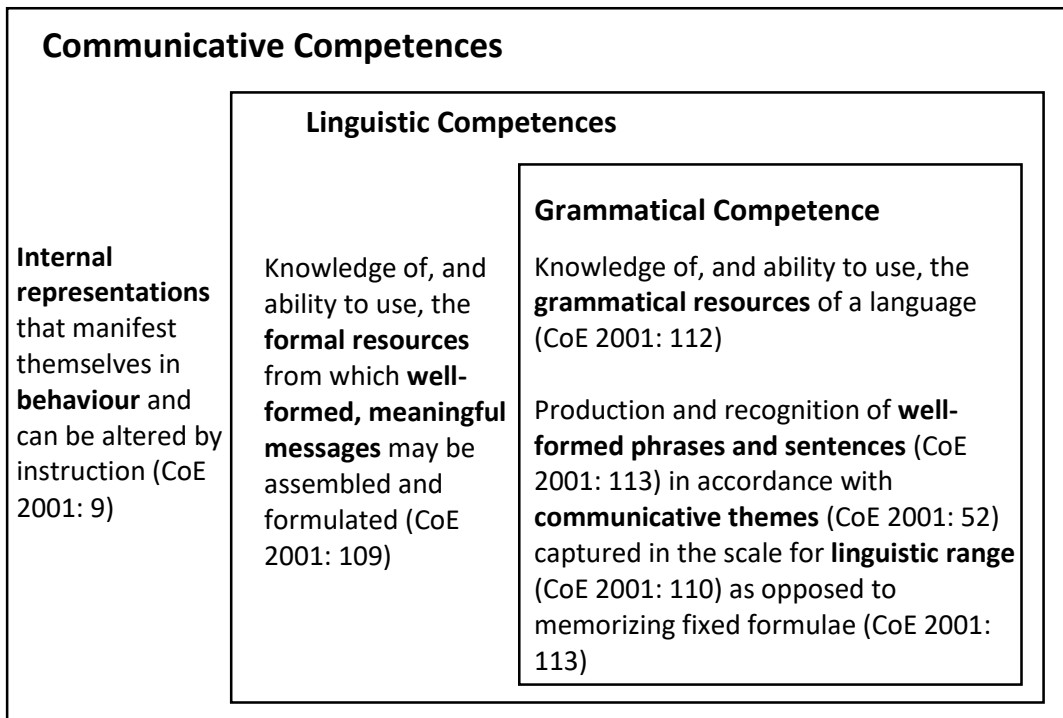


Figure 24: Grammatical competence in the CEFR in Context

Figure 24 shows the CEFR’s definition of communicative competences into which linguistic competences are integrated. One of the six sub-competences of linguistic competences is grammatical competence. Grammatical competence is defined as the “knowledge of, and ability to use, the grammatical resources of a language” (CoE 2001: 112). Grammatical resources are described as sets of principles governing the assembly of elements into meaningfully labeled and bracketed strings (sentences) (CoE 2001: 113). Sentences are assigned a primary role in grammatical competence as they are seen as a means to convey meaning (CoE 2001: 115). Further, grammatical competence encompasses the production and recognition of well-formed phrases and sentences in accordance with principles of linguistic range (see chapter 2.1.5). By combining grammatical competence with linguistic range, the authors of the CEFR want to ensure that the user does not assume that a learner purely memorizes and produces phrases and sentences as fixed formulae (CoE 2001: 113). The reader of the CEFR is directed to the descriptions of communicative themes (CoE 2001: 52) in order to

demonstrate that grammatical competence is situated within the action-oriented approach of the CEFR. However, the relationship between grammatical competence and the action-oriented approach or the communicative themes is not made explicit. It seems as if the authors of the CEFR simply assume that grammatical competence fits into their paradigm. Communicative themes mainly encompass situations in which language users might find themselves while using the target language. The authors of the CEFR also comment on the use of declarative and procedural knowledge in the context of grammatical competence and assume that both are included in their description of competence (CoE 2001: 13). In the elaboration of the term grammatical competence in the CEFR, the phrase “well-formed” can be found quite frequently. One might assume that this is an attempt to describe that the learner’s utterance may be comprehensible so that communication does not break down. However, it rather gives the impression that grammatical accuracy takes a primary role in grammatical competence (considering that the only scale for grammatical competence is the one for grammatical accuracy).

Grimm, Meyer & Volkman (2015: 9) highlight that in the CEFR, competence itself seems to be a fuzzy term that merges “knowledge of the language system and performance as its usage”. The authors of the CEFR seem to be aware that certain factors might interfere with the display of knowledge in performances (CoE 2001: 48). Harsch (2006: 30) explains, in this regard, that the CEFR’s definition of grammatical competence is based on Canale & Swain’s (1980) and Bachman’s (1990) models of language ability. The reader might recall that the CEFR is strongly influenced by the Threshold level (see chapter 2.1.1 for more detail). If Little (2007) is correct in stating that the Threshold Level is related to Hymes’ concept of communicative competence, then it might also be the case that a great deal of the four types of knowledge that Hymes⁷⁸ (1972) proposes was integrated into the CEFR. This seems reasonable, since Canale & Swain (1980) take up on Hymes’ notion of communicative competence. The models stated above will be briefly introduced at this point.

⁷⁸ Hymes’ notion of communicative competence is a reaction to Chomsky’s distinction between competence and performance that originates from his generative view of grammar as linguistic competence.

In Canale & Swain's (1980: 28) view, communicative competence includes a) grammatical competence, b) sociolinguistic competence, and c) strategic competence. The first competence, a) refers to the knowledge of the rules of grammar and b) to the knowledge of the rules of language use. More specifically, Canale (1983: 339) describes "Grammatical competence [as the] ability to use the 'language code' accurately, including correct lexis and spelling, accurate formation of words and sentences, and pronunciation" (additions by KH). In speaking, performance of grammatical competence might be displayed in the following way: "the FL speaker would be able to demonstrate proficiency in applying the grammatical rules that underpin the language, i.e., speak using accurate language, including adequate pronunciation" (summary by East 2016: 26, based on Canale & Swain 1980: 25). The accuracy-focus in their discussion that was probably taken up by the authors of the CEFR is quite evident.

Canale & Swain (1980: 6) argue that both, a) and b), interact, but are different from communicative performance. Communicative performance, in their view, refers to the "[...] realization of these competencies and their interaction in the actual production and comprehension of utterances (under general psychological constraints that are unique to performance)" (Canale & Swain 1980: 6). They further state that communicative competence is observable through communicative performance (Canale & Swain 1980: 29). Canale & Swain (1980: 30) argue that c), strategic competence relates to strategies that are employed when communication breaks down. These strategies might be related to grammatical competence when, e.g., paraphrasing is performed. The third competence, c) might also relate to sociolinguistic competence when role-playing strategies are employed.

Bachman (1990) discusses communicative language ability in the light of language testing. He claims that suitable language tests need to be based on a coherent theory of language ability and discusses that prior approaches to describing language proficiency have failed to produce operationalizable results (Bachmann 1990). Bachmann (1990: 81) states that his approach to communicative language ability is consistent with, as well as an extension of, earlier approaches to competence by, inter alia, Hymes (1972) and Canale and

Swain (1980). Bachman (1990: 84) uses the terms *knowledge* and *competence* synonymously and views competence as closely intertwined with performance. The following Figure represents Bachman's view of communicative language ability.

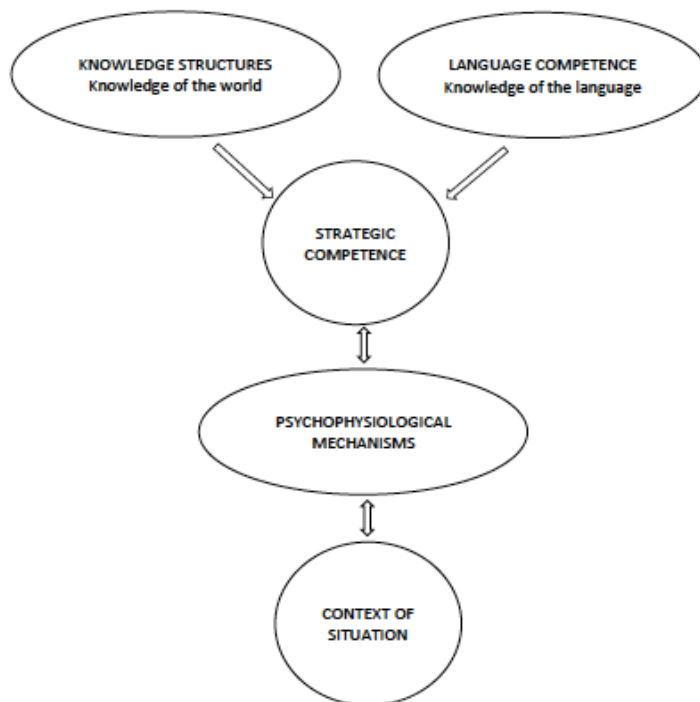


Figure 25: Bachman's View of Communicative Language Ability, taken from Bachman (1990:85)

In my view, the ellipsis that contains language competence and the circle that contains strategic competence are most relevant to this thesis because these are the factors that seem to have influenced the view of grammatical competence in the CEFR, that this thesis is concerned with, most strongly. This is why I will only focus on those two. For a full description of Bachman's view of communicative language ability, consult Bachman (1990: 81f.).

Bachman (1990: 87) breaks down language competence into a) organizational competence, consisting of grammatical as well as textual competence and, b) pragmatic competence that comprises illocutionary competence and sociolinguistic competence. A), Grammatical competence, Bachman (1990: 87) claims, includes all features of grammar that a speaker displays in usage. These are vocabulary, morphology, syntax and phonology. The other component under a) is textual competence. This is displayed by the use of

cohesive devices in a text that shows some rhetorical organization. Competencies under b) include illocutionary competence. Bachman (1990: 90) claims that speakers display this type of competence by the use of speech acts and different language functions, such as manipulative or heuristic functions. Sociolinguistic competence, also subsumed under b) in Bachman's view, encompasses sensitivity to differences in register, or dialect and the ability to interpret cultural references.

In Bachman's view, the component that contains strategic competence is influenced by both knowledge structures and language competence. Bachman (1990: 99f.) employs a broader account to strategic competence than Canale & Swain (1980). Bachman agrees with Canale & Swain in that his view of strategic competence also comprises strategies to employ when communication breaks down and strategies that the speaker uses to enhance the rhetorical effect of his/her utterance. Additionally, Bachman (1990: 100) makes recourse to Faerch & Kasper's (1983) psycholinguistic model of speech production that describes the planning process involved in communication. For this reason, Bachman includes *assessment, planning and execution* into his view of strategic competence. Although the terms that Bachman uses resemble the terms that Levelt (1989) uses in his blueprint for the speaker (see chapter 2.2.1.1), Bachman's ideas of *assessment, planning and execution* differ in that the latter focuses on the dynamicity of communicative acts. Therefore, *assessment* is about identifying the information and the context needed to achieve a communicative goal. The *planning* component in Bachman's view draws on organizational and pragmatic competences to achieve the speaker's communicative goal in relation to the communicative situation that the speaker finds him/herself in (see the relation to communicative themes as taken up by the CEFR). *Execution*, Bachman (1990: 103) explains "[...] draws on the relevant psychophysiological mechanisms to implement the plan in the modality and channel appropriate to the communicative goal and the context".

Bachman's view of communicative ability extends the one by Canale & Swain, especially in the area of strategic competence. Steininger (2014: 47) observes that the authors of the CEFR seem to point to Bachmann's model only

in chapter 9, whereas Canale and Swain's model is mentioned more often. He argues that this fact would contradict the action-oriented approach of the CEFR because Canale & Swain do not view knowledge, and the ability to use this knowledge, as closely intertwined (Canale & Swain 1980: 7). Bachmann (1990: 107f.) however, conceptualizes strategic competence as the link between knowledge and its use.

I do not consider it possible to fully determine as to which model the CEFR seems to follow more. It seems to be the case, though that the CEFR assumes at least some sort of interface between competence and performance as its authors highlight the situations in which the learner is supposed to display their competence because they advise the user of the CEFR to read the scale for grammatical accuracy in connection with the scales for linguistic range as well as communicative themes (see CoE 2001: 113 and CoE 2001: 52). Little (2007: 175) argues that the CEFR follows the definitional approach to describing linguistic ability taken up in the Threshold level⁷⁹ (van Ek 1975) that views linguistic performance to reflect more than purely linguistic knowledge. In assembling the Threshold level, van Ek (1975) relates his description of performance to Hymes' (1972) concept of communicative competence. Hymes opts for an integration of socio-cultural aspects into the competence and performance debate. Thus, Hymes (1972: 277f.) argues that "[...] a normal child acquires knowledge of sentences, not only as grammatical but also as appropriate. He or she acquires competence as to when to speak, when not, and as to what to talk about to whom, when, where, in what manner." Hymes claims that language form is related to linguistic competences, whereas the function of language comprises communicative competences. In his view, communicative competence is the interaction between grammatical, psycholinguistic, sociocultural and probabilistic competences. Hymes (1972: 281, emphasis in original) summarizes this by stating four types of knowledge that might be analyzed:

1. Whether (and to what degree) something is *possible*;
2. Whether (and to what degree) something is *feasible* in virtue of the means of implementation available;

⁷⁹ See chapter 2.1.1 for more details on the T-levels.

3. Whether (and to what degree) something is *appropriate* (adequate, happy, successful) in relation to a context in which it is used and evaluated;
4. Whether (and to what degree) something is in fact done, actually *performed*, and what its doing entails.

In the following section, a brief discussion of the notion of competence through a PT lens will be given. It needs to be borne in mind, however, that PT does not use the term competence but is rather concerned with language development.

3.3.2.2 Competence and PT

The scope of explanation of PT is necessarily narrower than the scope of the CEFR due to its theoretical psycholinguistic nature that explains the developmental path in SLA. PT primarily focuses on linguistic aspects and their representations in the learners' minds. Pienemann (2005b: 69) criticizes the research by White (1991), who equates research on acquisition only with research on linguistic knowledge. He (2005b: 70) argues that PT follows Kaplan & Bresnan's (1982) integrative line of thought in the discussion of which mental capacities underlie linguistic ability. Bresnan & Kaplan (1982) attribute special emphasis to the study of the language processor for exploring interfaces between competence and performance. Bresnan & Kaplan (1982) propose the Competence Hypothesis as a methodological principle in explaining the relationship between linguistic knowledge and its application in performance. Kaplan & Bresnan (1995: 1) explain

[...] that an explanatory model of human language performance will incorporate a theoretically justified representation of the native speaker's linguistic knowledge (a grammar) as a component separate both from the computational mechanisms that operate on it (a processor) and from other nongrammatical processing parameters that might influence the processor's behavior.

These claims are made in the context of developing the grammatical formalism LFG (see chapter 2.2.1.2). Pienemann (2005b: 70) considers LFG a coherent model and asserts that "[...] it provides a basis for relating linguistic knowledge to the processor." Kaplan & Bresnan (1995: 2) explain that a fundamental problem in developing theories of syntax is to explain the mapping between

semantic predicate argument relationships and the surface word/phrase configurations. This fundamental problem sets the framework for their discussion of linguistic knowledge and processing. Bresnan & Kaplan (1982) assume the existence of a direct correspondence between the rules of a grammar and the operations performed by the human language processor. Their idea is that the language processor is not equal to linguistic knowledge but able to operate on it, so that grammatical knowledge is only accessible through the linguistic processor (Pienemann 2005: 70). The language processor is understood as "the computational mechanisms that operate on (but are separate from) the native speaker's linguistic knowledge" (Pienemann 1998: 5). In the context of proposing the MCH (see chapter), Lenzing (2013) makes claims about the initial mental system of language learners. From her discussion, one can infer that linguistic knowledge refers to the mental representations in the learner's mind (see Lenzing 2013: 43ff) and that processing operations are different from linguistic (grammatical) knowledge.

As stated before, PT is not concerned with providing a model of competence, rather its modular approach aims at comprehensively describing and explaining, from a psycholinguistic point of view, the unification of grammatical features at any given point in the language acquisition process. Thus, PT's primary focus is on processing and not on describing linguistic knowledge. PT specifies the acquisition of the respective processing procedures. Thus, one can summarize that PT is concerned with the processing of grammatical knowledge and assumes that the language processor is the link between competence and displaying grammatical knowledge in performance:

Language acquisition studies that focus on linguistic competence therefore ought to place special emphasis on the interface between the processor and grammatical knowledge, since the latter is only accessible through the former, especially where it cannot be taken for granted that individual utterances are representative of the structure of the underlying linguistic system (Pienemann 2005: 70).

In the context of SLA, Pienemann argues that acquisition does not ultimately lead to production; judging by the above quote, however, it has to be viewed as a component part of performance.

The next chapter briefly discusses chances and challenges in combining PT and the CEFR with regard to their views of competence.

3.3.2.3 Competence: Chances and Challenges in Combining PT and the CEFR

The accuracy-based view in the CEFR is probably the one of the most prominent differences between the CEFR and PT. As explained in sections 2.2.4 and 3.3, PT does not make predictions about accuracy levels since it is concerned with explaining developmental patterns in SLA from a processing perspective. PT also does not use the term *competence* (Pienemann 2005b: 62) because PT is primarily concerned with processing. However, Pienemann argues that grammatical knowledge can be pursued within the PT framework because the language processor operates on linguistic knowledge. The processor, however, is not seen as equal to linguistic knowledge (Pienemann 2005c: 70). Therefore, the processor is ascribed a central role in describing grammatical knowledge as it is a prerequisite for putting knowledge to use. This processing-based perspective, i.e. the role of the processor for operating on grammatical knowledge, is quite compatible with the core aspect of the CEFR's view of grammatical competence. If the core aspect of grammatical competence is seen as the "knowledge of, and ability to use, the grammatical resources of a language" (CoE 2001: 112) in the CEFR, then the processor and its ability to work on grammatical knowledge fit into this definition quite well.

I argue that the accuracy-based view of the CEFR is deficient in that language learners necessarily make mistakes during their acquisition process. Even in 1974, Wilkins (1973: 14) described that where there is grammatical inaccuracy, communication can still take place. This statement can be read in relation to the prominent Communicative Language Teaching approach that focuses on meaning before accuracy. Even though this approach is currently dominant in language teaching, the CEFR seems not to integrate these ideas into their concept of grammatical competence, as it seems to be represented entirely through accuracy in its quantitative dimension.

In this context, the question arises as to how learners display linguistic knowledge. Pienemann maintains that the only valid point to retain that something has been acquired, is when it is displayed in production, hence the emergence criterion (Pienemann 1998, see section 2.2.4 for more details). Similarly, it is written in the CEFR that “[p]rogress in language learning is mostly clearly evidenced in the learner’s ability to engage in observable language activities and to operate communication strategies” (CoE 2001: 57). Observable behavior also seems to be focused on in the CEFR.

We can deduce at this point that the approaches taken to explaining linguistic knowledge and the display of that knowledge differ strongly in PT and the CEFR. This difference, in my view, mainly lies in the fact that PT aims at explaining the developmental path present in language learning through a processing perspective, whereas the CEFR superficially describes possible features of competence holistically. The aim of the CEFR is to make language professionals aware of the features and not to explain them. Both the CEFR and PT seem to assume some sort of interface between linguistic knowledge and its display. However, the focus of explaining (PT) and describing (CEFR) this interface differs strongly.

After having outlined the differences in the notion of competence (and to some extent performance) in PT and the CEFR, as well as potentials for combining them, I aim to review some ideas in the CEFR descriptors for grammatical accuracy, progression and processes in more detail and relate them to PT.

3.3.3 The Shape of the Emerging Linguistic System in Learners

Grammatical Range, as proposed in this thesis, might be envisaged in the light of possible options of feature unification at each level of development as explained by PT. In this context, accuracy criteria are less important than currently emphasized in the CEFR scale for Grammatical Accuracy because learner language necessarily contains ungrammatical features. Rather, the universal formal grammatical operations should be at the center of attention. What goes along with this, is that the development of grammatical structures cannot be

conceptualized based on a developmental path alone. Development encompasses two dimensions that Pienemann (1998: 231ff.) captures in his notion of *Hypothesis Space*; development and variation. Grammatical competence, I argue, must therefore be seen in relation to both areas: development and variation. For this reason, progression will be focused on in the next two chapters.

3.3.3.1 Progression in the CEFR

In the CEFR, it is stated that “[p]rogress is not merely a question of moving up a vertical scale” (CoE 2001: 17) and that the levels in the CEFR “[...] only reflect a vertical dimension” (CoE 2001: 17), whereas “[...] learning is a matter of horizontal as well as vertical process [...]” (CoE 2001: 17). It thus seems that both dimensions are covered in the qualitative part of the CEFR but not reflected in the quantitative part comprising the scales. This seems to be another internal contradiction in the CEFR.

Little (2014: 24) comments on progression in the CEFR scales and states that “A1, A2 and part of B1 are mostly concerned with informal communication in a wide range of everyday contexts, and reading and writing play a relatively minor role (BICS – Cummins 1979, 1991)”, whereas in “B1 we encounter more formal uses of language, and as we progress through B2 and C1 to C2 the development of proficiency is increasingly academic and literacy-based (CALP – Cummins 1979, 1991)”. Little’s quote shows that progression in the CEFR does not seem to happen in a linear, even fashion. Westhoff (2007: 676), however, reads the descriptors differently and states that indeed, from the descriptors for grammatical accuracy in the CEFR, a view of a linear progression in language acquisition can be perceived. He claims that the linear progression results from conceptualizing an outdated view of formal instruction in which “discrete grammatical items are presented one after the other” and that it seems that learning would happen in this fashion (Westhoff 2007: 676). North (2014: 101), who is significantly involved in the conceptualization of the descriptors in the CEFR, follows Westhoff’s argument by stating that

[I]n linearity in learner progress would assume that, given a constant investment of time and effort, the learner would advance up the levels at a more or less constant speed, with each level requiring more or less the same number of hours, so that one could predict where they [learners] would be in six months or two years later (additions added by KH).

He claims that this view is naïve in nature, as progress is “partly a question of developing competences in new areas” (North 2014: 101). North’s quotation is not directly linked to the descriptors for grammatical accuracy but language learning in general.

Westhoff (2007: 676) further interprets the CEFR descriptors for grammar to show a concentric development of foreign language competence that shows some progression, initially in lexical repertoire at lower levels and that it is characterized by formal correctness based on a limited lexicon. Later stages, Westhoff (2007: 676) observes, show some level of monitoring language use that is “[...] expected to develop gradually and concurrently with a broad array of grammatical issues.” In proposing a scale for *Grammatical Range*, I assume that a combined approach to PT and the CEFR can specify those grammatical issues for specific languages based on the universal processing procedures.

3.3.3.2 Progression in PT

Pienemann views progress in PT as the emerging processing procedures that allow for more and more complex processing operations. Pienemann would also abandon the idea of a linear progression in language development. He conceptualizes language development to happen in a cumulative and successive manner. As to the age factor, Pienemann (1998: 21) argues that “[t]he architecture of the Grammatical Encoder has to be constructed by child and L2 learners alike. There is no reason to believe that fundamentally different processing procedures have to be developed by the two types of learners.” Pienemann (2007: 13) does not link linear development to learners’ efforts, investment of time or speed, as North does (see section 3.4.2.1), but purely to the “[...] course of development of L2 linguistic forms in language production [...]”. Pienemann (2015: 129) further explains “applying the PT hierarchy to a

specific TL [target language] will not result in all grammatical features of the TL being lined up in a tight sequence like pearls on a string – reflecting a hierarchical step from one feature to the next.” Rather, PT challenges the notion of a single linear sequence in SLA because 1) several linguistic features are placed at the same level of processability and 2) learners develop their own versions of interlanguages containing features that are neither part of the L1 nor of the L2 (but which are constrained by the current level of processability) (Pienemann 2015: 129). This is why Pienemann (2015: 129) argues that “[...] learners progress through universal levels of acquisition (in terms of processing), yet the shape of their interlanguages at any one stage may vary”.

Also, Pienemann argues that language acquisition is not a unidimensional matter. It encompasses the dimensions of development and variation (see chapter 2.2.3 on Hypothesis Space) and can therefore not be captured in a vertical scale for grammatical accuracy.

3.3.3.3 Progression and Processes: Chances and Challenges in Combining PT and the CEFR

Above, I claimed that there seems to be a contradiction between the qualitative part and the quantitative part of the CEFR in terms of conceptualizing linearity in learner progress and that the scales in the quantitative part of the CEFR only reflect a vertical dimension. With regard to the descriptors for grammatical accuracy, Westhoff (2007) states that this vertical dimension in the scales implies the notion of linearity in a learner’s linguistic progression. From a PT viewpoint, this idea would be strongly discarded, as Pienemann argues that language development happens in a cumulative and successive manner and that it does not happen in a unidimensional fashion.

In the context of combining PT and the CEFR, the concept that covers development and variation is one that is only rarely referred to in the CEFR – Selinker’s idea of interlanguage. As discussed in chapters 2.2.3 and 2.2.6, interlanguages mirror the emerging grammatical system of learners. Keßler & Plesser (2011) call interlanguages ‘learner grammars’ to illustrate that learners

develop their own mental systems which necessarily contain non-target-like features. However, the notions of interlanguage and learner language are not considered in the description of grammatical competence in the CEFR, even though the authors of the CEFR mention Selinker's concept when they discuss learner errors and mistakes (see CoE 2001: 155). I argue that grammatical competence and interlanguage should be more closely linked, and that PT might help to do so in laying out a universal developmental path that describes the mental system of language learners.

What can be inferred from the discussion on progression and linearity is that the picture painted in the descriptors for grammatical accuracy seems to leave enough leeway to interpret a linear view of progression (see Westhoff 2007). I argue that this outdated view of linearity in linguistic progression can be overcome by integrating PT's principles of cumulative interlanguage development and by proposing a scale for *Grammatical Range*.

Both the CEFR and PT use the term *processes*. *Processes* as described by the CEFR seem, to some extent, overlap with the notion of processing taken up in PT. In chapter 4 of the CEFR, the authors state that processes are viewed as communicative processes and specify the users' actions involved in those processes. In terms of production, the speaker is required to "*plan and organise a message (cognitive skills); formulate a linguistic utterance (linguistic skills); articulate the utterance (phonetic skills)*" (CoE 2001: 90, italics in original). These cognitive subskills – rather than the overall communicative process - might match, to some extent, the assumptions of formulating a message as proposed by Levelt's Blueprint for the Speaker (1989) that is integrated into PT. At any rate, the terminology used in the CEFR to describe the production process matches the names of the processing components in Levelt (1989), i.e. the Conceptualizer, the Formulator and the Articulator. Here, PT, based on Levelt, might be able to complement and illustrate the production process as outlined in the CEFR (2001: 90) in more detail.

To summarize, the notion of processes, laid out by PT, are to a large extent compatible with those that the CEFR makes about grammatical competence – or rather – the ideas that were adopted by the authors of the CEFR

(for example the different competence models) and discussed by different researchers (see e.g. Westhoff 2007). From this discussion, it becomes evident that PT can be integrated into the CEFR because of PT's modular nature but that the ideas about progression in language acquisition differ.

3.4 Applied Issues in the CEFR and PT – Combined Assessment

In the following section, applied issues on combining PT and the CEFR in terms of assessment will be discussed briefly. I assume that a combined proficiency rating with Rapid Profile, based on PT, is beneficial in assessing a more accurate picture of a learner's grammatical ability. This is a further reason why I regard a scale for *Grammatical Range* as important.

3.4.1 Assessment in the CEFR

The most common way of assessment that places learners at different CEFR levels are psychometric rating procedures that commonly use assessment grids aligned with the CEFR. Raters are asked to match the learner language that they are presented with to the specifications in the assessment grids. Many studies report issues when it comes to human ratings (see e.g. Pienenmann et al. 1988; Brindley 1989, Pollitt & Murray 1993, Chalhoub-Deville 1995; Milanovic et al. 1996; McNamara & Lumley 1997; Brown 1995, 2000; Wisniewsky 2017a). These mainly encompass a) problems with inter-rater reliability (e.g. Michalska 2010) b) that the concepts in the assessment grid are arbitrary or used by the raters in an arbitrary way (e.g. Deygers et al. 2018) and c) that raters do not actually use the assessment grid but are biased by other factors (e.g. Wiesniewski 2017a). This ample criticism demonstrates that it is important to find alternative, more reliable ways to assess the ability of learners, especially when the individual's life choices are dependent on the rater's decision (such as an admission to a semester abroad, etc.). In the study presented later in chapter 4, I address the issue of inter-rater reliability in my data. In PT, the reliability issue is resolved by

using a strong acquisition criterion and the semi-automatic assessment tool, Rapid Profile.

3.4.2 Assessment in PT

PT uses a different approach to assessing language development. It is based on a discriminate distributional analysis and the emergence criterion (see chapter 2.2.4). These approaches seem so inherently different that Pienemann & Keßler (2007) formulated the scope-precision dilemma. The scope-precision dilemma specifies that rating scales aim at assessing a maximum scope of language whereas the SLA measures aim at a maximum precision in their assessment. I agree with Pienemann & Keßler (2007) that there is a discrepancy in assessment approaches, but I would argue that both measures should be combined. With the fully automatic PT-based interlanguage parser APES that Pienemann & Lanze are currently developing in an artificial intelligence environment, a feasible, valid and reliable SLA measure can be employed to back up human ratings (see chapter 2.2.7).

One factor in assessing language is to determine what counts as evidence of competence/performance/ability or development. In PT, this measure of development is clearly marked by the point of emergence (see emergence criterion in section 2.2.4). The use of an emergence criterion exhibits the advantage of having a clear cut-off point for measuring acquisition because it remains stable across different points of elicitation (see chapter 2.2.4 for more details). Pienemann (1998: 146 and 2015: 139) highlights that in determining evidence in SLA, there is a difference between (1) no evidence because there are no linguistic contexts, (2) insufficient evidence because of only a small number of linguistic contexts, (3) evidence for a non-application of a rule although contexts for the respective linguistic structure were present, and (4) evidence for rule application, i.e. sufficient contexts for rule application, are present and the linguistic structure in question is sufficiently applied. To accurately represent the

current state of learners' interlanguages, the assessor needs to bear these possibilities in mind.⁸⁰

3.4.3 Interfaces Regarding Assessment in PT and the CEFR

The CEFR sets out to most holistically describe what language use entails in that they set standards for different levels. A striking point is that North claims that the framework does not focus on documenting SLA (North 2014: 23). Whereas I understand that the CEFR does not take a particular view on language acquisition but tries to inform about different approaches, I am still surprised as to why acquisition research has such a low status in the framework (see also Alderson 2007, Hulstijn 2007), especially when its claim in assessment contexts is to be mainly concerned with validity. Combined assessments of PT and CEFR-based ratings may help to add to the validity issue and assess learner language according to the proposed scale for *Grammatical Range*.

When assessing grammatical competence, based on scales of the CEFR as suggested by the use of grids or rating scales, the choice of assessment tasks strongly determines the possibility for the learner to demonstrate his/her ability. Currently, test centers employ tests that mainly aim at one particular CEFR level. For this, assessment situations are constructed that match the descriptors of the respective CEFR level. In TELC exams, for example, learners who are assessed for Level A1 have to demonstrate that they can introduce themselves. For this, the learner needs to utter sentences like "My name is XXX". As shown above (chapter 2.2.2), from a PT perspective, a sentence like this might very well be classified as a formulaic pattern. To rule formulaic use out, a distributional analysis would need to be administered. The question is whether assessments based on the CEFR is willing to accept that for some of their tests, learners might display learned-by-heart linguistic structures rather than productive language. One could argue that the assessment of productive language use is not at the center

⁸⁰ One has to keep in mind that Pienemann's reasoning is mainly concerned with data elicitation for research purposes and not testing per sé. However, with Rapid Profile and the use of tasks for determining interlanguage development, the same criteria are applied.

of those early CEFR levels, but this would take us back to the discussion of what grammatical competence then actually entails.

In PT, a learner might produce a phrase such as, 'Me no live here', instead of its grammatically correct version "I don't live here" at stage 2 of the Processability hierarchy. According to the emergence criterion (Pienemann 1998), as laid out in chapter 2.2.2 of this thesis, it will not make a difference if the structure is not produced in a grammatically accurate manner. What is important in PT is that the underlying structural operation can be carried out by the learner. In their description of grammatical competence (see CoE 2001: 151), the CoE does not explicitly state accuracy as one of the criteria that language learners need to fulfil. Rather, the focus is put on the organization of "[...] sentences to convey meaning", following their communicative and action-oriented approach. However, in the comments for users of the framework, the CoE (2001: 152) suggests considering "the relative importance attached to range, fluency and accuracy in relation to the grammatical construction of sentences". What is more, the only scale that is given for grammatical competence is the one for "Grammatical Accuracy". Although it is suggested to read this scale in accordance with the scale presented for Linguistic Range, in my view, it sends the wrong message, namely that grammatical competence mainly comprises grammatical accuracy.

The Scope-precision dilemma, as spelled out by Pienemann & Keßler (2007, see chapter 3.5.2), specifies that ratings are employed for large-scale assessments that aim at a maximum scope, whereas Rapid Profile assessments aim at maximum precision. Thus, it seems as if both ways of assessment were not compatible. However, I assume that given the right choice of tasks, scope as well as precision can be accounted for. Thus, a combined assessment of ratings with Rapid Profile (or APES, see chapter 2.2.7) is quite possible. In this way, the specifications for *Grammatical Range* that I propose might be testable through a combined proficiency rating and the use of Rapid Profile (or APES), given that an assessment grid based on the scale for *Grammatical Range* is produced.

3.5 Summary

This chapter outlined that there are a number of conceptual differences in PT and the CEFR. These differences are mainly grounded in the fact that PT is a psycholinguistic theory that explains second language acquisition based on universal processing procedures, whereas the CEFR is a reference tool that may be consulted by language professionals. However, throughout this chapter, it became apparent that PT is able to extend the notion of grammatical competence in the CEFR by proposing to consider varying interlanguage shapes; a developmental and variational component in language progression in the form of a scale for *Grammatical Range*. By integrating the notions of PT, I assume that the CEFR will remain compatible with its action-oriented view towards language use, but also employ a more learner-centered view towards grammatical ability in language users. This chapter laid out the CEFR and PT in some detail. It described prior studies in the field of finding interfaces between SLA and the CEFR and discussed issues and potentials for integrating PT into the concept of grammatical competence in the CEFR in order to produce a scale for *Grammatical Range*. I will now go on to describe the empirical study that aims at finding interfaces between the CEFR and PT.

4. The Study

In the following chapter, I will lay out the details of the present study. The chapter sets out with a rationale that covers to what extent my study adheres to the gap in research that has been shown in current literature. I will then depict the aims of the study, present my research questions and describe by hypotheses. The description of the methodology and data analysis in relation to the twofold aims of the study will take up the majority of this section. I will conclude by describing and discussing the results of my analysis.

4.1 Some Words on the Rationale

According to the authors of the CEFR, “[t]he development of the learner’s *linguistic competences* is a central, indispensable aspect of language learning” (CoE 2001: 149, highlight in original). Grammar, as one aspect in the array of linguistic competences as described in the CEFR, should therefore be regarded as a building block of language and language learning. In my view, PT is able to capture and explain second language grammatical development in a comprehensive, theoretically profound and empirically grounded way. Its modular nature permits PT to be linked to other frameworks, and the CEFR with its open, non-exhaustive and undogmatic view is open enough for such a combination. The link between the SLA developmental schedule as proposed by PT and the CEFR in terms of grammatical accuracy is what the present study aims to explore. What is missing in the current version of the CEFR’s depiction of grammatical accuracy is a specification of which grammatical structures in SLA relate to which CEFR level. Hulstijn et al. (2011: 243) suggest that

[v]ocabulary appears to be the most important linguistic component at the lower levels. But which grammatical and phonotactic elements must a learner minimally control at these levels in the case of languages typologically as divergent as Chinese, Japanese, Finnish, and English? Note that research on these questions is particularly needed in the productive skills (speaking and writing).

My study takes up this need for exploration of the speaking skill as PT is primarily focused on oral language production. My claim is that PT can add to the CEFR in two ways: 1) implementing universal processing procedures into the CEFR scale and thus grounding the CEFR in language-independent SLA research, as well as 2) specifying which language-specific grammatical structures in the SLA process relate to which CEFR level. Hulstijn et al. (2011: 218) further state that “research is needed on how little linguistic competence is minimally required to perform tasks at the lower levels (A1, A2, and B1).” In addition to the research that is needed on the interplay between SLA and the CEFR, Hulstijn et al. (2007:16) also criticize that

[c]hapter 5 [of the CEFR] 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.

These quotations pinpoint the aims of my study. I argue that integrating the universal developmental schedule put forward by PT into the CEFR, adds to a) the empirical basis of the CEFR and b) to grounding the scale for grammatical accuracy in SLA research. Another issue that this thesis addresses is one that was criticized by Hulstijn et al. (2007: 17): “Furthermore, what the CEFR does not indicate is whether learner performance at the six functional levels as defined in Chapter 4 [of the CEFR] actually matches the linguistic characteristics defined in Chapter 5 [of the CEFR], and, more specifically, which linguistic features (for a given target language) are typical of each of the levels” (additions by KH). I intend to a) find interfaces between PT and the CEFR in terms of grammatical accuracy based on the scale for Global Oral Production and b) extend the scale for Grammatical Accuracy to cover *Grammatical Range* in order to focus more on the learner and the acquisition process. With a), specific linguistic structures at each of the CEFR levels can be discriminated (at least for those languages that PT currently covers) and with b), a more learner-centered view on grammar that matches the ideas of the qualitative part of the CEFR may be employed. Thus, spelling out a scale for *Grammatical Range* by implementing PT structures might give rise to a more discriminate view of grammatical ability informed by a universal, processing-centered view of SLA. Westhoff (2007: 676) argues that “[...] although the CEFR descriptors tell us a lot about what learners at a certain level can do, very little is said about what they should know in order to carry out these language tasks. In particular, the question of whether a certain level requires mastery of specific grammar items is left open.” The present study might add to the discussion of this shortcoming in that PT proposes an implicational hierarchy of processing procedures. I thus assume that the implicational nature of PT, when combined with the levels of the CEFR, can specify which CEFR level requires the acquisition of which (language-specific) grammatical items, and which universal processing procedure.

The overall aim of the study is to address the empirical basis of the descriptive machinery in the CEFR and to come up with a more learner-centered

view of grammatical competence in the CEFR. This learner-centered view cannot result in a scale for grammatical accuracy since accuracy is not a valid measure of grammatical ability (see chapter 2.3.2). Therefore, a scale that combines principles of PT and the CEFR for *Grammatical Range* is proposed. Several authors (e.g. Pienemann, Johnston & Brindley 1988, Harley et. al. 1990, Hulstijn 2007) have argued that for a description of communicative proficiency levels to be valid, it needs to be operationalized and grounded in, amongst others, empirical SLA research. Until the present date, the connection between SLA research and the CEFR has been elusive (see e.g. Hulstijn 2007). Wisniewski (2017a: 6) lays out three minimal prerequisites for empirical scale validity: 1) scales should be linked to models of communicative language ability and ideally mirror research findings from SLA, 2) scales should be relatable to empirical learner language, and 3) evidence as to the ability of human raters to apply the scales should be delivered. I assume that with the help of my study, it is possible to approximate the ideals put forward by Wisniewski on all three levels. Whereas issues 1) and 2) are directly covered in this thesis, issue 3) is addressed indirectly and should be investigated in more detail in future research.

After having described the rationale of the study and touched upon some of its aims, the research questions and hypotheses will be made explicit in the following section.

4.2 Research Questions and Hypotheses

In this study, I will explore the question “Is there a relationship between the six stages of language development as predicted by PT, and the six levels of communicative proficiency as described in the CEFR?” In order to shed light on the interfaces, 14 learners of English are assessed by means of the CEFR assessment grid for Overall Oral Production and Linguistic Profiles based on PT. In my view, interfaces between PT and the CEFR can only be explored via language assessment based on both frameworks. I argue that in order to find empirically-based interfaces, I first need to examine the role of grammar in

Overall Oral Production assessment based on the CEFR because a) PT is mainly concerned with grammatical features, b) PT mainly focuses on the production of learners, c) grammar in the CEFR is but one component part of oral production and d) an empirical account to finding interfaces can only be possible when using rating procedures based on the CEFR scales.

My research questions (RQs) are the following:

(RQ1) Are there correlations between PT and the CEFR?

This question aims to explore as to whether morpho-syntactic development, as explained by PT, is reflected in the CEFR. Since PT is a psycholinguistic theory that predicts morpho-syntactic development in SLA, another factor has to be determined in order to address the question about interfaces between PT and the CEFR: which role does *grammar* play in the CEFR descriptors? This question was focused on in the theoretical part of this thesis (see chapter 3). In order to empirically investigate this question, it is necessary to ask (1a) Which influence does grammar have on proficiency ratings with CEFR rating grids? In order to address the rater focus, a new direct, methodology is used. Raters are asked to rate two audio-files of authentic learner language with the grid for overall oral production. In one of the files, grammatical features were manipulated so that one sample is grammatically more accurate than the other sample (see chapter 4.3 for more details about this methodology). All other features in the sample are left untouched so that grammar is the only variable that was manipulated. The results of the ratings for both samples are compared. The research question that is connected to this is: Do raters rate the same audio sample on a lower CEFR level for overall oral production when the grammatical variable is manipulated in the sample? The rating results for the original samples without manipulations can then be used to address the superordinate research question (RQ1): How do CEFR rating results and profile results based on PT correlate when the rating and the profile analysis are carried out on the basis of the same samples? In other words, what kinds of connections can be found between the CEFR and standard developmental schedules?

Due to the study design that is explained in chapter 4.3, I am also able to shed light on assessment issues. In particular, the assessment issues are related to the influence of rater experience and assessment grid use on reliabilities of assessment procedures and their results. The second major research question (RQ2) thus is:

(RQ2) Do rater experience and assessment grid use influence rating results?

That is, do experienced raters behave differently from less experienced raters in terms of assessing learner language? Questions connected to this are: Are ratings more reliable when the CEFR assessment grid is used or do raters, who rely on pure intuition, perform equally well in terms of reliability of rating results? Do experienced raters produce more reliable results than less experienced raters? As outlined in chapter 3.4.1, studies have found issues in psychometric rating procedures that are due to the behavior of human raters. These issues encompass a) problems with inter-rater reliability (e.g. Michalska 2010), b) that the concepts in the assessment grid are arbitrary or used by the raters in an arbitrary way (e.g. Deygers et al. 2018), and c) that raters do not actually use the assessment grid but are biased by other factors (e.g. Wiesniewski 2014; 2017a). I want to explore these issues in my data because I argue in chapter 3.4.3 that combined assessment based on the CEFR and PT can lead to more reliable results that might be contextualized in SLA theory.

I put forward the following hypotheses:

(H1) There are correspondences between PT and the CEFR.

I assume that the correspondences are stronger at the lower CEFR levels at which language production (i.e. lexicon and grammar) is more restricted and less

elaborate⁸¹. This Hypothesis will be the basis for putting forward a combined scale for *Grammatical Range* based on PT and the CEFR.

I also put forward the following hypothesis H2:

(H2) Grammar plays a crucial factor in determining the CEFR proficiency level of a language learner, especially with less experienced raters.

I hypothesize that proficiency raters will rate samples with grammatical inaccuracies at a lower level, as compared to their more accurate correspondents, even if all other aspects of the learner language are the identical. Because of my study design, I am also able to discuss issues relating to the use of intuition in language assessment in contrast to the use of an assessment grid. I therefore hypothesize that the use of an assessment grid by proficiency raters produces more reliable results than ratings based on pure intuition. I do however assume that a higher level of experience does not add to a higher inter-rater reliability. To summarize, the present study encompasses two foci that need to be covered in order to shed light on interfaces between the CEFR and PT:

- a) which relations between PT and the description of language proficiency, as conceptualized by the CEFR, can be found, and
- b) how reliable are the rating results when distinguished between use of an assessment grid and experience level?

4.3 Methodology

Since this study encompasses two foci and many different steps that need to be accomplished in order to be able to address the global research questions, I present an overview of the overall procedure at this point. The overview contains references to the chapters that lay out the details of each step more closely.

⁸¹ Pienemann (1998: 232) explains this phenomenon with the concept of hypothesis space for development and variation in which he argues that the leeway of variational options that might be produced by language learners broadens when progressing in the developmental hierarchy.

4.3.1 Overview of the Procedures and Analyses

It is the aim of the study to explore interfaces between PT and the CEFR in terms of grammar in order to propose a combined scale for *Grammatical Range*. To address research focus a), I argue that in order for proposing a scale for *Grammatical Range* that combines PT and the CEFR, I need to determine the role of grammatical accuracy in proficiency ratings. This is to make sure that a combination of PT and the CEFR is (empirically) meaningful since PT is focused on the acquisition of morphosyntactic features. In order to shed light on the role of grammatical accuracy in proficiency ratings, I employ a direct approach to assessing which performance features raters attend to in oral assessments. For this, I edit sound files in a way that morphosyntactic features are deleted (see chapter 4.3.5 for information on the editing procedure) in the sound file so that an original (a grammatically more accurate file), as well as an edited, (a grammatically inaccurate file), is produced. The two files differ only in terms of morphosyntactic features. These features are determined in a prior study on the perception of grammatical inaccuracy (see chapter 4.3.2). Raters receive access to both the accurate and the inaccurate files (see chapter 4.3.7 on distribution of the sound files) and rate them with the same assessment grid (see chapter 4.3.9 on the assessment grid). The results for the edited and original file are compared and the effect of the grammatical variable is computed with the help of the Wilcoxon Signed Rank Test. If the grammatically inaccurate file is rated on a lower level than the original file, then I assume it must have been the grammatical variable which caused the rating results to be lower for the edited file than for the original file. If there is a difference between the two files, I can proceed to determine interfaces between PT stages and CEFR levels for grammatical accuracy in order to propose a combined scale for *Grammatical Range*. The methodology for exploring interfaces between PT and the CEFR, in terms of grammar, is outlined in the following:

I collected a body of 14 oral language samples of learners of English as a second language. These 14 language samples are recorded, transcribed and analyzed for the PT stage with the help of the computer program Rapid Profile

(see chapter 2.2.7 on Rapid Profile). The same audio samples are distributed to 53 proficiency raters (see chapter 4.3.7 for details on the distribution) who use the Global Oral Assessment Grid produced by the CoE (2009) to assess the CEFR level for the learners (see chapter 4.3.9 for an introduction of the assessment scales). The Rapid Profile results and the rating results are correlated with the help of the Spearman's Rank Order Correlation Test (see chapter 4.4.4). On the basis of the correlations, the descriptors of the Scale for Grammatical Accuracy and the processing procedures for each of the PT stages are combined, so that the scale for *Grammatical Range*, on the basis of both frameworks, can be produced. All descriptors in the scale for *Grammatical Range* are checked and those descriptors that refer to grammatical accuracy are deleted, so as to produce a learner-centered scale (see chapter 4.4.2).

The second focus of this study is concerned with the influence of rater experience and the use of an assessment grid on rating results. A future direction might be to ultimately develop a combined CEFR-based and PT-based assessment procedure. To address the influence of the rater, three groups of raters at different levels of experience are investigated: a) amateur raters without any experience with rating procedures who use intuition for their ratings, b) novice raters who had received a short training on CEFR-based ratings aligned to the suggestions by CEFR prior to the data collection, and c) expert raters who are currently affiliated to assessment centers (see chapter 4.3.8 for more information on the different groups). The three groups rate the same audio-samples and their ratings are compared in terms of variability of their results as well as within-group agreement by means of the Kruskal-Wallis H Test, the Mann-Whitney-U Test and the Kendall-W Test.

I will now summarize and explain the statistical measures that I use in this study.

Results on	Statistical Test
The effect of grammatical accuracy on ratings with the Overall Oral Production grid based on the edited and original samples	Wilcoxon Signed Rank Test
The relationship between PT stages and CEFR levels based on the original samples	Spearman's Rank Order Correlation Test
Rater Experience and Variability of Rating Results across rater groups based on the edited and original samples	Kruskal-Wallis H Test and Mann-Whitney-U Test
Agreement of raters within the different rater groups	Kendall's-W Test

Table 5: Overview of Statistical Test in Relation to Results

Table 5 shows that in order to determine as to whether grammatical accuracy plays a role in proficiency ratings of Overall Oral Production, the effect of the manipulations of grammatical accuracy in the edited samples is calculated with a Wilcoxon Signed Rank Test (see chapter 4.4.4.1 for more details). The Wilcoxon Signed Rank Test determines whether there are significant differences between the original and the edited samples in terms of the CEFR levels that were assigned by the raters. The Wilcoxon Signed Rank Test is the non-parametric equivalent to a dependent sample paired t-test (Dalgaard 2008: 99). It does not need a normal distribution in the data, because the data used for this test constitute ordinal scales. It tests the following null hypothesis: the average signed rank of two dependent samples is zero and thus indicates whether the samples are from the same population or not (see e.g. Dalgaard 2008: 99f).

In statistical measures, the type of scale used for calculations is one of the most important aspects since it reflects the nature of the data and strongly determines the choice of test that is applicable (see McCrum-Gardner 2008: 38). PT stages and CEFR levels can be plotted onto an ordinal scale. Ordinal scales measure non-numeric data in which the order of the features is important, but the difference between the points to be measured is not equal (as would be assumed for numerical data). For example, the difference between “agree” and “fully agree” cannot be determined numerically. In the same way, the difference between the linguistic features located at PT stage 1 and those located at PT stage 2 cannot be quantified. The PT scale thus constitutes an ordinal scale, so

non-parametric methods of measurement need to be employed. Non-parametric methods are generally argued to be “less powerful and less flexible than their parametric counterparts” (McCrum-Gardner 2008: 39), but are able to work with small data sets.

The results pertaining to the role of grammar in the proficiency ratings give the incentive to further investigate the relationship between PT and CEFR in terms of grammar. This relationship is investigated on the basis of the original samples and calculated using the Spearman’s Rank Order Correlation. The Spearman’s Rank Order Correlation Test determines correlations between the original samples at each PT stage (as analyzed prior to the ratings) and the ratings provided by both novice and expert raters (amateur raters are not included in this calculation because they did not use an assessment grid and were not trained in proficiency rating based on the CEFR). The Spearman’s Rank Order Test is the non-parametric equivalent to the Pearson Product-Moment Correlation. It tests the association between two ranked variables in terms of its strength and direction (Fieller et al. 1957: 470). The results of these correlations feed into the overall research question RQ1 and are the basis for suggesting a combined scale for *Grammatical Range* based on PT and the CEFR.

The Kruskal-Wallis H and Mann Whitney-U Tests investigate the effects of rater experience and the use of an assessment grid on the variability of rating results as well as rater agreement. These issues relate to RQ2. The Kruskal-Wallis Test is a one-way analysis-of-variance-by-ranks test (or H test). It is used to determine whether the three independent rater groups are the same or different on the variable of the rating results (Chan & Walmsley 1997: 1775). It thus tests differences across the three groups. The H-test determines whether there are differences across the three rater groups but does not specify between which groups exactly the differences are. Therefore, the Mann Whitney-U Test is used to determine if two of the rater groups come from the same population, i.e. it tests if the two independent groups are homogeneous (Nachar 2008: 14).

The Kendall-W-test is used to calculate the coefficient of concordance within the different rater groups (Legendre 2005). With this test, statements about the agreement of the raters within the three rater groups can be made.

This test helps to determine the variance within the amateur rater group that did not use an assessment grid for their rating in comparison to the agreement within the other two groups who used an assessment grid. The results of the last three statistical measures are supposed to determine whether there is a need to combine rating procedures with PT-based assessment.

In this context, I want to briefly comment on my approach of treating outliers in the data. I decided not to eliminate outliers in my data. In their discussion on Cronbach's Coefficient Alpha, Liu et. al (2010: 5) demonstrate "that coefficient alpha estimates were severely inflated with the presence of outliers, and like the earlier findings, the effects of outliers were reduced with increasing theoretical reliability." As described in the results section, my data generally show quite strong, significant results without having removed outliers. For my research questions, I consider the natural data sufficient and I assume that with the outliers present, the data reflect reality more strongly.

After having presented an outline of the overall procedure as well as a description of my analysis, I proceed to describe the details of the different methodological steps.

4.3.2 Pilot Phase for the Perception of Grammatical Inaccuracy

I will now proceed to give an overview of the pilot phase that relates to research focus a) and is supposed to investigate the perception of grammatical inaccuracy in oral learner data. This phase aims at determining the morphosyntactic structures that are to be deleted in the audio-files for the main data collection phase. As described in chapter 4.3.1, I argue that focus a) which relates so the question of relationships between PT and the CEFR in terms of grammar encompasses to determine the role of grammatical accuracy in global oral production ratings. To investigate the role of grammar in ratings, I employ a new direct methodology to explore the features that raters attend to in oral proficiency ratings. Brown et al. (2005: 6) discuss that "research into the cognitive processes employed in the rating of oral proficiency is extremely limited". Whereas a number of studies on rater cognition in assessing writing

have been published (see e.g. Cumming et al. 2001, 2002; Milanovic et al. 1996), little research is available on cognition processes in oral assessments. Most of the studies that are available on rater cognition in oral assessment use a verbal report methodology (see Pollitt & Murray 1993; Brown 2000). Brown, Iwashita & McNamara (2005: 7) state a strong limitation of verbal report strategies for the assessment of oral rating procedures as follows: “[...] the real-time nature of the assessment precludes the elicitation of concurrent reports, and limits, therefore, what can be inferred about the process of rating, as opposed to the performance features to which raters attend.” Therefore, my methodology comprises a more direct approach to assessing which performance features raters attend to in oral assessments. For this, I edit sound files in a way that morphosyntactic features are deleted in the sound file so that an original (grammatically more accurate) file as well as an edited (grammatically inaccurate) file can be produced. In order to determine which grammatical features should be deleted in the edited sound file, I employ a pilot phase that aims to determine which morphosyntactic features are perceived as highly non-target-like in oral learner language.

Two researchers participated in this pilot phase. In order to determine grammatical features that are perceived as non-target-like by people familiar with the English language, the two researchers recruited 10 advanced teacher trainee and English linguistics students. The 10 students were presented with five audio samples of authentic learner language that I provided. I conducted the five samples in prior studies based on the principles for diagnostic profiles based within the PT framework (see chapter 2.2.7 for more details). The 10 students listened to the different samples of learner language that each contains a variety of non-target-like structures. Additionally, each student received the global scale of the CEFR for an assessment of the five audio-samples. Each of the samples has a mean length of approximately 10 minutes. The researchers were asked to instruct the students in the following way:

“Please use this global scale to place the performance of the language learners on the CEFR levels. Take some time to familiarize with the descriptors and then listen to the recordings. Afterwards decide on a CEFR level. Once you have decided on a CEFR level, feel free to comment on anything that you would consider peculiar in this sample.”

In this phase of the study, it was not intended to yield any valid rating results by using the CEFR scale. The instruction was formulated in this manner, purely to have the students focus on a random task. The actual purpose was different: The two researchers should observe the students' reaction (smiles, laughs, smirks, rising eye-brows, etc.) to grammatically inaccurate structures. The researchers were asked to closely monitor the students while they performed their random rating task; focusing especially on their facial expressions or verbal comments when grammatical errors occurred in the speech sample. Each of those reactions were noted down next to the exact time that they relate to in the audio sample (see report in appendix 7.2). After the ratings were completed, the researchers interviewed the students as to what they thought were the most striking errors in each sample. Their answers were noted down as well.

Verbal comments as well as facial expressions are summarized in Table 6 below together, as it does not make a difference for this study whether they commented on an error or whether they reacted to one. The observations of the researchers can be summarized as follows. The column *No.* displays the number of students who reacted to an error type and the column *Error Type* displays a summary of the errors that the students reacted to. If less than five of the 10 students reacted to an error, I summarized the error types in one column, as I only intend to include those errors in the editing process of audio-files for the main data collection phase that were reacted to most often.

No	Error Type
10/10	a strong reaction to errors in S-V-agreement marking
08/10	a strong reaction to a lack of vocabulary
07/10	a strong reaction to incorrect use of various tenses (e.g. progressive form was missing)
07/10	a strong reaction to untarget-like pronunciation
<05/10	a strong reaction to transfer from German, incorrect use of prepositions, problems in question formation

Table 6: Results Perception of Grammatical Inaccuracy in Pilot Phase

All of the students showed a strong reaction to errors in S-V-agreement marking

in that they commented/reacted when the third-person-s was missing. eight of the students mentioned a lack of vocabulary. This means that the learner was pausing to search for words or using words other than the target language English. Seven reacted to incorrect use of various tenses. Furthermore, seven reacted strongly to untarget-like pronunciation (the incorrect pronunciation of the voiced dental fricative /ð/, as in **th**is, and the voiceless dental fricative /θ/, as in **th**ing, was mentioned most often). Less than five commented on transfer from German, incorrect use of prepositions and problems in question formation.

Based on these findings, I concluded that for my main data collection phase, I need to first and foremost delete the third-person-singular-s and the ing-form in the transcriptions. The transcriptions are used as a basis for producing the manipulated versions of the audio-files. In chapter 4.3.5, I will describe how the transcriptions were edited in more detail.

4.3.3 The Data

In the following, I will describe the data used in the study in more detail. This section focuses on the language learners of whom audio-recordings were compiled.

Twenty-two audio files, which feature 20 learners of English, were rated by 53 raters. The audio files can be divided into 14 original files (featuring 18 learners) and eight edited files. In order to produce the edited samples in which grammatical accuracy was manipulated, I compiled a corpus of 14 different original samples. At least two samples refer to each of the six PT stages.

22 files in total		
14 Original Files		8 Edited Files
PT stages	Number of files	Number of files
1	4	/
2	2 with 4 learners	/
3	2	2
4	2	2
5	2	2
5+	2	2

Table 7: Audio files divided into original and edited files

The corpus consists of two samples from the LARC data base (Ko02 & Ko03) with English language learners from an Indonesian background, four samples (Ko11, Ko12, Ko13, Ko14) of English learners with Finnish as L1 that were taken from a project conducted in Finland by Pienemann et. al. in 2008, four samples (Ko01, Ko07, Ko08 and Ko09) with German as L1 that were conducted by researchers affiliated to Paderborn University at that time, including myself. The participants in the samples Ko01, Ko07, Ko08 and Ko09 attend a lower middle school in Brandenburg and the data were elicited in 2014. I elicited another four samples at the language learning center (Zentrum für Sprachlehre) at Paderborn University in 2013 (Ko04, Ko05, Ko06 and Ko10). These latter four learners had already been placed at CEFR levels by the language learning center prior to this study. The L1s of the learners vary between German, Indonesian and Finnish.

An overview of the edited sound files is given in the last column. The procedure that is used for editing the audio-files is presented in chapter 4.3.5. Only eight edited sound files were produced. The reason for this is that learner language at PT stages one and two does not display enough variability in morphology to be edited. Thus, only samples for PT from stage three onwards are selected for the editing procedure. Two files per PT stages three to five plus are produced. This also explains why there are more original files than edited files. Another reason for the difference in the number of original and edited files

is because the manner of distribution of the audio-files to the raters had to be considered carefully (see chapter 4.3.7 for more detail).

The next chapter outlines the tasks that are used to elicit the PT stages of the learners.

4.3.4 The Tasks

The learners in the files carried out different semi-communicative tasks (see e.g. Ellis 2003), the learner language was transcribed, and the transcriptions were profiled by means of the semi-automatic profiling software Rapid Profile (see chapter 2.2.7 for more detail). While the tasks used for eliciting the PT structures differ in content in the files, the manner of elicitation is basically the same and follows the descriptions by Pienemann & Mackey (1993). Table 8 provides examples of the task types that are commonly used for a Rapid Profile analysis. The task design was based on a number of tasks that had been proven to be effective for eliciting a learner's PT stage (see e.g. Lenzing 2013; Pienemann 1998; Pienemann & Mackey 1993; Roos 2007).

Rapid Profile			
Task Name	Habitual Action	Spot the difference	Interview
Instruction	Describe the daily routine of Mr. and Mrs. Lee.	These are two pictures, they look similar, but they are not. Ask questions to find out about the differences.	I am a Martian and you are an earthling. You can ask me whatever you want to know about me.
Structures	SVO, adverbials, 3 rd -ps-sg-s	Do/Aux-fronting, WH-cop-?, Wh-Aux-2 nd -?	Do/Aux-fronting, WH-cop-?, Wh-Aux-2 nd -?

Table 8: Overview of the Task Set

In the Habitual Action Task, for example, learners are presented with a sequence of pictures of a person who performs daily chores. The learners are asked to describe the daily routine. This way, the production of declarative sentences with third person S-V agreement is supposed to be triggered. This task thus aims at

eliciting SVO structures, adverbials, as well as 3rd-person-singular S-V agreement. Spot-the-Difference tasks aim at question formation. In this task type, the participants are asked to find out the differences between two pictures by asking questions. One participant cannot see the picture that the other participant sees. Due to the differences in the pictures, the participants are given the opportunity to produce a variety of different interrogative structures at different PT stages. The third task that is used to elicit the developmental stage of a learner is an interview. In the interview, the participants are given the opportunity to interview a Martian. It is explained that one participant is a Martian who has travelled to the earth and the other participant is an earthling. Both participants do not know much about the other participant's lives so they are supposed to find out as much information as they can about each other. The participants are thus free to ask any question that comes to mind. This task is the most open task as the participants are not presented with any visual stimuli to trigger questions. The data elicitation for Rapid Profile usually starts with an Habitual Action Task that provides many visual incentives for speaking in order to let the participant ease into the situation. When the participant feels comfortable with the situation, more open-task formats are used.

The next chapter describes the audio-files as well as the editing procedure that is used to produce the edited versions of the original files in more detail.

4.3.5 The Audio-Files and the Editing Procedure

The learners were audio-taped completing either the semi-communicative tasks in pairs or together with a researcher. The recordings were transcribed and analyzed according to their PT level, with the help of the Rapid Profile Software by myself. I used the transcription for the RP analysis and not the audio-file itself in order not to miss any important linguistic structures. The RP analyses are later used for the correlations between CEFR levels and PT stages.

After the RP analysis, the morphological markers, determined in chapter 4.3.2, were deleted in the transcriptions. It is to be noted here that learner language which classifies as PT stages 1 and 2 does not display enough variability

in morphology to be edited. Pienemann (2005: 24) depicts how learners at stage 1 produce invariant forms in terms of morphology because a mere retrieval of lexical chunks from the lexicon, as envisaged by Levelt (1989), is assumed. This is why learners at PT stage 1 are excluded from the editing procedure. As a matter of precaution, I also excluded the learners at PT stage 2, who, in principle, should have been able to produce lexical morphemes (attaching the plural-s, for example). Since I did not perform a distinct distributional analysis in order to rule out the production of chunks, I decided to exclude these samples as well. I consider the morphology that learners can produce at PT stage 3 broad enough so that the audio-files at this PT stage could be included in the editing procedure.

I used the transcriptions of the samples from PT-stage 3 onwards to erase the morphological markers that were previously perceived as very inaccurate in the pilot phase (see section 4.3.2). The markers encompass the third person-singular-s, the past-ed and the ing- form. The transcription of the original audio file and the edited word document were then used to edit the audio file. An example of a transcription excerpt of an original file, with its correspondent edited file, is present in Table 9 below:

Original Transcription Ko06	a)	(um) (er) how many animals are on your picture↑
	b)	Okay (um) Mrs. Lee starts with (#) standing in his/(er) in her bedroom
Edited Transcription Ke06	a1)	(um) (er) how many animal ∅ are on your picture↑
	b1)	Okay (um) Mrs. Lee start ∅ with (#) stand ∅ in his/(er) in her bedroom

Table 9: Example of Editing Procedure in the Transcription

File Ko06 is an example of a learner at PT stage 5. The first row in the Table shows example a), the target-like plural marking for the noun animal. In a1) in line three under the edited transcription Ke06, the noun appears without the plural-s. Example b) shows a target-like example for third-person singular subject-verb agreement in “starts” as well as the ing-form “standing”. In b1), both the third-person-s as well as the -ing morpheme are deleted.

The editing procedure was performed in two different ways: i) an audio engineer used a computer software to cut off the morphological markers in the audio sample. The audio engineer, who is a native speaker of English, judged at the same time if the learner language still sounded natural after the morphemes were deleted in the recording. He thus made sure that no unnatural gaps or back-channeling was present in the recording. Three samples were edited in this way. For the other way of editing the audio-files ii), both the original and the edited transcription were given to different non-native speakers of English and I re-recorded both versions together with these non-native speakers in one session. In this scenario, I acted as the interviewer whereas the volunteers acted as the language learners. We focused on recording the original version first in order for the actors to get a feeling for the sample and then went on to record the edited version of the transcription. Five samples were edited in this way. In chapter 4.5.3.4, I investigate whether the different editing procedures have repercussions on the results. I claim that the two different procedures of editing both qualify as suitable for the study. The following Table presents an overview of the learner data.

File Original	PT stage	Source/Research Groups	According edited file	Edit	Gender
Ko01	4	Hagenfeld/ Göhrmann/Kröger	Ke01	Audio engineer	female
Ko02	4	LARC	Ke02	Audio engineer	female
Ko03	3	LARC	Ke03	Audio engineer	male
Ko04	5	Hagenfeld/ Göhrmann/Kröger	Ke04	Re-recording	female
Ko05	5+	Hagenfeld/ ZfS	Ke05	Re-recording	female
Ko06	5	Hagenfeld/ ZfS	Ke06	Re-recording	female
Ko07 (2 learners)	2	Hagenfeld/ Göhrmann/Kröger	none	/	Both female
Ko08 (2 learners)	2	Hagenfeld/ Göhrmann	none	/	Both male
Ko09	3	Hagenfeld/ Göhrmann/Kröger	Ke09	Re-recording	male
Ko10	5+	Hagenfeld/ ZfS	Ke10	Re-recording	male
Ko11	1	Finland project	none	/	male
Ko12	1	Finland project	none	/	male
Ko13	1	Finland project	none	/	female
Ko14	1	Finland project	none	/	female

Table 10: Sources of Learner Data

The original samples Ko04, Ko05, Ko06, Ko09 and Ko10 were re-recorded with the help of actors. Edited counterparts to the original samples are Ke04, Ke05, Ke06, Ke09 and Ke10 as can be seen from Table 10. The original versions edited by the audio engineer are samples Ko01, Ko02, Ko03 (their corresponding edited versions are Ke01, Ke02 and Ke03). The samples for PT stages one and two, that were excluded from the editing procedures, are termed Ko07, Ko08 and Ko11 to Ko14.

4.3.6 Piloting the Edited Samples

This pilot phase serves two purposes: a) to elicit whether the edited data sounds unnatural or staged, and b) as a trial for the order and manner of distributing the sound files to the raters in the actual data collection phase, described in chapter 4.3.7.

To ensure that the edited data do not sound unnatural, the files were trialed with two teacher trainees who major in English. One of the teacher trainees is a native speaker of English. The students had access to three samples at a time via Dropbox. They were asked to listen to the audio-files fully and then rate them with the help of the Overall Oral Production Grid based on the CEFR. Just as with the pilot phase outlined in chapter 4.3.2, the participants were given the following instruction:

“Please use this global scale to place the performance of the language learners on the CEFR levels. Take some time to familiarize yourself with the descriptors and then listen to the recordings. Afterwards, decide on a CEFR level. Once you have decided on a CEFR level, feel free to comment on anything that you would consider peculiar in this sample.”

I told the participants that this was a pilot phase for my study and that they were free to comment on anything that they thought could be important for the actual data collection phase. In this way, I hoped that they would be more likely to report if they thought something was peculiar about the language samples. After the rating, the participants each sent an email to me with their results and a comment. I then proceeded to delete the samples from the online cloud

“dropbox” and to make a new batch of samples available to them. This procedure was repeated four times until all samples were rated. Again, these ratings were not intended to yield any valid CEFR levels for the samples. Rather, the aim was to find out whether the participants thought that the language samples sounded somehow unnatural, and also to see if they recognized the similarity in the samples. This phase thus acted as a trial for the order and manner of distributing the sound files to the raters in the actual data collection phase described in chapter 4.3.7. Neither of the two trainees gave any comments on unnaturalness or similarity of the data during this pilot phase.

The manner of distributing the files is presented in the following chapter.

4.3.7 Distributing the Files to the Raters

The manner and rotation in which raters were presented with the data fulfilled two functions, namely a) to keep the workload on the raters manageable and, at the same time, to present the raters with at least two samples at each PT stage, and b) to reduce the likelihood that raters recognized the samples that were edited and match them to the original samples. If raters recognized the audio-sample, one could argue that this might bias their rating. To ensure that the raters are presented with a sufficient, as well as even, number of audio-samples for each of the ratings weeks, I chose four samples on stage 1 and stage 2, instead of two audio-samples (two originals and two edited ones) for the other stages of the PT hierarchy for distribution to the raters. The reason for this is that no edited versions for samples at PT stage 1 and 2 were produced. The rater cohort in this study is assigned to two different groups receiving two different data sets. This means that amateur, novice and expert raters (see chapter 4.3.8) are again subdivided into two subgroups. The reason for this is to keep the workload for each rater to a minimum but to ensure that samples at all of the six PT stages are rated. Both groups did receive audio-samples at all 6 PT stages.

The raters received the samples over a course of four weeks. On Mondays they were provided with a link to access the samples via the online storage tool

'dropbox'. In 'dropbox', I created folders for each week that contained the audio-files of the data presented in Table 11 below. The files were coded *learner XY*, as is displayed in the brackets. To illustrate this, a rater assigned to group A received a link to a folder that contains three audio-files, labeled learner 1, learner 2, learner 3, on a Monday. The rater then has five days to analyze the audio-samples and was contacted via email on a Friday to submit her/his results. In that way, the rater received two to three files per week and their work amounted to a total of four weeks. After I had received the rating results for one week, I deleted the link from 'dropbox' so that the raters were not able to go back to the files to listen to them again. Additionally, I instructed raters to delete the files after the ratings, should they have had downloaded them. This way, I tried to reduce the chance that raters were able to go back to the files and check for the level they had assigned to a file earlier. Table 11 displays the mode of presentation of the audio files.

Rater Group A	Week 1	Week 2	Week 3	Week 4
	Ko12 (Learner 1), PT: 1	<i>Ko05</i> (Learner 4), PT: 5+	Ko09 (Learner 8), PT: 3	Ke02 (Lerner 11)
	Ke09 (Learner 2)	Ko08 (Learner 5 + 6), PT: 2	Ke04 (Learner 9)	Ke05 (Lerner 12)
	<i>Ko04</i> (Learner 3), PT: 5	<i>Ko02</i> (Learner 7), PT: 4	Ko11(Learner 10), PT: 1	
Rater Group B	Week 1	Week 2	Week 3	Week 4
	Ko14 (Learner 13), PT: 1	<i>Ko10</i> (Learner 16), PT: 5+	Ko03 (Learner 20), PT: 3	Ke01 (Learner 23)
	Ke03 (Learner 14)	Ko07 (Learner 17 + 18), PT: 2	Ko13 (Learner 21), PT: 1	Ke10 (Learner 24)
	<i>Ko06</i> (Learner 15), PT: 5	<i>Ko01</i> (Learner 19), PT: 4	Ke06 Learner 22)	

Table 11: Mode of Presentation of Audio Files to Raters

To recapitulate, KoXX refers to all the original audio-files and KeXX represents the edited versions. The PT stages for the original samples are given, too. The audio-files (in italics) display all the edited versions of the audio-files. The file names

(printed in bold) show the case in which an edited file as presented before the raters had had access to the original version. The original samples, for which the edited versions were presented later, are highlighted in italics. I highlight whether the original version (italics) or the edited version (bold) was distributed first, because I assume that the order of presentation might have an influence on the rating results. This issue will be discussed later when presenting the outcomes of the study in chapter 4.

Since the number of files, when grouped together, is too small to allow for statistical randomization, I manually put the sequence of presentation to the raters into a structured order. Each rater group (A and B) was thus presented with at least one sample at each of the six PT stages.

In week one, both subgroups receive one sample at a lower PT stage, for which no edited version was produced, as well as one edited sample. In week 2, three original samples are presented to the raters. Two of their edited corresponding versions are given in week two. The dropbox folder for week 3 then contains the edited, original version respectively, of the samples that were presented in week 1. The underlying pattern is presented in Table 12 below:

Week 1	Week 2	Week 3	Week 4
/	C	D	A°
D°	/	/	C°
B	A	B°	

Table 12: Pattern of Sample Order for File Distribution

The small circle indicates the edited samples, and the remaining plain letters represent the other original samples. Samples that have no edited relative are depicted by a slash “/”. In using this scheme, I attempt to make sure that the raters had at least one week of time between the ratings of the original and the edited version of a file. The sequence of the samples within the week itself was also scrambled. Take sample Ke09 in rater group A as an example. It was presented to the raters in second position in week 1, whereas its original relative can be found in first position in week 3. Although I was not able to control for the

sequence in which the raters actually listen to the samples, I tried to influence them to use my proposed sequence by labeling the files chronologically: learner 1, learner 2, learner 3, etc. in the dropbox folder for each week. This can be seen in Table 12 above.

4.3.8 The Rater Groups

Chapter 4.3 introduced the two research foci of the present study. Focus b) relates to investigating how reliable the rating results are when raters are trained and when they use an assessment grid as opposed to pure intuition for their ratings. The claim I made in chapter 3.4.1 is that the results of human ratings are strongly influenced by the behavior of the raters, thus impeding the reliability of the results. I propose a combination of CEFR-based ratings and linguistic profiles based on PT can make results more reliable. To substantiate the claim of a need for the combination of the two assessment procedures, I investigate the role of the use of an assessment grid as well as the effect of rater training; i.e. the level of rater experience in more detail.

In order to shed light on whether the use of an assessment grid and/or the level of experience of raters have repercussions for the results of the rating in terms of variability, I divided the rater population into three different sub-groups, i.e. amateur raters, novice raters and experienced raters. I will now describe the classification of the three rater groups in more detail. Table 13 gives an overview of the participants in each of the three groups.

Name	No. of participants	Male/female
Amateur rater group	23	17 females 6 males
Novice rater group	22	18 females 8 males
Expert rater group	10	7 females 3 males

Table 13: Participants in Rater Groups

The amateur rater group and the novice rater group consist of teacher trainee students at Paderborn University. These two groups were selected, because at least an upper intermediate level of English language ability in both groups can be assumed. The entrance level for teacher training in English at Paderborn University currently is the CEFR level B2. Tentatively, one can thus assume a vaguely equal amount of prior experience with the language and with content knowledge. No native speakers of English were amongst raters in these two groups. To be more specific, the amateur raters, as well as the novice rater group, consist of students enrolled in undergraduate studies; Bachelor of Education, at Paderborn University who took part in two different teacher education seminars in the summer term of 2016.

The amateur rater group consists of 23 students, six male and 17 female students, enrolled in their fourth semester on average. The amateur rater group was given no instruction on how to rate learner language whatsoever. They were asked to assign 6 random letters from low proficiency to high proficiency (letters D to I) to the samples. I refrained from using the letter A-F because I assume that is generally associated with the top cut-off point and the raters should be free from any such associations for their ratings. The term proficiency itself or ways of language assessment were not discussed. A more detailed account to the instrument this group used for their rating will be given in section 4.3.9.

The novice rater group which was also enrolled in Bachelor of Education studies, includes 22 students of which four are male and 18 are female. In the summer term of 2016, they were enrolled in the fourth semester of their studies on average as well. In contrast to the amateur raters, the novice raters received instruction on the basis of the CEFR and on how to use the Global Oral Assessment grid following the model of the Manual for Aligning Tests with the CEFR (see CoE 2009). The Manual provides familiarization activities with the aim, amongst others, to “encourage increased transparency on the part of examination providers; [...] as well as the transparency of the content of examinations (theoretical rationale, aims of examinations, etc.)” (CoE 2009: 1). The authors of the Manual thus try to gear towards the need for more standardized procedures of using the CEFR as a reference tool for examination

alignment. Additionally, they try to adhere to issues that arise in the arbitrary way that examination providers refer to the CEFR in their tests, although they explicitly state that: “[t]he manual was not conceived as a tool for linking existing frameworks or scales to the CEFR, but the sets of procedures proposed might be useful in doing so” (CoE 2009: 2). There seems to be some inherent contradiction in this statement. Nevertheless, in the Manual, a familiarization seminar for test administrators is proposed in order to get a deeper understanding of the CEFR scales. Activities comprise 1) preparatory activities before the seminar, 2) activities at the familiarization seminar, and 3) a quantitative analysis of the CEFR scales and a preparation for rating the productive skills (see CoE 2009: 18).

Following this, the training of novice raters encompassed three 90-minute guided sessions as well as about 30-60 minutes of self-study and training. The time frame used for familiarization with the CEFR exceeds the specifications in the Manual. The Manual suggests calculating at least 180 minutes of familiarization training (CoE 2009:23). The authors of the Manual recommend the following course of action for familiarization with the CEFR:

<u>Familiarisation</u>	
•	can be organised independently from any other training activity, and can be recycled at the start of the Specification and the Standardisation activities.
•	takes about three hours:
-	Brief presentation of the CEFR Familiarisation seminar by the coordinator (30 mins)
-	Introductory activity (d-e) and discussion (45 mins)
-	Qualitative activity (f-g) including group work (45mins)
-	Preparation for rating (h-i) (45 mins)
-	Concluding (15 mins)

Figure 26: Time management for familiarization activities, taken from CoE (2009: 23)

In order to train the novice raters, the students were asked to gather as much information about the CEFR and its scales as possible and to bring notes of that knowledge back to the guided sessions. This was done before the workshop started. Additionally, students were supposed to read through the Global Scale of the CEFR and highlight the most important aspects in the scale. They were also supposed to read through the salient features in the CEFR section 3.6 (pp. 33-36)

as a preparatory activity. The content of the workshop is presented in Table 14 below. I present the page numbers of the CoE document in the last column so that the reader can relate each of the activities that the workshop provided in this study to the activities in the Manual.

Slot	Time frame	Activity	Relation to activity in Manual (CoE 2009: 18-23)
Preparation	30-60 mins	Gather information about the CEFR and highlight most important points in Global Scale.	Activity a), p. 18
Week 1: (Introductory Activities)	90 mins	Presentation of information gathered by participants. Brief input session by coordinator. Sorting Table A1 and highlight key elements in color. Self-assess quality of own foreign language(s) with the global CEFR scale.	Activity d), p. 20 Activity e), p. 21
Week 2: (Qualitative Analysis)	90 mins	Sorting individual descriptors from Global Oral Production scale. Reconstruction of CEFR global scale in which important elements were deleted and checking the outcome against those of others. Summary the most important aspects of the scale in own words. Brief introduction to proficiency rating procedures by coordinator.	Activity f), p. 21 Activity g), p. 21
Week 3: (Preparation for Rating)	90 mins	Reconstruction of the rating grid. Discussion of the reconstruction. Use of the grid with two samples and discussion of the rating.	Activity h), p. 22 Activity i), p. 22

Table 14: Overview of Novice Rater Training

Since the manual advises users to “select activities from each group at the start” (CoE 2009: 17) of the training, I refer to the letters that the Manual assigned to the activities in the last column of the table above. Generally, it was the aim of the workshop to become more familiar with the content specifications of the different CEFR levels and to gain some experience in rating oral language samples. In week 1, the participants presented the information they had gathered before the workshop to each other. The information was collected on

the blackboard and made available to the participants after the seminar in the form of a photo protocol. The workshop coordinator provided a brief input on the sources and development of the CEFR, its aims, as well as the horizontal and vertical dimension in the document. The participants then read through pages 33-36 of the CEFR in order to highlight the salient features in the CEFR scales. In a following quantitative analysis (CoE 2001: 21), the scales were focused. The participants were asked to sort the descriptors of the Global CEFR Scale (CoE 2001: 24), compare their results with those of other participants and highlight the most important descriptors contained in the scale. Then, the participants were instructed to assess their own proficiency in any of their foreign languages with the help of the Self-Assessment Grid (CoE 2001: 26f.). The Self-assessment grid is presented below.

		A1	A2	B1
U N D E R S T A N D I N G	Listening	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 phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g. very basic personal family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.	I can understand the main points of clear, standard speech on familiar matters, regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.
	Reading	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters	I can understand texts that consist mainly of high frequency everyday or job-related language. I can understand the description of events, feelings and wishes in personal letters.
S P E A K I N G	Spoken Interaction	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).
	Spoken Production	I can use simple phrases and sentences to describe where I live and people I know.	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.
W R I T I N G	Writing	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.	I can write short, simple notes and messages relating to matters in areas of immediate needs. I can write a very simple personal letter, for example thanking someone for something	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.

Figure 27: Self-Assessment Grid part one, taken from CoE (2001: 26f.)

		B2	C1	C2
U N D E R S T A N D I N G	Listening	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. I can understand television programmes and films without too much effort.	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
	Reading	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.	I can understand long and complex factual and literary texts, appreciating distinctions of style. I can understand specialised articles and longer technical instructions, even when they do not relate to my field	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.
S P E A K I N G	Spoken Interaction	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.	I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skilfully to those of other speakers.	I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently and convey finer shades of meaning precisely. If I do have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it.
	Spoken Production	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	I can present a clear, smoothly-flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.
W R I T I N G	Writing	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.	I can express myself in clear, wellstructured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind.	I can write clear, smoothly-flowing text in an appropriate style. I can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

Figure 28: Self-Assessment Grid part two, taken from CoE (2001: 26f.)

After the self-assessment, the difficulties in assessing their proficiency were discussed in the group. As a homework exercise, the participants were asked to compare the CEFR scales for Spoken Production (CoE 2001: 58) and Spoken Interaction (CoE 2001: 74).

In week 2, the participants were made aware of the finer shades of meaning in the descriptors by comparing the different types of scales. Then, the participants were presented with the Global Scale (see chapter 2.1.3) in which some of the descriptors were missing. They were supposed to reconstruct the Scale by filling in the blanks with the appropriate descriptors. The outcome was discussed with the peers and checked against the original Global Scale. Following this, the participants were asked to summarize the Global Scale in their own words. Afterwards, the coordinator gave a brief introduction to proficiency rating procedures. Homework for the next workshop day was to read the Global Oral Assessment Grid carefully (CoE 2009: 185).

In week 3, all the descriptors in the Global Oral Assessment Scale were discussed. The students were presented with snippets of the descriptors and asked to sort them according to the appropriate CEFR level in the grid. This reconstruction was then discussed. Afterwards, the participants were presented with an audio-file similar to those in the actual study and asked to rate the learner's CEFR level. The audio-file displays the features of the standard diagnostic assessment procedure for Rapid Profile assessment. The results of the students' assessments were discussed in a plenary session. The participants were then given access to another audio-file which they were supposed to rate at home. The result of their rating was submitted to the coordinator who gave individual feedback on their rating. In using the familiarization activities as suggested in the Manual (2009), it was ensured that all study participants had received the same amount of instruction on the CEFR and proficiency ratings. Based on this limited experience with proficiency ratings, I consider this rater group to be novice raters. The only common basis the raters in the amateur and the novice group have, is that they were all teacher trainees at undergraduate level at a German university at the time of the data collection. All of the students chosen to take part in the study claimed to not have had any experience in grading/rating learner language before.

The third group, the experienced rater group is a rather heterogeneous group of raters who are affiliated to either TELC, Cambridge, IELTS or the Zentrum für Sprachlehre (ZfS) at Paderborn University. The latter group frequently

administers proficiency placement tests for the Deutscher Akademischer Austauschdienst (DAAD). Raters qualify for this group after having executed at least 10 ratings prior to the study and were currently involved with proficiency ratings along the lines of the test centers mentioned above. The experienced rater group consists of 10 raters, three male and seven female of different age groups. Six out of the 10 raters are native speakers of English. Table 15 gives an overview of the different rater affiliations at the time of the data collection.

Affiliation	Rater Code
TELC	aE02, aE03, aE04, aE05, bE01, bE03,
Cambridge/IELTS	bE04
ZfS Paderborn	aE01, bE02, bE05

Table 15: Overview Expert Rater Affiliations

It was important to ensure that this rater group has had experience with either rating procedures administered by TELC, Cambridge or the DAAD because these proficiency tests are aligned with the CEFR descriptors and calibrated towards assessing CEFR levels on different scales and skill levels. This way, a sufficient amount of familiarity with both the CEFR descriptors and the proficiency rating procedure can be assumed.

After having described the participant groups, I outline the assessment grids that the novice and expert raters used to assign the CEFR levels to the audio-files of learner language that I provided. I also briefly outline the assessment table that amateur raters used.

4.3.9 The Rating Schemes and the Novice Rater Training

This chapter describes the assessment grids that the different groups used for their assessment.

The amateur raters did not assign the actual CEFR level label, but rather a random letter (D for the lowest level of language proficiency and I for the highest level of language proficiency) to the audio-recordings. This method is used with the amateur rater group, because the aim is to investigate a most intuitive

approach to rating the samples without any previous instruction on how to administer ratings of learner language. This intuitive approach is later compared to the use of an assessment grid by the amateur and the novice rater group. Table 16 gives an overview of the letters that the amateur rater group used for their assessment of the audio-files.

Letter	CEFR level
D	Below A1/A1
E	A2
F	B1
G	B2
H	C1
I	C2

Table 16: Letters used by amateur raters and according CEFR level

The amateur rater group was asked to assign six random letters to the audio-recording. No content specifications for the letters are given and thus, no connection to the CEFR levels can be assumed. Please note that since this group did not use descriptors to quantify the level content, but only the 6 arbitrary letters, no beginning or cut-off point for the scale could be formulated. Therefore, D might represent both “below A1” and “A1”. The other rater groups were able to distinguish between “below A1” and “A1”.

The second and the third group, the novice and the expert raters used the grid presented in Table 17 below for their assessment of the audio-files.

C2	<p>Conveys finer shades of meaning precisely and naturally.</p> <p>Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.</p>
C1	<p>Shows fluent, spontaneous expression in clear, well-structured speech.</p> <p>Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy, errors are rare.</p>
B2+	
B2	<p>Expresses points of view without noticeable strain.</p> <p>Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.</p>
B1+	
B1	<p>Relates comprehensibly the main points he/she wants to make.</p> <p>Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, [sic!] sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.</p>
A2+	
A2	<p>Relates basic information on, e.g. work, family, free-time, etc.</p> <p>Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.</p>
A1	<p>Makes simple statements on personal details and very familiar topics.</p> <p>Can make him/herself understood in a simple way, asking and answering questions about personal details, provided the other person talks slowly and clearly and is prepared to help. Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words.</p>
Below A1	Does not reach the standard for A1.

Table 17: Global Oral Assessment Scale, taken from CoE (2009: 184)

This Assessment Grid is produced by the CoE authorities and presented in the official Manual for Relating Examinations to the CEFR (2009). It is aligned to the Overall Production Scale of the CEFR (see section 4.3.8). This grid, along with the

complementary grid, was used in the study because a) all of the raters were supposed to use the same assessment grid, b) it originates from an official CoE source so that a maximum of alignment between the rating rubric and the CEFR descriptive scale may be assumed, c) it covers grammatical accuracy as one aspect of five in overall oral production so that the role of grammatical accuracy in these ratings may be elicited (see below for more details) and d) PT focuses mainly on oral production of language learners. I assume that because of these aspects, claims about the compatibility of the CEFR and PT can be made. These points are discussed below.

a) In particular, the expert raters come from various different rating backgrounds, ranging from TELC raters to Cambridge and DAAD raters (see chapter 4.3.9 for more detail). Since these assessment centers use different rating rubrics calibrated to their assessment ideals, I wanted all raters to mainly rely on one specific rating grid and not on their own criteria. With the presentation of this rating rubric (Tables 16 and 17), I opted for controlling the use of the same rating criteria across the raters in order to ensure that the same assessment criteria are used. Additionally, I asked the raters to briefly describe the technique they employed while rating the audio-files. With this question, I aimed to elicit whether some raters might additionally use another rating rubric. If that had been the case, those raters would have been excluded from the study. None of the raters reported using additional criteria or rating grids. However, it cannot be determined in how far their rating experience and usual rating procedure was implicitly applied to my data.

b) The Global Oral Assessment Scale was produced by the CoE and published in the Manual for Relating Language Examinations with the CEFR (CoE 2009). Since the regular scales presented in the CEFR are descriptive scales and not rating grids, I asked the raters to use this assessment scale that was particularly designed for the assessment of Global Oral Production based on the CEFR descriptive scale for Oral Production.

However, Deygers & Gorp (2015) showed that a CEFR-based rating scale that was constructed together with raters for usage during their ratings, did not guarantee a uniform interpretation of the descriptors, despite high inter-rater

reliabilities. Harsch & Rupp (2011) therefore argue that one needs a high level of analytic detail in CEFR-based scales in order to compensate for the broadness of the initial descriptors. However, to my knowledge there is no better assessment scale available that can be used for this study, so I need to rely on the official document. I consider this assessment scale useful and appropriate for my study because it comprises the language features Range, Accuracy, Fluency, Interaction and Coherence. Accuracy, in this regard, refers to the accuracy of grammatical features in the learner language (CoE 2009: 185). This can be seen in the complementary assessment grid for Global Oral Production below that was also used for the rating.

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn-taking, referencing, allusion making, etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organizational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional, or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his (sic!) remarks in order to get to keep the floor and to relate his/her own contributions skillfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, (sic!) and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns or expressions, there are a few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground around confirming comprehension (sic!), inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some jumpiness in along contribution.
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is evident, especially in	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

	family, hobbies and interests, work travel and current events.		longer stretches of free production.	part of what someone has said to confirm mutual understanding.	
A2+					
A2	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulations are very evident.	Can ask and answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like “and (sic)”, “but” and “because”.
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like “and” or “then”.

Table 18: Complementary Grid for Global Oral Assessment, taken from CoE (2009: 185)

The column *Accuracy* in the complementary grid displays the assessment standards for grammatical accuracy. A rising amount of control over gradually more complex grammar can be seen from Level A1 to Level C2. Level A1 is characterized by the learner displaying limited control over a few grammatical structures and sentence patterns in a memorized way (see CoE 2009: 185). For Level C2, the rater using this grid is asked to assess a consistent grammatical control of complex language when attention is otherwise engaged.

One could argue that it may be more appropriate to use a grid for grammatical accuracy in this study to investigate research question H2 (see chapter 4.2); i.e. interfaces between the CEFR and PT. However, I was not able to find a rating grid specifically produced for the CEFR scale for grammatical accuracy by the CoE. Therefore, I decided to use the Global Oral Assessment Grid that encompasses grammatical accuracy as one feature. This fact, combined with points c) and d) described below, lead me to conclude that the Global Oral Assessment Grid, along with its complementary grid, is appropriate for this study.

c) By using the grids described above, I assume that it is assured that *Accuracy* is only one feature that the raters should assess. In the grid, no ranking of importance of one language feature over another features is visible. This way, I assume that it can be elicited whether the accuracy of grammatical features plays an overriding role in language proficiency ratings that aim at Overall Oral Production, including Range, Fluency, Interaction and Coherence. This relates to my hypothesis H2; that grammar plays a crucial role in determining the proficiency level of a learner. If it is the case, that grammar plays a primary role in the ratings, it shows that it is important to reevaluate the status of the CEFR scale for grammatical accuracy and to underpin it with empirical SLA research.

d) PT as a theory of SLA is mainly concerned with the oral production of language learners. This is why the data elicitation for PT-based profiles mainly happens based on semi-communicative tasks (see section 2.2.7 for more detail) that implicitly trigger the production of grammatical features along the PT hierarchy. An issue in this regard is whether the task-based design used in the audio-files (that is used for eliciting PT-related structures) is appropriate for general language proficiency ratings. To explore this issue, I played back three of my recordings to an experienced Cambridge rater before the main data collection phase and asked him whether the data was dense enough to be rated by means of the Global Oral Assessment Scale. He stated that a rating is possible. Additionally, I asked all of the raters who participated in my study to comment on the way they approached the rating, i.e. their rating techniques, and to comment on any peculiarities in the audio-files. One rater in the expert rater group reported the following:

bE04	<p>In a different note, there are some questionable testing techniques being used during the making of the recordings, though I am guessing that the way that the rateable language is collected isn't really relevant for your study. Nevertheless, if we take the final recording, there is a possibility that this candidate might do better and produce more authentic language doing a standardised testing rubric like Cambridge or TOEFL. Because of the tasks that the candidate is required to do, I don't really feel the candidate is given an opportunity to talk at greater length. Nevertheless, so far, I feel that in each case the recordings contain enough rateable language to justify the level ratings that I have given.</p>
------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Rater bE04 comments on recording Ko06 on PT stage 5. This file was rated a B1 level by this rater. Although he comments on the uncommon way of data elicitation, he admits in the end that the data contain enough ratable language. For this reason, I concluded that the use of the task-based data for CEFR-based ratings is feasible.

After having described the audio-files, the way in which the files were manipulated for grammatical accuracy, the raters who participated in the study and the order in which the data was distributed to the raters, the results of the study will be presented in following chapter.

4.4 Results

The following chapter describes the results of the empirical investigation of a) the difference between assigned CEFR levels to original and edited speech samples, b) the relationship of the CEFR levels and PT stages for the original speech samples, and c) differences between the three rater groups in relation to their assessment of the speech samples. This is accompanied by group differences in terms of inter-rater reliability. Both statistical measures for issues a) - c), and a qualitative analysis of aspects of point b) will be given in this chapter.

In total, 53 raters participated in this study. They were subdivided into three groups: 22 amateur raters, 21 novice raters and 10 expert raters. The raters rated 22 audio-files of oral language samples of learners of English in total with the help of the Global Oral Production Grid provided by the CEFR. The 22 audio-files encompass 14 original files that represent learners at PT stages 1 to 5+ and eight edited files at PT stages 3 to 5+. The PT stages had been determined prior to this study by means of Rapid Profile. Thus, 424 ratings were made by the 53 raters in total (see Table 19 for more details).

It is to be noted that all results have to be treated with caution because the number of raters, who participated in this study, is relatively small. Also, only two files represent each PT stage and therefore the relationship between PT

stages and CEFR levels that is depicted in this study needs be regarded as tentative.

4.4.1 The Effect of Grammatical Accuracy – Original and Edited Speech Samples

In the following section, I will present the results on how the raters rated the original speech samples and their edited version, i.e. effect of the grammatically edited speech samples on rating results in contrast to the results of their original versions. To recapitulate: The idea behind this way of approaching the data is – in comparison to, e.g. immediate-retrospection (see e.g. May 2006, Weir et al. 2009) or verbal protocols (see e.g. Edorsy 2004, Joe et al. 2011) - to employ a more direct way of measuring the influence of grammatical accuracy in oral production ratings. I thus compare the CEFR levels assigned to the original speech samples to those CEFR levels allocated for their edited corresponding versions by the raters. The edited speech samples differ from the original ones only in terms of grammatical accuracy. The raters used the Global Oral Assessment Grid that comprises the features Range, Fluency, Accuracy and Coherence. One can see that *Accuracy* is but one language feature included in this scale. Should the edited speech samples be rated on a lower level than its corresponding original, then the Grammatical Accuracy-variable might be assumed to determine this difference.

I will present the results for each of the levels of experience of raters and depict in how far the results are different for each of the rater groups. Please note here that the files for PT stages 1 and 2 had to be excluded for this calculation because there were no edited samples present for those stages (see Table 19). Learners at PT stages 1 and 2 do not produce interlanguage features that display enough morphological variation to be edited.

All raters (n= 53)			Amateur raters (n=22)			Novice raters (n=21)			Expert raters (n=10)		
Edi	Edi	Edi	Edi	Edi	Edi	Edi	Edi	Edi	Edi	Edi	Edi
<	=	>	<	=	>	<	=	>	<	=	>
Org	Org	Org	Org	Org	Org	Org	Org	Org	Org	Org	Org
84	54	74	28	22	38	40	19	25	13	13	14
Total = 212			Total = 88			Total = 84			Total = 40		
z= -3.259			z= -1.420			z= -3.669			z=- .092		
p= .001			p= .156			p= .000			p= .927		

Table 19: Results of Rating - Comparison of Original Files and Edited Files

53 raters participated in total; this group is depicted in the most left column. Next to the column showing “All raters”, the numbers for the three rater sub-groups are given. The two rows below the rater groups depict the relationship between the original and edited sample. Here, “Edi<Org” represents all cases in which the edited version of the speech sample is rated on a lower level than its corresponding original sample. “Edi=Org” represents the cases in which both samples are rated on the same level. “Edi>Org” shows in how many cases the edited samples are rated on a higher level than the original. The second-last row gives the total number of original-edit pairs that were rated. A Wilcoxon Signed Rank test was run to determine whether there are significant differences between the original and the edited samples in terms of assigned CEFR levels. The test uses a p-value of > .05. The data show a statistically significant difference in medians for the original and the edited speech samples at $z = -3.259$ and $p = .001$ across all three groups. The last row shows the p-values and z-scores indicating whether the difference in assigned CEFR levels to the original and edited recordings is statistically significant.

Of the 212 original-edit pairs that were rated by all three experience level groups in total (see most left columns), 84 edited files are rated on a lower CEFR level than the original sample. Seventy-four samples are rated on a higher CEFR level than their edited version. In 54 cases, there is no difference between the rating for the original and edited language sample.

It is interesting then to investigate whether the different rater groups show any differences in assigning CEFR levels to original and edited speech samples for Grammatical Accuracy. The amateur rater group who had received no instruction on how to assess learner language and who had not used an assessment grid, rated 88 original-edit pairs in total. 22 samples show no difference in assigned levels. It is to be kept in mind that these raters did not assign CEFR levels but only a range of 6 letters; *D* representing the lowest level of proficiency and *I* representing the highest level of proficiency, see Table 16. Twenty-eight edited speech samples are rated on a lower level than their original counterpart by the amateur rater group. Thirty-eight edited samples are rated higher than their corresponding original version. This yields a statistically insignificant result of median difference at $z = -1,420$ and $p = .156$.

The novice rater group behaves differently. This group had received the minimum amount of instruction, as suggested by the Manual of Relating Examinations to the CEFR (CoE 2009), prior to the rating and used the assessment grid for Overall Oral Production. The novice rater group displays a significant difference in medians between the original and edited sound files at $z = -3,669$ and $p = .000$. In this group, there are 25 ties in assigned CEFR levels to original and edited files. Forty edited speech samples are rated lower than their original counterpart and 19 original samples are rated to be on a higher level than their respective edited version.

The results of the expert raters, the smallest group of participants with $n = 10$ and 40 original and edited speech sample-pairs that were rated, shows no statistically significant difference between original and edited versions: $z = -0.92$ and $p = .927$. In total, there are 14 ties, 13 instances in which the edited samples are rated lower than the original samples and 13 cases in which the original is rated higher than the edit.

After each rating procedure, I asked the raters to comment on any peculiarities that they might have noticed during the ratings. With this question, I intended to elicit whether a rater might have recognized that they had listened to a similar audio-file after having received a corresponding original or edited speech sample. Interestingly, only a few raters reported back to me that they had

noticed a degree of familiarity with some of the files. None of the raters, however, realized that the files were edited in some form. Six out of 53 raters asked whether I had mistakenly uploaded the same file that I had already presented to them via Dropbox before. They, however, did not realize that it was an edited speech sample that they had received. Since the raters were not able to get hold of the audio-files that they had rated before in order to check for their rating (see chapter 4.3.7 for more details on the manner of distribution of the files), I conclude that the six samples that the raters commented on should not be excluded from the analysis.

The next chapter employs a more qualitative comparison of original and edited speech samples.

4.4.2 Qualitative Comparison of Original and Edited Samples

The following table provides an overview of the *range* and *mode* of all CEFR levels assigned to audio-files as well as the accorded percentages of the *mode* for each of the original-edited pairs. I present this table because it shows the tendencies with which raters assigned the CEFR levels to the original and the edited recordings, i.e. whether raters tended to place the edited speech samples on higher or lower levels. This tendency is reflected better in terms of the *range* of levels assigned to the recordings, rather than based on the *mode* of the assigned levels alone.

Range refers to the dispersion of all CEFR levels that raters gave for the respective recording. Both, the highest and the lowest CEFR levels are shown in this column. However, with the levels presented in *Range*, it does not mean that all CEFR levels in between the highest and the lowest level were *de facto* assigned by the raters. It may be the case that the lowest CEFR level given to a sample is A1 and the highest is B2, but no rater gave a B1 level for this sample. *Mode* represents the most frequent level that was assigned to a recording by the respective raters. The Table is subdivided into the three experience levels of rater groups. The brackets indicate how often the CEFR level had been assigned by raters. The percentage column indicates the percentage of agreement on the

level displayed in the *Mode* column within the rater group. Additionally, the PT stages for the original samples are displayed in the second column. It is to be borne in mind that no edited samples were generated for files at PT stages 1 and 2, which is why those files are not given in the table below. To recapitulate, the amateur raters did not use the actual CEFR level grids and labels for their ratings but six arbitrary letters (see Table 16 for more details).

As amateur raters only used the letter labels, an issue is that no distinction between “below A1” and “A1” could be made. The other rater groups were able to distinguish between these levels. To compare the rater groups, “below A1” and “A1” are summarized together and treated as “A1”. For the same reason as mentioned above, the amateur rater group was not able to assign plus levels, whereas the other groups were able to use plus levels when they felt that the learner in the audio-file performed beyond the descriptors for one CEFR level, but not yet according to the next level. As explained in chapter 4.3.7, the three rater groups were again subdivided into subgroup A and group B so as to reduce the workload for each rater. This is why the number given in the *mode* column has to be read in conjunction with the number of participants in the sub-groups (see section 4.3.7 for more details). See file Ke01 for example. The expert rater group comprised 10 raters, but this group was split in half so that only five raters rated file Ke01. For Ke01, all expert raters agreed on the level B1 which is why the % column displays 100% agreement on mode B1 for this edited file. The full Table is presented below.

		Amateur raters (n=22)			Novice raters (n=21)			Expert raters (n=10)		
File	PT	Range: low-high	Mode	%	Range: low-high	Mode	%	Range: low-high	Mode	%
Ko03	3	D-F	F (5)	62,50%	A1-A2	A2 (8)	72,73%	A1-A2+	A1 (2) A2+ (2)	40,00%
Ke03		E-F	E (4) F (4)	50,00%	A2-B1	A2 (6)	54,55%	A2	A2 (5)	100,00%
Ko09	3	E-H	F (7)	50,00%	A2-B1	A2 (8)	80,00%	A1-B1	A2 (3)	60,00%
Ke09		D-F	F (7)	50,00%	A1-A2+	A2 (6)	60,00%	A2-A2+	A2 (4)	80,00%
Ko01	4	E-H	F (4)	50,00%	A2+-B2	B1 (6)	54,55%	A2-B1+	A2 (2) B1 (2)	40,00%
Ke01		F-H	F (3) G (3)	37,50%	A2-B1+	A2 (4)	36,36%	B1	B1 (5)	100,00%
Ko02	4	D-H	F (6)	42,86%	A2-B1	A2 (4)	40,00%	A1-A2+	A2 (3)	60,00%
Ke02		E-G	F (6)	42,86%	A1-B1	A2 (3) B1 (3)	30,00%	A2+-B1	B1 (4)	80,00%
Ko04	5	F-H	H (6)	42,86%	B1-B2	B1 (6)	60,00%	B1+-B2+	B1+ (2) B2+ (2)	40,00%
Ke04		F-I	G (5) H (5)	35,71%	A2+-B2	B1+ (4)	40,00%	B1-B1+	B1 (3)	60,00%
Ko06	5	F-I	G (5)	62,50%	B1-B2	B1 (5)	45,45%	B1-B1+	B1 (3)	60,00%
Ke06		E-G	G (6)	75,00%	A2-B2	B1 (5)	45,45%	A2-B1+	B1 (3)	60,00%
Ko05	5+	G-H	G (5)	35,71%	B1-B2	B1 (4) B2 (4)	40,00%	A2-B2	B1 (3)	60,00%
Ke05		G-H	G (6)	42,86%	A2-B2	A2 (5)	50,00%	A2-B2	B1 (2)	40,00%
Ko10	5+	G-I	H (4)	50,00%	B2-C2	C1 (6)	54,55%	B2-C1	C1 (3)	60,00%
Ke10		E-I	G (4)	50,00%	A2-C1	B1 (3)	27,27%	B1-B2	B2 (3)	60,00%

Table 20: Results Ratings - Range and Mode for CEFR Ratings for PT stages

As another illustration for this table, consider the original file Ko03 that represents a learner at PT stage 3. *Mode* shows that, with five indications, the amateur raters assigned the letter F to this file. The *range* that was assigned to this file encompasses letters D and F. Ko03's corresponding edited file Ke03 was assigned a *range* between letters E and F. These levels are also the most frequent levels given by this group; four times E and four times F. So, whereas *mode* for the original sample only displays letter F, *mode* for the edited sample also shows letter E. It thus seems that the amateur rater group displays a tendency to rate sample Ke03 on a lower level in comparison to the original sample Ko03.

The novice rater group behaved differently with the same sample set (K03). They generally assigned lower levels for the original sample (range A1-A2), with eight of the novice raters agreeing on the A2 level. The mode for its edited version Ke03 is also A2 (6), but the range that was assigned by the novice raters is one level higher (A2-B1) than which was given to the original (A1-B2).

At two indications each, the expert rater group assigned levels A1 and A2+ to the original audio file Ko03 which, at the same time, represents the range of levels given for this recording. The edited version seems to find more agreement amongst the expert raters, as all of them assigned the A2 level to this recoding. To summarize, when taking *range* and *mode* as a point of departure, only a very small tendency to down-rate the edited file Ke03 can be seen with the amateur rater group and the expert rater group. No difference between Ko03 and Ke03 is visible with the novice rater group.

After having illustrated the Table above, the following section compares the original and edited files based on *mode*.

4.4.2.1 Comparison of Original and Edited Files based on Mode

Mode represents the CEFR levels that were assigned to the audio-files by the raters and rater sub-groups most frequently. Based on mode, files Ko10 and Ke10 is the only pair in which a tendency to rate the edited version lower than the original file can be observed across all three sub-groups.

Mode for Files Ko10 and Ke10			
Amateur Raters	Ko10	H	
	Ke10	G	= -1 level
Novice Raters	Ko10	C1	
	Ke10	B1	= - 2 levels
Expert Raters	Ko10	C1	
	Ke10	B2	= - 1 level

Table 21: Mode for Ko10 and Ke10

All rater groups seem to agree that the original file, Ko10, can be placed on level H; i.e. C1 respectively. It is to be borne in mind though that only four amateur raters, six novice raters and three expert raters represent mode in this instance, so all results and their implications have to be viewed rather cautiously. The amateur and the expert raters rated Ke10 one level below the original version. The novice raters assigned the B1 level to this file, which is minus two levels. The learner in file Ko10 is located at PT stage 5+.

For files Ko05/Ke05 and Ko01/Ke01, only the novice rater group rated the edited sample one stage below the original sample:

Novice Rater Group			
Ko05	Ke05	Ko01	Ke01
B1 (4), B2 (4)	A2 (5)	B1 (6)	A2 (4)

Table 22: Comparison of Mode of files K05 ans K01 by the Amateur Rater Groups

The amateur rater group is the only group which rated a difference in file Ko03/Ke03: Ko03: B1 (F) (5), Ke03: A2 (E) (4) and B1 (F) (4). These results will be discussed in the chapter on uneven profiles (4.4.3.3).

Based on mode alone, there is only one instance, i.e. Ko10, in which all rater groups rated the edited speech sample lower than the original. This might be due to the strong discrepancy between the elaborate vocabulary, discourse and phonology present in the original sample, and the poor morphology found in the edited speech sample. In the data, there are four more files for which the edited version was down-rated, but not across all groups. Generally, the amateur

and the novice raters seem to tend to rate the edited files on a lower level than the original in comparison to the expert raters.

Next, a comparison between the original and edited files based on range will be presented.

4.4.2.2 Comparison of Original and Edited Files based on Range

Range is based on all indications of CEFR levels assigned to the audio-files by the raters and rater sub-groups. When using range as a criterion for the comparison of original and edited speech samples, the tendency in the direction of rating results becomes clearer. These results tie in with the results on the agreement within the rater groups presented in section 4.5.3.2. In general, a tendency to rate the edited speech sample lower than the original can be seen. The chapter above described this tendency based on mode, i.e. the CEFR level that was assigned most often by the raters (and rater sub-groups). When only taking mode into account, the picture of the tendency is not fully painted because mode neither covers the top levels that were given to one sample, nor the bottom levels assigned to the sample. Range however encompasses top as well as bottom levels. Thus, it can be seen whether raters tended to rate the edited sample towards the top levels or the bottom levels on the continuum of levels that can be possibly assigned to the audio-file. Considering all subgroups, mode shows that eight edited samples were rated on lower levels than the original sample. When taking range into account for all sub-groups, the edited speech sample was rated lower 12 times.

The dispersion of the levels assigned by the amateur raters to the audio-files is presented below. For the sake of comparability, I refer to the CEFR levels for this group and not the letters that they had *de facto* used for their rating (see section 4.3.9 for more details). In the amateur rater group, the files Ko09/Ke09, Ko02/Ke02, Ko06/Ke06 and Ko10/Ke10 show a difference in assigned levels as determined by range, see Table 23. That is, in six out of 16 cases, either the top or the bottom levels assigned to the audio files are rated on lower levels than the

original files. By top levels, I mean the highest levels that were assigned to each file across the different raters and bottom levels refer to the lowest levels that were given by raters. For samples Ko09/Ke09 as well as Ko06/Ke06 both top and bottom levels are rated lower than the original. In the case of Ko02/Ke02, only the top level that was assigned to the edited version was rated lower than the original (B2 instead of C1 for the original). This group shows an equal number of down-ratings in the edited sample for top and bottom levels (three times top level and three times bottom levels) as indicated by the ellipses in the tables.

Amateur Raters															
Ko03	Ke03	Ko09	Ke09	Ko01	Ke01	Ko02	Ke02	Ko04	Ke04	Ko06	Ke06	Ko05	Ke05	Ko10	Ke10
A1	A2	A2	A1	A2	B1	A1	A2	B1	B1	B1	A2	B1	B1	B2	A2
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
B1	B1	C1	B1	C1	C1	C1	B2	C1	C2	C2	B2	C1	C1	C2	C2

Table 23: Original - Edit Comparison based on Range for Amateur Raters

Table 24 below depicts that the novice rater group displays a tendency to rate the bottom levels lower than the top levels. In seven out of 16 cases, the novice raters rated the bottom levels for the edited speech samples lower than their corresponding original sample. In the case of speech sample pairs Ko09/Ke09 and Ko10/Ke10 both bottom and top levels assigned to the edited sample vary from the original sample. That is, A1 instead of A2 and A2+ instead of B1 for Ke09, as well as A2 as compared to B2 and B1, instead of C2 for Ke10.

Novice Raters															
Ko03	Ke03	Ko09	Ke09	Ko01	Ke01	Ko02	Ke02	Ko04	Ke04	Ko06	Ke06	Ko05	Ke05	Ko10	Ke10
A1	A2	A2	A1	A2+	A2	A2	A1	B1	A2+	B1	A2	B1	A2	B2	A2
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A2	B1	B1	A2+	B1	B1+	B1	B1	B2	B2	B2	B2	B2	B2	C2	C1

Table 24: Original - Edit Comparison based on Range for Novice Raters

It is the novice rater group that shows the strongest tendency to rate the edited samples on a lower level than the original samples (see also section 4.4.2.1).

The expert raters rated seven/16 edited files on lower levels. They seem to down-rate the top levels (five times) rather than the bottom levels (twice). The expert rater group thus tends to alter their rating rather downwards the scale

than upwards the scale, like the novice raters seem to do as can be seen from Table 25 below.

Expert Raters															
Ko03	Ke03	Ko09	Ke09	Ko01	Ke01	Ko02	Ke02	Ko04	Ke04	Ko06	Ke06	Ko05	Ke05	Ko10	Ke10
A1	A2	A1	A2	A2	B1	A1	A2+	B1+	B1	A2	A2	A2	A2	B2	B1
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
A2+	A2	B1	A2+	B1+	B1	A2+	B1	B2+	B1+	B1+	B1+	B2	B2	C1	B2

Table 25: Original - Edit Comparison based on Range for Expert Raters

When using range for the comparison of edited and original samples, it can be seen that more edited samples are rated on lower levels than based on mode alone. The expert rater group shows a tendency to alter the top levels for their ratings of the edited files downwards the scale. The novice raters seem to rather rate the bottom levels downwards and the amateur raters tend to do both; rate top and bottom levels downwards. As discussed in Eckes (2005), general types of raters who tend towards severity or leniency, halo, or central tendency are factors in human ratings that are independent of the rating grids used. These factors seem to simply be part of human nature and need to be considered in standardized rating procedures. To what extent every single rater tended towards one of these characteristics cannot be determined on the basis of the present data. However, general tendencies can be found in the data as described above. These tendencies are likely to have influenced the rating results.

The following chapter examines the differences between the rater groups in more detail.

4.4.3 Discussion of Results on Grammatical Accuracy

Above, I have laid out that all results presented in this thesis have to be treated with caution because the number of raters in this study is quite small. Especially the results of the expert rater group need to be seen as tentative, because of the inherent heterogeneity in this group. It is desirable to employ future studies with bigger and more homogenous expert rater groups. Also, the explanations that I presented for the results (see section 4.4) are all generated post-factually and an

empirical investigation of these explanations could not have been employed based on the data. In the following chapter, I want to comment on further factors that might have caused the variability in the results due to the effect of the manipulation of grammatical features in the edited speech samples as compared to the original speech samples.

4.4.3.1 Discussion of Original and Edited Files across Rater Groups

When comparing the effect of the manipulation of grammar in different rater sub-groups, major differences can be seen. It is to say that the amateur rater group and the novice rater group are rather homogeneous groups (see chapter 4.3.9 for more detail) whereas the expert rater group consists of only 10 raters with varying assessment backgrounds. The only criteria that the expert raters had to fulfill, are that they had to have assessed at least 10 speakers according to the CEFR, and that they were affiliated to an assessment center at the point of data collection. Thus, the raters come from different backgrounds, such as TELC, Cambridge, and the ZfS at Paderborn University. The heterogeneity and small number of participants in the group might be responsible for the variability present in their ratings. This poses a threat to the comparability of their results with those of the other groups. Nevertheless, I assume that general tendencies in the rating behavior are visible in the data. I present the amateur and the expert raters first and only then comment on the novice raters, because the latter group shows differences as compared to the other two groups.

The amateur rater group rated the audio files based on intuition; i.e. they did not use an assessment grid and had had no experience in proficiency ratings prior to the study. This group tends to rate the edited sample on a higher level than the original sample:

Amateur Rater Group (n=22)		
Edi < Org	Edi = Org	Edi > Org
28	22	38
$z = -1.420, p = .156$		

Table 26: Edit-Original Rating in Amateur Rater Group

It seems to be the case that the raters in this group tend to revise their rating towards higher levels when presented with an edited sample as they rated 38 cases on higher levels than the original files. This is quite surprising since one would have assumed that language samples containing more grammatical mistakes/errors would most likely not be rated to be better than samples with less mistakes/errors. Since the amateur raters did not have an assessment grid to rely on, it might have been hard for them to compare the learner language and fully determine as to what a proficiency levels comprises. Maybe one day they felt that D was characterized by features X, and another day they added or deleted features to their internal rating scheme. One could argue that these ratings therefore might rely on most arbitrary factors which cause this tendency towards higher levels for the edited sample. However, in the other sub-groups, there are also instances in which raters rate the edited sound file on higher levels than the original version. That is why I assume that the use of intuition in proficiency ratings, in this study alone, cannot be held responsible for rating edited samples on higher levels than original samples.

The expert rater group also quite frequently rated edited sound files on higher levels than the original, at 14 times as can be seen from Table 29 below.

Expert raters (n=10)		
Edi < Org	Edi = Org	Edi > Org
13	13	14
z=- .092, p= .927		

Table 27: Edit-Original Rating in Expert Rater Group

It might also be the case that raters did not even subconsciously react to the changes in the data. However, finding explanations as to why the amateur rater group rated edited files on higher levels than their original files does not seem purposeful since the data do not permit any clear conclusions. This finding should be investigated in future research. A qualitative account with interviews might be most beneficial for this. However, it seems to be the case that it cannot be

attributed to the lack of assessment criteria alone which causes the arbitrary rating results in the amateur rater group.

The expert rater group that is presented in Table 27 above produces most random results when comparing the assigned CEFR levels to original and edited sound files. All possibilities account for about a third of the data; Edi < Org: 13, Edi = Org: 13, Edi > Org: 14. Again, it cannot be determined why raters would counterintuitively rate edited speech samples on higher levels than original speech samples that contain less errors and mistakes. However, the arbitrariness in the results for this group might be due to the fact that it is very small and most heterogeneous. Only 10 expert raters participated in the study. They also come from different assessment backgrounds: six raters from TELC, one from Cambridge and three from the ZfS at Paderborn University. Additionally, six raters are native speakers and four non-native speakers. The minimum criterion that qualifies raters for the expert rater group is that they had conducted at least 10 assessments according to the CEFR prior to the data collection phase. Plus, they needed to be affiliated to an assessment center at the time of the rating (see section 4.3.9). I used these criteria to ensure that the raters were sufficiently trained and still active in assessments. One problem is that all assessment centers use different kinds of descriptors and assessment grids that they calibrate themselves towards the CEFR levels. It cannot be determined to what extent the raters unconsciously applied those criteria to the ratings that they are familiar with. Also, it would have been beneficial to judge the tendency in rater personality towards harshness or leniency. Eckes (2005) found that these individual tendencies influence rating results quite strongly. Another issue is that professional raters usually assess speakers according to one specific CEFR level and specific skills. The goal of the assessment determines the use of tasks and the content of the tasks. The unfamiliar layout of tasks that were used in this study and that are geared towards eliciting a maximum number of morphosyntactic structures according to PT might compromise the rating results for this group.

The novice rater group shows the clearest tendency in this study when it comes to the effect of grammatical manipulations in the sample. This group was trained in the use of the Global Oral Assessment Grid (as suggested by the Manual for Aligning Examinations with the CEFR published by CoE in 2009) and has had a minimum amount of instruction on ratings. This group is familiar with the layout of the sound files because they were trained in rating these tasks (see chapter 4.3.8). The novice rater group tends to rate the edited samples on lower levels than the original samples. Table 28 shows that in 40 cases, the edited sample was rated lower than the original sample with a significant p-value.

Novice raters (n=21)		
Edi < Org	Edi = Org	Edi > Org
40	19	25
z= -3.669, p= .000		

Table 28: Original-Edit Rating in Novice Rater Group

One could assume that the low level of experience with ratings and learner language makes this rater group prone to focus on aspects that are intuitively easy to assess. I assume that these aspects comprise features for which right or wrong can easily be determined. Grammar, I suppose, is one of those aspects since it is quite easy to assess whether a grammatical structure is produced correctly or not. I believe that concepts like fluency, range, coherence and interaction – which are the subcategories in the Assessment Grid for Overall Oral Production – are more fluid concepts which cannot easily be quantified. In my view, it is harder to assess whether a learner is able to “express him/herself fluently and spontaneously, almost effortlessly”, as specified in the assessment grid for Fluency at C1 level (CoE 2009: 185), than to assess “Uses some simple structures correctly” as stated in the Grid for Accuracy (CoE 2009: 185). Raters who are quite new to the assessment of language learners might therefore cling more tightly to the factors that seem easier to assess than the others. I assume that this is the reason why the manipulation of grammatical features shows the

most influence in the novice rater group. Again, this is a post-factual explanation and needs to be investigated in future research.

In the following section, some general aspects, which might have had an influence on the ratings of edited and original speech samples, will be discussed.

4.4.3.2 Rater Cognition and Rater Experience

As stated in chapter 4.3.2 above, only a small body of research on rater cognition processes in oral assessments is available. This study employs a new methodology that aims at determining which performance features raters focus on. While the results for the amateur rater group and the expert rater group are rather inconclusive, the novice rater group shows significant results in terms of the effect of the manipulation of the grammatical variable in oral learner language data. In her study on rater focus in oral assessment, Brown (2000) shows that raters tend to focus on factors that are not specified in the assessment grid, such as fluency and pronunciation or communicative skills. She also found that raters put different criteria to use in order to specify what it means to fulfill a task demand. Some raters employed a narrower strategy in looking for specific linguistic features and some raters focused more on the context of learners' responses. All of these aspects might have played a role in the ratings of the original and edited speech samples in this study. A qualitative approach that is not based on group means but focuses on the individual rater to determine why such diverse results occur, may be beneficial. However, the focus of this study is to investigate the role of grammar in oral proficiency ratings in order to evaluate how fruitful it is to combine a theory on the development of morphosyntactic structures with the CEFR. In this regard, the findings for the novice rater group concur somewhat with Pollitt & Murray (1993). Pollitt & Murray (1993) assessed raters for the Cambridge Assessment Spoken English oral interview. These raters focused more on grammatical knowledge at lower levels and sociolinguistic competence and stylistic devices on higher proficiency levels. The focus on grammatical aspects can be confirmed in this study for the novice

rater group. In chapter 4.4.2.2, I showed that the novice raters tended to rate the A levels lower than those samples on higher levels. Therefore, the experience level of raters might also play a role in terms of which features raters attend to in oral proficiency ratings. Davies (2016) investigated the role of rater training and experience for consistency in rater scoring. He found that generally, training resulted in increased inter-rater correlation and agreement as well as improved agreement with established reference scores. However, experience that raters gained after training appeared to have little further effect on raters' scoring consistency (Davies 2016: 117). Similarly, Isaacs & Thompsons (2013) found no differences between an experienced rater group and a novice rater group in terms of scoring consistency for oral interviews. May (2006) used verbal protocols and immediate retrospection to assess rater orientations of an experienced and an unexperienced rater. She found that the experienced rater referred more to accuracy, whereas the inexperienced rater referred more to fluency aspects in their ratings. Isaacs & Thompsons (2013) found that the same was true for their data but comment that the inexperienced raters did not have the meta-language readily available to pinpoint the exact errors. These findings diverge from the findings in this study. However, the differences in findings might be due to the methodology employed. In the present study, the novice raters did not need to use meta-language in order to comment on their assessment. Rather, it might be argued that the novice rater focus is apparent due to the tendency to rate the edited speech samples on lower levels than the original samples.

From the comments that some of the raters gave, it can be seen that a focus on grammatical items was present with at least some of the data. The only comments that experienced rater aE02 (expert group) gave for file Ko05, for example, were about grammatical issues.

aE02	Good grammar/fluency/vocabulary/pronunciation. Some minor grammar slips (mainly present progressive) but otherwise using complex sentences and grammar structures.
------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------

Experienced rater bE01 however, was more distinct in her rating and weighed the different performance aspects against each other.

bE01	There were more grammatical mistakes than I would like to have heard at B2 'I think that he should to work;', 'Behind of him' Her fluency and interaction were very good. The exercises were perhaps not designed to test a B2 candidate, but I felt there was enough extension in her language (the Martian exercise particularly) to justify a B2 overall. But I would definitely only have marked Accuracy as B1 and possibly a B1, too, for Range (lack of complex sentence forms)
------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

She distinctly states that she would have rated Accuracy as a B1 level but decided to rate the overall performance as a B2, because the learner's overall performance was better than B1. From her comment, it seems that she abides to the rating grid and views grammar as only one aspect that does not outweigh the other performance factors. It can be seen in these comments that the expert raters within the group employed different foci in their rating procedure. While aE02 only states her general impression of the learner, bE01 distinctly refers to the descriptors in the CEFR. These differences in the approach taken to the rating would need to be assessed in more detail in future studies, because they might have had an influence on the inconclusive results in the experienced rater group.

In the following section, the effect of uneven profiles will be discussed.

4.4.3.3 The Role of Uneven Profiles

Samples that display varying degrees of proficiency in different language areas are referred to as uneven profiles. The CoE (2009: 43) itself presents a "flat profile" for rater training in the Manual for Relating Language Examinations, but Hulstijn (2011: 243) reports on a personal conversation with Green that uneven profiles are the rule and flat profiles an exception. Therefore, it is questionable as to why the CoE (2009) suggests training raters using flat profiles. In this regard, Hulstijn (2011: 243) asks the question:

[...] how linguistically imperfect (in terms of vocabulary, grammar, pronunciation/intonation, articulation speed) can performance on a C1 task be without failing as a communicative act, and to what extent can weaknesses in one component of linguistic competence be compensated with strengths in another component at a given CEFR level?

In this quote, Hulstijn makes explicit that uneven profiles and their effect on proficiency ratings are issues that need to be more closely addressed in language assessment research. I assume that the effect of an uneven profile is strongly visible in some of my language samples.

I would like to examine the samples at the higher PT stages more closely. Audiofiles K04, K06 and K05, K10 relate to PT stages 5 and 5+. The actors in samples Ko06/Ke06 and Ko10/Ke10 display a highly native-like accent approximating British and American varieties of English. Rater bE02, for example comments on sample Ko06 “The student has good pronunciation (exposure to native speakers?)” (brackets in original). The pronunciation would thus have probably been rated a lot higher than the other performance language features (accuracy, coherence, etc.); especially with the edited file. Therefore, I consider these samples instances of uneven profiles.

Based on *mode*, file Ko10/Ke10 is the only pair in which a tendency to rate the edited version lower than the original file can be observed across all three sub-groups.

Mode for Files Ko10 and Ke10			
Amateur Raters	Ko10	C1 (H)	
	Ke10	B2 (G)	= -1 level
Novice Raters	Ko10	C1	
	Ke10	B1	= - 2 levels
Expert Raters	Ko10	C1	
	Ke10	B2	= - 1 level

Table 29: Mode for Ko10 and Ke10

All rater groups seem to agree that the original file, Ko10, can be placed on CEFR level C1 (or letter H respectively). It is to be borne in mind though that only four amateur raters, six novice raters and three expert raters represent mode in this instance, so all results and their implications have to be viewed rather cautiously. The amateur and expert raters rated Ke10 one level below the original version. The novice raters assigned only the B1 level to the edited file, which is minus two levels as compared to the original sample. The learner in file Ko10 is located at

PT stage 5+. His language seems rather elaborate and he displays a number of discourse features and quite specific vocabulary, even some humoristic instances. This can be seen in the excerpts of the transcriptions below.

Ko10	she seems to be some kind of secretary (#) (um) well and at one pm she obviously quits (#) working/stops working and (er) does shopping afterwards (um) what kind of engine does your Mars spaceship have↑
	I would sell it [<i>the spaceship</i>] to someone else and make millions of dollars with it (#) and then I would sell you [<i>the Martian</i>] and make even more money (#) with you because you are an alien (#)
	Okay so I was wondering how you want to go from now↑ and (#) where to↑

Ko10 and Ke10 were re-recorded with the help of an actor. The actor shows a very strong, native-like accent, tending towards the American variety. Winke & Gass (2013) have shown that accent plays a decisive role in the rating of oral language data. In a qualitative study, they showed that especially non-native speakers are prone to be biased by foreign accents to such an extent that test reliability is influenced. Accent familiarity however, can counter these biases as shown by non-native speaker raters. More than half of the expert raters (06/10) in this study are native speakers of English. It seems to be the case that the strong mismatch between the elaborate use of vocabulary and discourse features, paired with the native-like accent has produced the strongest results in terms of rating the edited language sample at a lower level than the original sample. I assume that this is because the accent is native-like, but the grammar is inaccurate. It thus seems to be the case with this learner that the mismatch between the pronunciation and the inaccurate grammar is the cause for rating the edited sample at a lower level. This becomes especially evident when comparing Ko10 to Ko06. Ko06 is a learner on PT stage 5. Her vocabulary and discourse connectors seem to be quite advanced but a lot less elaborate than those used by Ko10.

Ko06	Okay (um) Mrs. Lee starts with (#) standing in his/(er) in her bedroom and (er) after that (er) at 8 (er) eight o'clock (er) she goes to breakfast and (um) sits there with (#) her family
	I have two (um) also a rabbit and a (#){Gans} she is running in front of the playground (um)

	(um) what color do you (er) do you (#) have your (um) house↑ and (er) the top↑
	and now you stay in berlin↑ so what is the reason↑

The actor with whom I re-recorded Ko06, however, shows a very strong, native-like British accent. Based on mode alone, there is no tendency to rate the edited file for this learner lower than the original file. All sub-groups rated original and edited file on the same level.

Mode for Files Ko06 and Ke06		
Amateur Raters	Ko10	B2 (G)
	Ke10	B2 (G)
Novice Raters	Ko10	B1
	Ke10	B1
Expert Raters	Ko10	B1
	Ke10	B1

Table 30: Mode for Ko06 and Ke06

The raters might have perceived less of a mismatch between the vocabulary, phonology, and the erroneous grammar of this learner in the edited file compared to Ko10. This lack of a mismatch might be responsible for not rating the edited sample lower than the original in this case.

It might be assumed that Ko10/Ke10 displays a strong uneven profile and the results show that all groups rated the edited version of this sample at a lower level than the original. The approach of using original-edited speech samples might therefore be valuable for investigating the effect of uneven profiles more closely in future research.

The next section will discuss whether it makes a difference in the rating results, if the original or the edited file was presented first to the raters.

4.4.3.4 The effect of the locus of presentation of files

This section investigates whether it makes a difference to the rating results if either the edited or the original files were presented before their corresponding

files. That is, if the edited file was presented before the original file or vice versa. If there were tendencies for the edited files that were presented before the original files to be rated differently; i.e. with a stronger/weaker agreement amongst the raters, then the locus of presentation of the files to the raters might be responsible for the variability in the results. This would pose a threat to the comparability of the rating results with regard to original and edited files.

To investigate this in more detail, it may be beneficial to investigate whether there is a difference in the ratings when the edited speech sample is presented before the original sample to the raters, or vice versa.

File	Amateur Raters	Novice Raters	Expert Raters
Ko03/ Ke03	↓		↓ ↑
Ko09/Ke09			
Ko01/Ke01	↑	↓	↑
Ko02/Ke02		↑	↑
Ko04/Ke04	↓	↑	↓
Ko06/Ke06			
Ko05/Ke05		↓	
Ko10/Ke10	↓	↓	↓

Table 31: Overview of Direction of Rating Tendencies for Original and Edited Files across Groups

The column most left depicts the original-edit sound file pair. The arrow pointing downwards indicates that the edited sample was rated on a lower level than the original sample. The vertical dash marks a tie in the rating results for both samples and the arrow pointing upwards shows that the edited version was rated at a higher level than the original. When a dash as well as an arrow appears, then there is an equal number of raters who appointed either the same level as well as a lower/higher level (see Table 31 for more details). Table 31 is based on the mode of ratings and not on range.

Raters had access to the edited files Ke03 and Ke09 before they were given the original files. Ke03 and Ke09 are depicted at the top of Table 31 presented above. Both files represent learners at PT stage 3. One can see that the results are somewhat inconclusive. For Ko09/Ke09 none of the rater groups

display a difference in levels assigned to the edited and original sound file; i.e. all are rated on the same level. For Ko03 and Ke03, the amateur rater group assigned the same as well a lower level for the edited sample based on mode. The novice rater group shows a tie for these samples and for the expert rater group, mode represents a higher as well as a lower CEFR level for the edited version Ke03. For all the other files presented below Ko03/Ke03, the original sample was presented before the edited version. Here, the same arbitrary results can be seen. With Ko06/Ke06, for example, all rater groups assign the same mode level to the original and edited sample. Samples Ko02/Ke02 and Ko04/Ke04 show that rater groups gave either a tie, a rating on lower or a rating on higher levels for these samples. For this reason, I assume that the locus of presentation of the edited files to the raters does not make a difference in rating results.

Next, it will be discussed whether the different types of manipulation of morphosyntax have an effect on the rating results.

4.4.3.5 The effect of the different types of manipulation of edited files

One might assume that the effect of the grammatical manipulation gives clearer results for those samples that were edited by the audio-engineer because it is only the morphological features that were erased. In the re-recordings with the actors, one could argue that some aspects, such as speed, tone or pitch might differ in some instances. If this were the case, one might assume more consistency to rate the edited sample on lower levels than the original samples across, as well as within, the different rater experience groups. As can be seen from Table 10 above, samples Ke01/Ke03 were edited by the audio engineer. The table below presents the range of levels that were assigned to the edited recordings. The column *No. of levels* depicts in how many levels the range for the edited versions differed as appointed by the raters. The column *PT Orig.* shows which PT stage of the original sample the edited version corresponds to. For the edited versions, no PT stages were elicited. The files edited by the audio-engineer are highlighted in grey and presented at the top of Table 32.

File	PT Orig.	Amateur Raters		Novice Raters		Expert Raters	
		Range	No. Levels	Range	No. Levels	Range	No. Levels
Ke03	3	E-F	1	A2-B1	1	A2	0
Ke01	4	F-H	2	A2-B1+	1,5	B1	0
Ke02	4	E-G	2	A1-B1	2	A2+-B1	0,5
Ke09	3	D-F	2	A1-A2+	1,5	A2-A2+	0,5
Ke04	5	F-I	3	A2+-B2	1,5	B1-B1+	0,5
Ke06	5	F-H	2	A2-B2	2	A2-B1+	1,5
Ke05	5+	F-H	2	A2-B2	2	A2-B2	2
Ke10	5+	E-I	4	A2-C1	3	B1-B2	1

Table 32: Comparison of edited versions generated by audio-engineer and through re-recordings

If there were a measurable effect of the type of manipulation of the audio-files, then I would assume that the range (no. of levels) for the files that were recorded with the help of the actors strongly differs from the ranges (no. of levels) that were assigned to those files manipulated by the audio-engineer.

The amateur rater group varied in about one to two levels in terms of the recordings edited by the audio-engineer. For the other samples, the amateur raters varied in up to four different levels (see file Ke10). However, one can see that the higher the PT stage, the more variability in the no. of levels is visible. This tendency is true for the other subgroups as well. Thus, it might be concluded that the higher the PT stage, the more variability in the rating results. Ke09 is a sample that was re-recorded with an actor. Its original correspondent is located at PT stage 3, just as the file Ke03, which was edited by the audio-engineer. The amateur rater group only varied in about two samples. Thus, the two different ways of editing the sound-files, namely by the audio-engineer and by re-recording doesn't seem to have an effect in this group.

The same tendency seems to be true for the novice and the expert rater group when comparing samples Ke03 and Ke09 (original version both assessed for PT stage 3). The differences in the levels assigned to both files are nearly the same. One can see from the table that except for File Ke10, the number of levels contained in range varies in about the same size for both types of manipulations

when the level of learner language is considered. I thus assume that there is no effect of the different type of manipulation.

The next chapter summarizes the results discussed in the previous sections.

4.4.3.6 Summary of the Effect of Manipulation in Morphosyntax

To summarize, the results of the effect of manipulating grammatical accuracy in edited speech samples vary to some extent. Raters rated the edited speech samples on lower levels, the same level or even higher CEFR levels. With significant differences in medians between original and edited speech samples, the tendency to rate edited speech samples on lower levels is most strongly visible within the novice rater group (40 x lower, 25 ties, 19 higher). In the expert rater group, a down-rating tendency was least observable. The expert group rated (nearly) an equal number of files on lower, the same as well as on higher levels (13 lower, 14 ties, 13 higher). However, this group is a) the smallest in size and b) the most heterogeneous group in terms of their background. These factors might cause the non-significant results. Interestingly, the amateur rater group who based their ratings on pure intuition shows a similar tendency as the expert rater group. The amateur raters seem to rate the edited speech sample on a higher level than the original sample (28 lower, 22 ties, 38 higher). An explanation regarding the differences between the groups can only be given tentatively and needs to be investigated in future research. However, it might be the case that novice raters tend to rely quite strongly on cues that they are very familiar with and that are rather easy to assess, i.e. grammatical accuracy. The expert raters seem not to rely on grammatical cues as much, but to also take the other features (range, fluency, interaction and coherence) into account. It might be the case that it is harder to employ a right or wrong approach to these features which results in more variability (in terms of rating files on lower and higher levels) in the levels assigned.

When considering all the groups together, the edited samples are rated on lower CEFR levels than the original samples with significant results (see section 4.4.1). Although the expert raters seem not to be affected by the grammatical variable as much as the other groups, I conclude that grammar does play a role in the ratings of Global Oral Proficiency. I argue that the inconclusive results for the expert raters is mainly due to the heterogeneity inherent in this group. For this reason, I hypothesize that there is a need to complement the CEFR with theoretical and empirical research on the acquisition of grammatical features. At the same time, these results show that PT, as a psycholinguistic theory of SLA with a focus on oral production, should be investigated as complementary to the CEFR. Therefore, the next chapter discusses the results of the relationship between CEFR levels and PT stages. It also proposes a combined PT-CEFR scale for *Grammatical Range*.

4.4.4 Relations between CEFR levels and PT stages

This section of the chapter deals with the relationship between CEFR levels and PT stages. The learners' PT stages were determined prior to the study with the help of the Rapid Profile computer interface. Two samples representing each PT stage were chosen. For PT stages 1 and 2, two additional learners were selected because for these stages no edited samples were generated. An even number of files at each PT stage should be distributed to the raters in order to establish a certain amount of consistency for the rating procedure itself (see chapter 4.3.7 on the distribution of files for more detail). I will show the results for the expert and the novice groups together and novice raters and expert raters individually, compare their data and depict which CEFR level generally relates to which PT stage. It is important to note here that the amateur raters' results will be left out as they did not specifically assign CEFR levels but used a random letter to quantify their rating of the learner language in the audio-files. Also, the amateur raters did not receive any training in the CEFR and rating procedures. The part of the study in which they took part, relates to a different hypothesis and is discussed in chapters 4.4.1 - 4.4.3.5. Additionally, only the original files, not the edited files,

are used to calculate correlations because only the original files represent actual learner language.

Table 33 below shows which PT stage relates to which CEFR level in the data set. The left-hand column shows the PT stages for the original samples determined prior to the ratings. The second column represents the CEFR level that was most frequently assigned by both groups (mode), novice and expert raters. The third column shows the mode of the assigned CEFR level in terms of percentages. Similarly, the CEFR levels and percentages for each sub-group are given in the remaining columns. The CEFR levels in which the novice and the expert rater group differs, are highlighted in bold. I will first describe the results for both groups together.

In 60% of the cases, novice and expert raters combined assigned audio files at PT stage 1 the A1 level. A1 was also most frequently assigned to audio files at PT stage 2 at 73,9%. At 61,8%, audio files at PT stage 3 were given an A2 level. Audio files at PT stage 4 and 5 were both rated on the B1 level (35,3% for PT stage 4 and 44,1% for PT stage 5). Interestingly, the B2 level was not assigned to the audio-files often enough to be included in this overview. Audio files at PT stage 5+ were most often rated a C1 level (34,3%).

PT stage	CEFR both (n=31)	%	CEFR Novice (n=21)	%	CEFR Expert (n=10)	%
1	A1	60	A1	64	A1	57,1
2	A1	73,9	A1	77,6	A1	66,7
3	A2	61,8	A2	70,8	A2	38,5
4	B1	35,3	B1	41,7	A2	46,2
5	B1	44,1	B1	50	B1+	46,2
5+	C1	34,3	C1	36	C1	42,9

Table 33: Results Rating - General Overview of Relations between PT stages and CEFR levels

The novice rater group shows the exact same picture for correlations between PT stages and CEFR levels. This is reasonable, because the novice rater group

consists of twice the number of raters than the expert rater group. The percentages for each of the levels are higher in the novice rater group compared to both groups. With the novice raters, PT stage 1 seems to relate to CEFR A1 at 64%. PT stage 2 also relates to CEFR level A1 at 77,6%. At 70,8%, raters rated audio-files at PT stage 3 on CEFR level A2. B1 was assigned to audio-files at both PT stages 4 and 5. At 36%, C1 seems to be related to PT stage 5+. The group of expert raters deviates from this pattern in that they assigned PT stage 4 the A2 level and PT stage 5 the B1+ level at 46,2 % in both cases. The percentages at the lower levels, i.e. for the audio-files displaying earlier learners of English, are generally higher than those for learners at later levels. This matches the discussion about the effect of different rater types in section 4.4.5. The expert raters seem to converge less with their ratings at PT stage 3 than at PT stages 4, 5 and 5+.

A Spearman's Rank Order Correlation Test was run to determine correlations between the original samples at each PT stage (as determined prior to the ratings) and the ratings provided by both novice and expert raters. Spearman's Rho shows a strong positive correlation at $\text{Cor}(\text{PT}, \text{CEFR}) = .864$ with a p-value of $p = .000$. Correlations become significant at $p = .01$ for this test. Since Preston (2009) suggests that for some calculations, Kendall's Tau-b is more appropriate to capture the strength of correlations, a Kendall's Tau-b correlation coefficient was calculated for assigned CEFR levels and PT stages. Preston (2009) argues that Kendall's Tau-b can handle tied data and usually has smaller values than Spearman's rho. It is also based on concordant and discordant pairs, making it less sensitive to errors than Spearman's Rho (see e.g. Preston 2009). Thus, the p-values of Kendall's Tau-b are considered to be more accurate with smaller sample sizes. Kendall's Tau-b yields a strong positive correlation at $.823$ with $p = .000$. This correlation is not strikingly less significant than the Spearman's Rank Order Correlation Coefficient. Both tests show strong positive correlations. In the next chapter, I will discuss the results on the relationship between PT stages and CEFR levels.

4.4.4.1 Discussion of Results on Relations between CEFR and PT

Table 34 gives an overview of the different results of the studies on PT and the CEFR that were presented in chapter 3.2. The results of the present study are highlighted in grey in the last column. The results for the present study are based on the ratings of both novice as well as expert rater groups. All the instances in which the present study converges with the results of a previous study are printed in bold.

	Lenzing & Plesser (2010)	Michalska (2010)	Hagenfeld (2017)	Present study
PT	CEFR level	CEFR level	CEFR level	CEFR level
1	Below A1	/	/	A1
2	Below A1; A1	/	/	A1
3	A1	A1; A2 ; B1; B2	/	A2
4	A1; B1	A2; B1 ; B2	/	B1
5	B1 , B2, C1	B1 ; B2; C1	B1 , B2, C1	B1
5+/6	C1	B2; C1	/	C1

Table 34: Overview Previous Studies on the Relationship between PT and the CEFR

It is only the assessment of level A1 for PT stage 1 that does not confirm the results of any previous studies, but I assume that this is due to the fact that previously only Lenzing & Plesser (2010) have looked at learners at these early PT stages. The results for PT stage 2, i.e. CEFR level A1, converge with Lenzing & Plesser (2010). The result for PT stage 3, i.e. CEFR level A2 converges with Michalska (2010). The CEFR level B1 for PT stage 4 converges with both Michalska (2010) and Lenzing & Plesser (2010), and the same is true for the results for PT stage 5. This might not be surprising because previous studies found a variety of CEFR levels corresponding to these PT stages. For example, a) Michalska (2010) investigated the inter-rater reliability differences between Rapid Profile assessments and CEFR ratings and therefore found a variety of levels for the different PT stages and b) Lenzing & Plesser (2010), as well as Hagenfeld (2017), also found three CEFR levels to correspond to PT stage 5. All studies (except Hagenfeld who did not investigate PT stage 6) seem to concur that PT stage 5+/6

seem to relate to CEFR level C1. One can see that the results of the previous studies and the present study agree to some extent, despite the different methodologies. However, it is precisely because of the different methodologies and hypotheses in the studies that the relationship between them has to be treated with caution: all studies used different sets of rating grids that are more or less calibrated to the CEFR levels. I consider the present study especially valuable in this regard, because the assessment grid used by the raters was produced and calibrated by the CoE itself. Therefore, I assume a maximum of level of fit between the descriptors and the assessment grid.

The comparability of the present results to the results by the studies that investigate interfaces between SLA and the CEFR other than through a PT lens is also problematic. Firstly, this is the case because most of the studies focus on language-specific rather than language-universal aspects. Also, most studies focus on written rather than oral production (see Prodeau et al 2012; Gyllstad et al. 2014; Thewissen 2013) and thirdly, the measurements used to investigate the interfaces between SLA and the CEFR also vastly differ (see Díez-Bedmar 2015; Williams 2007, Wisniewski 2017a). The problem of comparability can be depicted using the following example: Prodeau et al. (2012) use a corpus-based design in investigating written French to investigate language-specific criterial features that distinguish level B1 from level B2. They were not able to find any distinguishing features and reason that this is due to the gradual nature of grammatical development (Prodeau et al. 2012: 63). PT, as the theoretical framework used in the present study, also assumes a gradual development of grammar, but is based on universal processing procedures that help to specify language-specific developmental markers in oral production. However, my study did not find any relationship between CEFR level B2 and PT stages, because the B2 was not rated often enough to have an influence on the correlation.

In this study, the raters used data that had been gathered with the aim of eliciting the PT stage of a learner. Therefore, the learners performed in communicative tasks as described in chapter 4.3.4 above. It cannot be determined to what extent the rather unusual data had had repercussions on the rating result because raters might not be used to the data. However, before the

study started, I asked different expert raters if it were possible to rate the data with the Global Oral Assessment Grid. As stated in chapter 4.3.9, expert raters commented on the unfamiliarity of the tasks, but also stated that they felt that sufficient ratable structures were present in the data. In fact, none of the raters reported that they were not able to place a file on the Global Oral Assessment Grid.

Wisniewski (2017a: 4) argues that it is “[...] preferable to have every text explicitly rerated post hoc to establish a more direct link to CEFR.” Although, I did not use texts, but oral language samples, this procedure was used in this study. One could argue that, in order to establish a more direct link between the CEFR and PT, an assessment grid for grammatical accuracy would need to be produced and the study should be repeated with this scale. However, since there was no assessment grid for grammatical accuracy available to me that had been provided by the CoE, I decided to use the scale for Global Oral Production instead, because grammar is only one aspect in this scale and PT primarily focuses on oral production.

It is important to bear in mind that rating procedures in themselves face some issues. While rating procedures aim at systematizing human judgements, they are not free from subjectivity. Wisniewski (2017a: 7) maintains that one cannot assume that ratings actually mirror the rating scales because it has not been fully established as to what raters actually do while rating (see also Connor-Linten 1995: 763). The results presented in sections 4.4.1-4.4.3.6 on the role of grammatical accuracy add to this discussion. Thus, the claim that there is a relationship between PT and the CEFR which relies mostly on the ratings carried out by different rater groups, has to be treated with caution. Chapter 3 in which I laid out some theoretical interfaces might ease out this issue to some extent.

The next chapter presents a scale on the basis of the study results that combines PT and the CEFR.

4.4.4.2 A Combined Scale for Grammatical Range

On the basis of the study results presented above, the following combined scale for *Grammatical Range* might be proposed based on the overall relationship between PT and the CEFR. I will first present the universal scale for *Grammatical Range*, then compare this scale to the scale for General Linguistic Range and afterwards present the language-specific scale for *Grammatical Range for English*. It is important to bear in mind that the descriptors used in this scale are informed by the original CEFR descriptors for grammatical accuracy and corroborated with descriptors that relate to the universal processing procedures as spelled out by PT. The formulation/wording of the PT-related descriptors is not empirically tested or scaled as the original descriptors are (see North 1996; North & Schneider 1998). If a combined scale for *Grammatical Range* were to be put into practice, future research would need to determine the appropriateness of the item formulation. It should also be noted that this study uses only two language samples at each PT stage to determine the relationship between PT and the CEFR. Future research needs to use a bigger data set and test the hypotheses put forward in this thesis. Therefore, the scale for *Grammatical Range* needs to be regarded as tentative.

CEFR	Descriptors for Grammatical Accuracy	Combined Descriptors for Grammatical Range	PT Phenomena	PT
C2	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	/	/
C1	Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot.	Consistently maintains a high degree of grammatical control ; errors are rare and difficult to spot. Productive language use broadens to structures that require subordinate clause procedure operations.	Subclause-procedure	5+
B2	Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	/	/
	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstandings.	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstandings.	/	/
B1	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express.	Communicates in familiar contexts; has generally good control over the target language. Errors occur, but it is clear what he/she is trying to express. Productive language use comprises the unification of grammatical information on an inter-phrasal level.	S-procedure	5
	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	Uses a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	VP- procedure	4
A2	Uses some simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget (sic!) to mark agreement; nevertheless, it is usually clear what he/she is trying to say.	Uses simple structures, but still systematically makes basic mistakes – for example tends to mix up tenses; nevertheless, it is usually clear what he/she is trying to say. Productive language use comprises grammatical information unified on a phrasal level.	Phrasal Procedure	3
A1	Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.	Produces a few chunks and formulaic sequences in a learnt repertoire. Assigns the grammatical category. The sequences might show a level of accuracy, but this is due to their unanalyzed nature.	Category procedure	2
			Word / lemma access	1

Table 35: Empirically Motivated Proposed Scale for Grammatical Range in Comparison to Grammatical Accuracy

In spelling out the descriptors for *Grammatical Range*, I remained as close to the original voice of the CEFR scale as possible. However, I deleted all the instances that point towards grammatical accuracy as I assume that grammatical accuracy does not capture language development appropriately. Since some PT stages relate to two CEFR levels, I put both of them together in one line (see e.g. level B1). Here, an intra-level hierarchy might be assumed but this needs to be determined in future research. I used the universal processing procedure terms to specify the locus of grammatical information exchange possible at each level. Since a) learners are able to employ creative strategies to solve developmental problems, b) the use of unanalyzed material cannot be ruled out at each level, and c) the emergence criterion needs to be used to determine whether a grammatical structure has been acquired or not, all references to accuracy were deleted. Therefore, I used the phrase “productive language use might comprise...” before specifying the processing procedure. This phrase is supposed to indicate that learners can potentially employ different strategies to circumnavigate developmental problems. Another issue here is that, in principle, a test for each of the processing procedures would need to be employed.

A1 descriptors seem to relate to PT stages 1 and 2 and encompass descriptors for unanalyzed chunks and a few formulaic patterns. To indicate that many learners display some misleading level of accuracy in this early phase of development, the descriptor ‘The sequences might show a level of accuracy, but this is due to its unanalyzed nature’ is added. In terms of PT, only lemma access is possible at this time in the development. Therefore, the combined descriptors for *Grammatical Range* at level A1 include the phrase the ‘few simple’ grammatical structures from the ‘grammatical accuracy’ descriptors. This is to show that little production might be assumed in this early phase of language use. Further, this A1 descriptor is combined with Lenzing’s (2013: 163f.) typology of formulae (see chapter 2.2.2 for more information). Following Lenzing (2013: 163f.), the descriptors use the umbrella term ‘formulaic sequences’ that cover those expressions learned by heart, because they might occur as fixed expressions in textbooks. Formulaic patterns, according to Lenzing, are unanalyzed chunks with an open slot which learners fill. As opposed to

holistically-stored expressions, the filling of these slots requires at least some sort of assembling the lexical items into strings, although they are not analyzed at this stage yet. At this level, no feature unification as envisaged by PT would be assumed.

CEFR level A2 would relate to PT stage 3 as determined in this study and descriptors might comprise the unification of grammatical features on a phrasal level. In terms of grammatical structures, Prodeau et al. (2012: 53) argue that the A2 level is defined by “the absence of finiteness and utterances are organized in terms of topic and focus components in that order. Beyond the basic variety, the learner variety includes grammatical elements” from the B1 level onwards. PT’s predictions would be somewhat congruent with Prodeau et al. in stating that at the A2 level, the learner would mostly still rely on category procedure operations. It is only from the phrasal procedure stage that might be located at level B1, that the production of phrases with information exchange within the phrase would be assumed.

The B1 descriptors are subdivided into two parts in the original version for Grammatical Accuracy. I kept this division but deleted the words “reasonably accurately” so *Grammatical Range* encompasses “Uses a repertoire of frequently used ‘routines’ and patterns associated with more predictable situations”. Similarly, I deleted the phrase “though with noticeable mother tongue influence” in the higher B1 descriptors because a mother-tongue influence might have been noticeable at earlier stages as well and, in my view, there is no logical reason as to why this descriptor occurs at level B1 for the first time. The phrase “Productive language use comprises the unification of grammatical information on an interphrasal level” is used so as to bridge the gap to the S-procedure and the VP-procedure. It remains subject to future research to determine if the intra-level distinction between higher and lower B1 descriptors relates to the VP-procedure and the S-procedure. From a PT viewpoint however, stage 5 is quite advanced and one would assume that learners who have reached this stage might be able to communicate in more situations than only “predictable situations” as defined in *Grammatical Range*. As of yet, PT is not concerned with determining communicative acts and situations. That is why the assumptions about linguistic

production in specific communicative situations cannot be proven. However, the lower B1 level that states “Uses a repertoire of frequently used ‘routines’ and patterns associated with more predictable situations” would generally be associated with lower PT stages and not with Wh-Copula-S(x) or Copula-S(x) structures. These structures need inter-phrasal information exchange and are thus quite advanced from a processing point of view. I therefore suggest treating this relationship very cautiously.

For CEFR level B2, no relationship to a PT stage was found. This is why the original descriptors are simply copied. B2 descriptors in the original version do not specifically state grammatical accuracy but rather grammatical control and were thus not edited in the scale for *Grammatical Range*.

CEFR level C1 seems to relate to PT stage 5+. Therefore, the descriptors are supplemented with the phrase “productive use broadens to structures that require subordinate clause procedure operations”.

C2 does not occur in the present study which is why the descriptors are copied for this level as well.

4.4.4.3 Comparing Scales for Grammatical Range and General Linguistic Range

The authors of the CEFR advise the reader to read the scale for Grammatical Accuracy in relation to the scale for General Linguistic Range. It might therefore be useful to revisit the scale for Linguistic Range here in order to check for its compatibility with the proposed scale for *Grammatical Range*. Table 36 below places both descriptors next to each other for comparison:

CEFR	Descriptors for General Linguistic Range	Combined Descriptors for Grammatical Range	PT
C2	Can exploit a comprehensive and reliable mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity...No signs of having to restrict what he/she wants to say.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	/
C1	Can select an appropriate formulation from a broad range of language to express him/herself clearly, without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical control ; errors are rare and difficult to spot. Productive language use broadens to structures that require subordinate clause procedure operations.	5+
B2	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	/
	Has a sufficient range of language to be able to give clear descriptions, express viewpoints and develop arguments without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstandings.	/
B1	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates in familiar contexts; has generally good control over the target language. Errors occur, but it is clear what he/she is trying to express. Productive language use comprises the unification of grammatical information on an inter-phrasal level.	5
	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies, and interests, work, travel, and current events, but lexical limitations cause repetition and even difficulty with formulation at times.	Uses a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	4
A2	Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words.	Uses simple structures, but still systematically makes basic mistakes – for example tends to mix up tenses; nevertheless, it is usually clear what he/she is trying to say. Productive language use comprises grammatical information unified on a phrasal level.	3
	Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information. Can use basic sentence patterns and communicate with memorized phrases, groups of a few words and formulae about themselves and other people, what they do, places, possessions etc. Has a limited repertoire of short memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations.		
A1	Has a very basic range of simple expressions about personal details and needs of a concrete type.	Produces a few chunks and formulaic sequences in a learnt repertoire. Assigns the grammatical category. The sequences might show a level of accuracy, but this is due to their unanalyzed nature.	2 1

Table 36: Comparison Descriptors for General Linguistic Range and Proposed Descriptors for Grammatical Range

For descriptors at A1 level, there is no noticeable mismatch between General Linguistic Range and *Grammatical Range*. The “simple expressions” referred to in General Linguistic Range might embody the unanalyzed chunks in *Grammatical Range*. Level A2 is subdivided into lower and higher descriptors in Linguistic Range on the left-hand side but the original descriptors for Grammatical Accuracy are not distinguished there. Consequently, *Grammatical Range* does not either. The lower descriptors are quite compatible with *Grammatical Range* because the “[...] brief everyday expressions [...], basic sentence patterns [...] memorized phrases” may relate to the simple structures and phrasal sentence procedure operations in *Grammatical Range*. The higher A2 descriptors give no qualitative descriptors that refer to accuracy, but rather describe situations in which the language might be used under the heading of “repertoire of basic language”. One could argue that structures based on phrasal procedure operations can be subsumed under basic language so there does not seem to be an incongruity between those descriptors either. The lower B1 descriptors for *Grammatical Range* were not altered in comparison to Grammatical Accuracy so they should be congruent with General Linguistic Range. In fact, there seems to be a large overlap between the descriptors for both scales (General Linguistic Range and Grammatical Range) although General Linguistic Range seems to put some emphasis rather on lexical/vocabulary aspects at this level. Adding the descriptor “Productive language use comprises the unification of grammatical information on an inter-phrasal level” at the higher CEFR level B1 in the scale for *Grammatical Range* is also congruent with General Range descriptors because these refer only to “sufficient range of language” and “reasonable precision”. Thus, I assume that they are open enough to encompass the VP- and the S-procedure. The B2 descriptors are disregarded at this point because no relationship between CEFR level B2 and PT stages was found in this study. The “broad range of language” that is referred to in the descriptors for General Linguistic Range at C1 level can be argued to comprise subordinate clause procedure operations as specified in the descriptors for *Grammatical Range*. The C2 level is not covered in this study, so these descriptors will be ignored at this point.

From Table 36 it can be seen that there is no discrepancy between the combined descriptors for *Grammatical Range* and General Linguistic Range. I thus assume a good compatibility of both scales so that the new scale for *Grammatical Range* does not violate the assumptions in the CEFR.

4.4.4.4 Grammatical Range for English

A combined scale for *Grammatical Range* has advantages for the assessment of the specific target languages. Based on the universal processing procedures, researchers within the PT framework have spelled out specific grammatical features for various languages. Raters might find it handy to use those language-specific grammatical features to determine the level of *Grammatical Range*. A combined scale for English might consequently look as follows:

CEFR Level	Combined Descriptors for Grammatical Range for English	PT Phenomena and Examples	PT Stage
C2	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	/	/
C1	Consistently maintains a high degree of grammatical control ; errors are rare and difficult to spot. Productive language use broadens to structures that require subordinate clause procedure operations.	<u>Cancel inversion</u> <i>I wonder what he wants</i>	5+
B2	Good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	/	/
	Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstandings.	/	/
B1	Communicates in familiar contexts; has generally good control over the target language. Errors occur, but it is clear what he/she is trying to express. Productive language use comprises the unification of grammatical information on an inter-phrasal level.	<u>Inter-phrasal morph.</u> SV-agreement (<i>The mouse plays Volleyball</i>) <u>Neg/Aux-2nd-?</u> <i>Why doesn't he go home?</i> <u>Aux 2nd</u> <i>What do you collect?</i>	5
	Uses a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	<u>Wh-Copula-S(x)</u> <i>What is your number?</i> <u>Copula S(x)</u> <i>Are there boots?</i>	4
A2	Uses simple structures, but still systematically makes basic mistakes – for example tends to mix up tenses; nevertheless, it is usually clear what he/she is trying to say. Productive language use comprises grammatical information unified on a phrasal level.	<u>Phrasal morphemes</u> Det+N (<i>two ears</i>) <u>Adverb-first</u> <i>Today he stay here.</i> <u>Wh-SV(O)-?</u> <i>What you like?</i> <u>Do-SV(O)-?</u> <i>Do you have a sun?</i>	3
A1	Produces a few chunks and formulaic sequences in a learnt repertoire. Assigns the grammatical category. The sequences might show a level of accuracy, but this is due to their unanalyzed nature.	<u>Lexical morphemes</u> Plural -s (<i>pets</i>) Past -ed (<i>played</i>) <u>Canonical Word Order</u> <i>The mouse play Volleyball</i>	2
		Invariant forms/formulae	1

Table 37: Overview of potentially combined descriptors to propose a scale for English, based on Grammatical Range (examples taken from Lenzing 2013: 144; based on Pienemann 2005: 24)

Again, the same limitations as for the scale for *Grammatical Range* apply to the scale for English. The wording of the descriptor items was not empirically tested (as done by North 1996; North & Schneider 1998). For assessment purposes, the scale needs to be transformed into an assessment grid and calibrated against the

other CEFR descriptors. Considering, however, that the results in chapters 4.4.1-4.4.3.6 found that raters generally seem to unconsciously attend to grammar, it might be worthwhile to make them to listen for specific grammatical features as determined by PT so that a more reliable placement on CEFR levels can be achieved. What needs to be kept in mind though, is the use of the emergence criterion in which raters would need to be trained sufficiently.

To summarize, I assume that the proposed combined scale for *Grammatical Range*, based on PT principles and the scale for Grammatical Accuracy, taken from the CEFR, employs a more learner-centered, theoretically sounder and empirically grounded view to Grammatical Accuracy in comparison to the original scale for Grammatical Accuracy. Furthermore, I hypothesize that using the combined scale for *Grammatical Range* as a basis for spelling out language assessment grids will lead to greater validity in assessments because PT specifies grammatical features at each stage of development for specific target languages that raters can assess. This is, provided that they are familiar with emergence criterion, especially since this study found that raters tend to subconsciously attend to grammatical features in oral language production.

In the following chapter, the variables of rater experience and the use of an assessment grid in relation to the variability of the rating results in this study will be discussed. This chapter lays out implications for general language assessment, rater training and experience.

4.4.5 Rater Experience and Variability of Rating Results

Another aim of this study is to investigate to what extent rater experience has an influence on the reliability of rating results. Due to the study design, I was also able to test how the variability of rating results differ if the raters use an assessment grid or administer their rating based on intuition.

To recapitulate, the amateur rater group did not use an assessment grid, and only assigned a random letter to the sound files distributed to them. The novice rater group received training sessions on the CEFR and rating procedures as modeled by the *Manual of Relating Language Examinations with the CEFR*

(CoE 2009). The expert rater group included 10 raters with varying backgrounds, all of which were affiliated to assessment centers at the point of data collection. The results for rater experience and the use of an assessment grid can be seen in Table 38 below.

Kruskal-Wallis H test	X²=32.933, p= .000		
Mann-Whitney U test	Amateur/ Expert	Amateur /Novice	Novice /Expert
p-values	p=.000	p=.000	p=.926

Table 38: Variability due to Rater Experience and Assessment Grid Use

A Kruskal-Wallis H Test was run to determine differences in the CEFR scores assigned to the speech samples by the three experience levels of raters; amateur raters, novice raters and expert raters. Distributions of the CEFR scores were similar for all groups, as assessed by visual inspection of a boxplot diagram. The Kruskal Wallis test yielded a result of $X^2=32.933$, $df\ 2$, $p=.000$. A post hoc Mann-Whitney U test was run to determine where exactly the differences between the three groups can be found.

At a value of $p=.000$, the difference between assigned CEFR levels by amateur and expert raters is statistically significant ($z=-4.625$, 2-tailed). The same results can be found for amateur and novice raters ($z=-5.207$, $p=.000$). However, there was no statistically significant difference between novice and expert raters ($z=-.093$; $p=.926$) in the assigned CEFR levels. This indicates that the amateur group, who did not use an assessment grid, performed differently from those groups who were given the Global Oral Assessment grid (novice and expert raters). Thus, one might assume that the use of arbitrary letters instead of the assessment grid is the factor which yields the difference in the scores assigned to the audio-files. Another indication is that novice raters, who had received a minimum amount of training prior to the rating, do not perform much differently than raters with more experience. This confirms the results presented by Isaacs & Thompsons (2013) and May (2006).

A follow-up test, the Kendall-W Test was run to investigate the agreement within the three rater groups. A Kendall-W Test that calculates the coefficient of concordance, was run to determine the level of concordance within three

groups; amateur raters, who did not use an assessment grid, novice raters who received some instruction on the CEFR and rating procedures, and expert raters, who were affiliated to language testing agencies at the time of data collection. Please note that I had divided each group into subgroups A and B to reduce the workload for each rater during the data conduction phase. Both subgroups thus rated a different set of samples. Subgroups A and B should still be comparable because each group received the same number of samples with a comparable mean length, in the same manner of distribution, and at the same PT stages. Kendall's coefficient of concordance was calculated for both subgroups within the three levels of experience. The percentages that can account for the variability of all cases based on Kendall's W are given in the following table. First, each subgroup is presented on its own and then the amalgamated percentages for the whole group are presented.

Amateur Raters (n=22)		Novice Raters (n=21)		Expert raters (n=10)	
Subgroup	Subgroup	Subgroup	Subgroup	Subgroup	Subgroup
A (n=14)	B (n=8)	A (n=10)	B (n=11)	A (n=5)	B (n=5)
W=.,00 p=.825	W=.,00 p=.915	W=.,00 p=.875	W=.,00 p=.911	W=.,00 p=.874	W=.,00, p=.923
82,5%	91,5%	87,5%	91,1%	87,4%	92,3%
87%		89,3%		89,8%	

Table 39: Results Agreement within Sub-groups

Kendall's-W for subgroup A of amateur raters yielded a significant result and the raters agreed in their assessments of speech samples, $W=.,00$ $p=.825$. This indicates that the agreement between the 14 raters can explain 82,5% of all possible variability. Agreement in this subgroup can be assumed to be good. For subgroup B of amateur raters, the test yielded a significant result and the raters agreed in their assessments of speech samples, $W=.,00$ $p=.915$. This indicates that the agreement between the eight raters can explain 91,5% of all possible variability. Agreement in this subgroup can be considered very good. The test for subgroup A of novice raters yielded a significant result. The raters agreed in their assessments of speech samples, $W=.,00$ $p=.875$. This indicates that the

agreement between the 10 raters can explain 87,5% of all possible variability. Agreement in this subgroup is thus considered to be good. For subgroup B of novice raters, the test yielded a significant result. Novice raters (B) agreed in their assessments of speech samples, $W=.,00$ $p=.911$. This indicates that the agreement between the 11 raters can explain 91,1% of all possible variability. Agreement in this subgroup can thus be regarded as very good. For the subgroup A of expert raters, Kendall's-W yielded a significant result. Expert raters (A) agreed in their assessments of speech samples, $W=.,00$ $p=.874$. This indicates that the agreement between the five raters can explain 87,4% of all possible variability. Agreement in this subgroup might be considered as good. For subgroup B of expert raters, the test yielded a significant result. These raters agreed in their assessment of speech samples, $W=.,00$, $p=.923$. This indicates that the agreement between the five raters can explain 92,3% of all possible variability. Agreement in this subgroup can therefore be considered very good.

Interestingly, the agreement in the amateur rater group is nearly as high as in the other groups, although they did not use an assessment grid and although the Mann-Whitney-U test yielded a significant difference in scores assigned to the audio files across groups (see Table 41). This might be because the amateur raters did not use the CEFR grid for their rating but arbitrary letters. However, in this data set, it seems as if the raters who use intuition for their ratings produce results that are nearly as reliable (in that they agree as much) as the other subgroups who used an assessment grid. This finding will be discussed in section 4.4.5.1 in more detail. The results also indicate that inter-rater reliabilities in subgroups B are consistently higher than those in subgroups A, although it is not possible to find the reason for this based on the data set. A question that arises is whether the samples in subgroups B were somehow easier to assess for the raters or whether the different number of raters within the subgroup played a role. As indicated above, in the sampling phase, close attention was paid to control as many variables as possible so that the samples distributed to the raters did not differ in too many aspects. A more qualitative analysis of this finding will be given below.

In the following part, I will display the dispersion of the assigned CEFR levels within each rater level group and show how much the groups varied in assigning the CEFR levels to recordings. The amateur rater group was able to assign letters from D to I for their ratings, which results in the opportunity to assign six letters in total. Novice and expert raters were able to allocate levels from below A1 to C2 including the plus levels up to level C1+. This results in the opportunity to assign 12 levels (below A1; A1; A1+; A2; A2+; etc.) in total. Twenty-two recordings were rated in total, 14 of which are original recordings and eight of which are recordings edited for grammatical features. The column to the most left indicates the three different experience levels as well as the group size; n. The column labeled *number* indicates in how many instances of the assigned CEFR levels the raters differed. The *Audio-file(s)* column shows which audio file the different CEFR levels were assigned to. For example, amateur rater group appointed file Ko14 the very same letter, the novice rater group assigned four different CEFR levels to the files Ke02, Ke01, Ko10, the expert rater group assigned four different CEFR levels to the file Ke05. The files Ko07 and Ko08 are called Ko07A, Ko07L and Ko08K, Ko08J because these are files in which the interviews were done in pairs. The raters, however, rated each of the learners individually. Thus, the letters A, L, J, K indicate which file it is that the different numbers of levels were assigned to.

Groups	Number	Audio-file(s)
Amateur Raters N=23	1 level	Ko14
	2 levels	Ko12, Ko07L, Ko07A, Ke03, Ko07L, Ko07A, Ko13,
	3 levels	Ko04, Ke05, Ko08J, Ke09, Ke01, Ko11, Ko03, Ke06, Ko10
	4 levels	Ke02, Ke04, Ko08K, Ko09, Ko01, Ko06, Ke10
	5 levels	Ko02
Novice Raters N=22	1 level	/
	2 levels	Ko08K, Ko08J, Ko09, Ko11, Ko12, Ko03, Ko07A, Ko07L
	3 levels	Ko02, Ko04, Ko05, Ke09, Ko01, Ke03, Ko06
	4 levels	Ke02, Ke01, Ko10
	5 levels	Ke04, Ke05, Ke06
	6 levels	Ke10
Expert Raters N=10	1 level	Ke01, Ke03, Ko13
	2 levels	Ke02, Ke09, Ko11, Ko12, Ke06, Ko07A, Ko07L
	3 levels	Ko02, Ko04, Ke04, Ko05, Ko08K, Ko08J, Ko09, Ko01, Ko03, Ko06, Ko10, Ke10, Ko14
	4 levels	Ke05

Table 40: Amount of assigned CEFR levels to audio-files across groups

Table 40 shows that the amateur raters place a high number of files on two, three and four different levels. The novice raters place the highest number of files on two and three different levels. The expert raters also place the highest number of samples on two as well as three different CEFR levels. Generally, it can be seen that all the rater groups seem to vary mostly in about two (novice raters) or three (amateur and expert raters) levels. The uppermost number of variances can be seen in the novice rater group who rated language sample Ke10 at six different CEFR levels. The second highest number can be found in the amateur rater group with five different letters ascribed to audio-file Ko02. It needs to be kept in mind that the amateur raters could only assign six different letters, whereas the other two groups could give plus levels as well, resulting in the opportunity to allot 12 different levels in total. Thus, the amateur rater group used nearly the full range

of their scale to ascribe to audio-file Ko02. In the expert rater group, the highest number of assigned CEFR levels amounts to four different levels for file Ke05. More information on this trend will be given in section 4.4.5.1 below. This analysis underlines the statistical analysis that the rater groups, in themselves, show high agreement.

4.4.5.1 Discussion of Results on Rater Experience and Variability of Rating Results

The following chapter discusses the results on rater experience and variability. It starts out with discussing reliabilities in connection with rater experience and concludes by commenting on different rating techniques employed by the raters.

4.5.5.2 Rater Experience and Agreement

When comparing the rater groups with each other, and as investigated by the use of the Mann-Whitney-U test (see section 4.4.5), the amateur group, who did not use an assessment grid, performed differently from those groups who were given the Global Oral Assessment grid (novice and expert raters) for their assessment. This difference might be attributed to the use of intuitive knowledge for the rating procedure. Interestingly, the agreement within the amateur rater group is nearly as high as the agreement within the other groups, even though they did not use an assessment grid (see section 4.4.5). These results converge with Davies (2016) and Isaacs & Thomson (2013). Both studies found that rater experience is not necessarily a factor that leads to a more consistent application of rating criteria or stronger inter-rater reliabilities. Similarly, Deygers et al. (2018) investigated 82 CEFR-based oral ratings on the B2 level and conclude that “[t]he results show that using the same language proficiency scales as the basis for rating scale criteria may lead to superficial correspondences or a perceived equivalence but does not necessarily lead to greater comparability of shared criteria.” They thus conclude that a stronger inter-rater reliability is perceived

when using an assessment grid but that the grid does not lead to a more consistent application of the rating criteria. Human proficiency ratings thus seem to be inherently biased and the use of assessment criteria is an attempt to reduce the variability and to counter human biases. However, even these efforts seem to fail to some extent. It might thus be worthwhile to combine oral rating procedures with more reliable assessment instruments. In the case of the proposed scale for *Grammatical Range*, I propose that oral language assessments can easily be combined with a Rapid Profile Analysis. These combined assessments might lead to a greater extent of reliability in language assessments.

4.4.5.3 Positive and Negative Wordings of Descriptors

Galaczi et al. (2011) express concerns about the positive wording of the CEFR descriptors and the brevity of certain CEFR scales in the rating scale construction and rater training. They found that the positive wording is a potential risk in language testing and that raters tend to be more comfortable with negative wordings in rating grids. This finding can be underpinned by my study. After the rating procedure, the amateur raters were given a scale with letters D-I and were asked to specify the criteria that they think they had applied to the rating for each level. I analyzed their comments and listed the raters according to their tendency to use negative or positive wording. I did this analysis because I assume that the favoring of positive or negative wording in the descriptors is a potential source of variability in rating results. As negative instances I counted those formulations that refer to faulty grammar, mistakes, lack of fluency, pausing, unusual pronunciation and lack of, or inappropriate vocabulary, etc. Positive wordings are those that refer to managing language features; fluent use of language, correct grammar, appropriate word choice, etc. All the instances that cannot be attributed to be either positive or negative are disregarded in this analysis.

Only six out of the 22 amateur raters generally tend to use more positive wording for their criteria than negative wording. However, all of the six raters used negative phrases/wording to formulate their criteria on lower proficiency

levels, i.e. levels D and E. Rater bA07, for example, provided the following rating criteria.


High proficiency	Level	Criteria
	I	The learner speaks in full sentences without making grammatical mistakes, His choice of words is appropriate, and the pronunciation is native-like.
	H	The learner speaks in full sentences, almost without making grammatical errors. His choice of words is appropriate. Only English is used.
	G	The learner speaks in rather short sentences. He is sometimes unsure about applying grammatical rules appropriately. Only English is used.
	F	The learner does not speak in full sentences. He is not able to apply grammatical rules. Only English is used. The pronunciation is very inappropriate.
	E	The learner is not able to produce statements in proper English. He has no awareness of the grammatical rules. Not only English is used. The pronunciation is hardly understandable.
Low proficiency	D	The learner is not able to find the right vocabulary without any help. There is no awareness of grammatical rules. The use of English words is very exceptional (sic!). The pronunciation is not understandable.

Figure 29: Rating Criteria given by bA07

For level D, amateur rater bA07 states that the vocabulary choice is limited. She refers to an explicit awareness of grammatical rules. Explicit awareness of grammar was not at focus in the tasks, rather the tasks aimed at implicit rule application. One might assume that this rater is not aware of the difference between explicit and implicit knowledge, so it can be inferred that this rater means that the learner does not apply the grammatical rules at this low level of proficiency. By the phrase “the use of English words is exceptional”, this rater most likely means that mostly the first language is used at this level. To level E, rater bA07 attributes more English words but still no correct grammar. At level F, this rater argues, only English is used but not in full sentences and grammar is still incorrect. Level G seems to be characterized by short sentences, but grammar is only “sometimes” incorrect, and, to rater bA07, word choice is appropriate at this level. Learners at the highest level, according to bA07, show a native-like pronunciation, appropriate word choice, no grammatical mistakes and they use full sentences. Here, the more positive wording is evident.

This rater states a number of factors that I assume are quite easy to assess: appropriate vocabulary, faulty grammar and first-language use. It is

rather evident that first language use plays a major role in what rater bA07 believes are the assessment criteria she had applied during the ratings. Native-language use is not an aspect that is considered in any of the oral assessments that I am aware of, because assessment centers usually assume only target-language use in their language tests. However, first language-use seems to be a fact that was easy to assess for this rater and that, to her, distinguishes levels D and E from levels F onwards. From the criteria that amateur rater bA07 accumulated, it might be deduced that vocabulary use, grammar, first language use and unusual pronunciation are factors that seem rather natural for the assessment of oral language. The rating criteria provided by amateur rater bA08 confirm this trend:


High Proficiency	Level	Criteria
	I	The learner speaks in full sentences without making grammatical errors, his choice of words is appropriate, and the pronunciation is native-like. Only English is used.
	H	The learner speaks in full sentences and makes only a few grammatical errors, his choice of words is mostly appropriate, and the pronunciation is almost native-like. Only English is used.
	G	The learner speaks in full sentences, makes a lot of grammatical errors and the choice of words is not always appropriate, but his statements are fluent and comprehensible. The speaker has a slight accent, there are some word demands in the learner's first language.
	F	The learner attempts to speak in sentences but makes a lot of grammatical errors and the choice of words is often not appropriate. His statements are not fluent and occasionally difficult to comprehend. The learner has an accent and uses his first language a few times.
	E	The learner attempts to speak in simple sentences, (sic!) but makes a lot of grammatical errors and the vocabulary used is limited. The learner has a strong accent when speaking English and uses his first language occasionally.
Low proficiency	D	The learner does not speak in full sentences and uses his first language a lot. The English vocabulary used is very limited, the learner uses his first language a lot and has a strong accent when speaking English.

Figure 30: Rating Criteria Rater bA08

Most of the criteria that she thinks she has applied, are concerned with vocabulary, grammar, first language use, pronunciation and fluency. It thus seems that these are the features that she naturally thinks the data can be assessed on. The criteria fluency, accuracy and range (referring mostly to

vocabulary use) are also stated in the Global Oral Assessment Grid (CoE 2009). Further criteria in the assessment grid provided by the CoE, are interaction and coherence. These cannot be found in either the criteria provided by bA07 or bA08. That these criteria are missing is most likely due to the task set that was used in this study. The tasks aim at eliciting the developmental stage of learners and not necessarily their proficiency levels. Therefore, it needs to be kept in mind that the tasks have most likely strongly influenced the rating criteria that the amateur raters provided. However, I assume that their criteria do indeed provide a hint as to what seems to be focused on intuitively by proficient language users who are asked to assess non-native speakers. These findings might have repercussions for assessments in teaching English as a foreign language-contexts.

Rater aA04 provides a systematic account to his rating criteria.


High Proficiency	Level	Criteria
	I	Fluent use of language; correct grammar; appropriate and variable vocabulary; correct pronunciation; across a high number of different situations/vocabulary fields; idiomaticity; no recourse to L1
	H	Mostly fluent use of language; mostly correct grammar; mostly appropriate vocabulary; mostly correct pronunciation; across a variety of different situations/language fields; almost no recourse to L1
	G	Partially fluent; few grammar insecurities; limited but mostly appropriate vocabulary; predominantly correct pronunciation of standard vocabulary; across a select (sic!) number of situations/vocabulary fields; limited recourse to L1
	F	Hesitant use of language (pauses); grammar insecurities; limited and (at times) inappropriate vocabulary; pronunciation insecurities; across a limited number of situations/vocabulary fields; limited recourse to L1
	E	Inarticulate stagnant use of language; faulty grammar; very limited and oftentimes inappropriate vocabulary, faulty pronunciation; across very few situations/vocabulary fields; frequent recourse to L1
Low proficiency	D	Extremely limited ability to use the language independently; incorrect basic grammar; extremely limited basic vocabulary; incorrect pronunciation of standard items; no situational adaptivity; very frequent recourse to L1

Figure 31: Rating Criteria by Amateur Rater aA04

In his criteria grammar, vocabulary and pronunciation are also mentioned but he also examined situational adaptivity and independent language use. Amateur rater aA04 seems to apply a rather systematic approach to his rating criteria

because he uses expressions at a continuum that ranges from “extremely limited” at level D to “variable” at level I in terms of vocabulary use. He thus uses gradations in his formulation of criteria along the levels. This is striking because the assessment criteria stated in the Global Oral Assessment Grid are not characterized by this systematicity because no gradations are consistently mentioned along the levels and also, new criteria are stated at higher levels that were not mentioned at earlier levels. This was criticized by Green, for example (Green 2012: 60). The incrementality that rater aA04 seems to have assessed thus shows more consistency and systematicity than the actual assessment grid produced by the CoE.

To summarize, the amateur raters seem to tend to listen for positive and negative features in the language sample. They seem to quantify the features by mostly using words or descriptors like “shows no XXX, show a lot of XXX, limited XXX, not XXX”. It can be assumed that it is easier to determine whether something is done, not done or done in a limited way than to use descriptors like “some”, “fairly even” and so on which are presented in the assessment grid by the CoE. Even the amateur raters seem to display some kind of systematicity in their rating criteria that they formulated after the ratings. The criteria that the amateur raters noted down differ from those given in the Global Oral Assessment Grid mainly by the use of the first language and a lack of criteria referring to coherence and interaction.

The inferences that are made in this chapter need to be examined in future studies because, as stated above, the criteria the amateur raters produced are most likely influenced by the general outlook of the task set that the raters were presented with. The tasks that were used in the audio-files are geared towards assessing the developmental stage of the learner within the PT framework (see section 4.3.4). If the tasks covered different activities, the rating criteria that the amateur raters noted down might have looked differently. However, this general tendency to listen for positive and negative criteria was visible with at least one rater from the expert rater group who also stated his rating strategy as follows:

aE05	I use the same technique for all listening, that is listening for negatives (eg. poor grammar) and for positives (eg. good words). Also communication, do I understand what the person is telling me? Not the individual words, but the sentences and paragraphs. There may be mistakes, but do they detract from what the person is saying. As well as using the criteria grid. This may sound a bit rambling, but I try to focus on the whole thing, rather than the positives and negatives.
------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

In this data sample, all amateur raters referred to grammar as one component part for judging the learners' oral performance. This fact underpins my hypothesis that grammar intuitively plays an important role in oral language assessment.

The next chapter displays how different rating techniques were used despite instruction to use the Global Oral Assessment Grid.

4.4.5.4 Different Rating Techniques

Deygers et al (2018) found that even trained raters interpret the same test-specific criteria differently. Deygers et al.'s findings can be underpinned in this study by the example of the expert rater group. The expert raters were asked to comment on their rating techniques in the first week. This was an open question and the raters were free to be as specific as they wanted.

From the analysis of the expert rater comments, I deduce that the raters employed vastly different rating techniques, although all of them can be regarded as experienced raters and all of them, in principle, should have used the assessment grid. Some raters only briefly described that they used the rating grid that I had asked them to use, such as rater bE03 who was affiliated to TELC at the time of data collection:

bE03	Here are my results that I based on the table C1 which was attached to the email detailing the study.
------	-------------------------------------------------------------------------------------------------------

Other raters' comments were more detailed, for example, aE05 describes that

aE05	I use the same technique for all listening, that is listening for negatives (eg. poor grammar) and for positives (eg. good words). Also communication, do I understand what the person is telling me? Not the individual words, but the sentences and paragraphs. There may be mistakes, but do they detract from what the person is
------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

	saying. As well as using the criteria grid. This may sound a bit rambling, but I try to focus on the whole thing, rather than the positives and negatives.
--	------------------------------------------------------------------------------------------------------------------------------------------------------------

His comment was also used in chapter 4.4.5.3 on positive and negative wording of descriptors, but it also shows in to what extent he employs a different technique than rater bE03. Rater aE05 was affiliated to TELC at the time of data collection. He describes that he uses the assessment grid - which he seems to think points towards positives and negatives - but also tries to focus on the presentation of the language as a whole. Factors seem to include to listening for gist and communication “do I understand what the person is telling me?” and “[...] mistakes, but do they distract from what the person is saying”. Other than what the amateur raters stated (see section 4.4.5.3), rater aE05 rather focuses on the communication than the absence or presence of specific language features and communicative functions. It seems as if this trained rater rather uses his overall impression of the learner to base his assessment on than to adhere to the details laid out in the assessment grid.

Another expert rater who was also affiliated to TELC at the time of the data collection, is aE02. She seems to use a very different approach for her assessment than rater aE05:

aE02	Read the CEFR criteria – again and again. If I’m examining at one particular level (or a dual level) exam I jot down examples of utterances that meet the criteria - and those that might not. In an exam situation I welcome a short discussion with a fellow examiner (preferably a global English one, as I’m a native speaker). I usually like (time permitting) to go over the first people examined/assessed, as the first time round is a bit of a calibration exercise.
------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

From her comment, it seems that rater aE02 focuses primarily on the descriptors and tries to match the language that the learner produces to the descriptors. She also seems to like to discuss her rating with a fellow examiner so as to calibrate their assessment criteria to each other.

What can be seen from the comments displayed above is that although all of these raters are affiliated to and were trained by the same assessment center, they seem to use very different approaches in their ratings. Whereas

aE02 and bE03 used the rating grid, rater aE05 rather supplements the rating grid with his general impression of the communicative ability of the learner.

After having presented and discussed the results of this study, the following chapter concludes this thesis and states directions for future research.

5. Conclusion and Future Directions

It was the aim of this study to find interfaces between the CEFR and PT in order to add to the theoretical, descriptive, as well as empirical basis of the CEFR since the lack of these three instances in the CEFR has been heavily criticized.⁸² I hypothesized that PT can complement the CEFR in terms of grammatical development. My overall aim was to propose a scale for *Grammatical Range* that is based on correlations between PT stages and CEFR levels. *Grammatical Range* is grounded in the original scale for Grammatical Accuracy presented in the CEFR. *Grammatical Range* encompasses the universal processing procedures put forward by PT and thus takes a universal processing perspective on grammatical development. It does not contain any references to grammatical accuracy since it was established that accuracy does not mirror development and that an accuracy criterion would lead to false assumptions about the productiveness of learner language (see chapter 2.2.4). The scale for *Grammatical Range*, in my view, is compatible with the action-oriented approach taken in the CEFR and tries to incorporate the concept of interlanguage that PT works with. Despite some differences in the notions of competence, progression, universality, accuracy and emergence (see chapter 2.2.4), I argue that the modular approach taken in PT can be integrated into the open, descriptive and undogmatic CEFR. This integration is especially viable since PT's universal processing procedures substantiate the CEFR's representation of grammatical competence and makes it more learner-centered (see chapter 3).

⁸² See chapter 2 for a more detailed discussion of the criticism connected to the lack of a theoretical and empirical basis in the CEFR.

I argue that in order for PT to be able to complement the CEFR, I would first have to determine the role of grammar in CEFR-based ratings for overall oral production. I consider oral production especially useful for this study, as PT, in its current version, is mainly concerned with the oral production of morphosyntactic features in SLA. Moreover, as of today, grammar seems to be a somewhat neglected area in the CEFR. In order to investigate the role of grammar in the CEFR and proficiency ratings based on the CEFR, I employed an innovative methodology that elicits the rater focus in a direct manner. I edited recordings of oral learner language for morphological markers in such a way that the markers that had been determined as very inaccurate in a prior study were deleted in the recordings. This way, I compiled a corpus of 22 oral language samples, eight of which were edited for grammatical accuracy. The samples thus only differ in one feature, i.e. grammatical accuracy. Raters were presented with both the original and the edited files and in 84 cases, the edited file was rated on a lower CEFR level than the original file (see chapter 4.4.1). The results thus show that generally, grammar seems to be a strong determiner in rating overall oral proficiency, although it is only one out of five possible performative skills to be rated for (see chapter 4.4.1 for more details). These results allow me to conclude that grammar as a factor in oral proficiency ratings should be awarded more attention and that it therefore makes sense to investigate interfaces between the CEFR and PT, because PT is mainly concerned with grammatical development. Chapter 4.4.4 shows the results on interfaces between the CEFR and PT. Oral language samples were rated according to the CEFR and analyzed with Rapid Profile so as to analyze their PT stage. The CEFR levels and PT stages were correlated. In chapter 4.4.4, it is illustrated that there are significant correlations between both frameworks, especially at the lower learner levels. These significant correlations lead to proposing a combined scale for *Grammatical Range*. The proposed scale for *Grammatical Range* is formulated as close to the original voice of the CEFR scale for grammatical accuracy but is combined with the universal processing procedures as put forward by PT. Also, all references to accuracy criteria in the descriptor items are deleted. They are deleted because I argue that learner language is necessarily inaccurate during the acquisition

process and that therefore, a conceptualization of grammatical competence only in terms of accuracy (in the quantitative part of the CEFR) paints a learner-unfriendly picture.

Due to my methodology, I was also able to comment on methodological issues in proficiency ratings. I examined three groups of raters: amateur raters who used intuition for their ratings, novice raters who were trained in the CEFR and the use of assessment grids (as suggested by CoE 2009) and expert raters who were affiliated to language assessment centers at the time of the data collection. Chapter 4.4.7 presents the results on rater experience. My data suggests that there is no significant difference between novice raters and experienced raters in terms of the variability of the rating results. Thus, it seems not to make a difference if raters are newly trained or if they have had a lot of experience in proficiency ratings. I also examined whether it makes a difference to use an assessment grid in proficiency ratings as compared to intuition. Amateur raters did not use an assessment grid, whereas novice and expert raters used the same grid. It was shown that indeed, the amateur rater group produced more variable results than the other two groups. However, when investigating inter-rater reliability, the amateur rater group in itself produced results almost as reliable as the other two subgroups. This is a finding that needs to be investigated in further studies.

In general, the findings presented in this thesis need to be treated tentatively. The reason for this is that at 53 raters, the number of participants is relatively small. Also, only two files represent each PT stage, so the correlations between the PT stages and CEFR levels rely only on the two learners. What is more, no relationship between a PT stage and CEFR level B2 was found, as raters simply did not assign this level often enough to have an effect on the correlation. Thus, the correlations between PT and the CEFR should be investigated more closely in future studies. It would also make sense to turn the data elicitation process around. That means that the learners would need to be assessed for the CEFR first and only afterwards assessed with Rapid Profile. This way, the ratings might not be affected by the unfamiliarity of the tasks that are usually used for

eliciting PT stages (see chapter 4.3.4). The scale for *Grammatical Range* should be validated in future empirical and theoretical research.

I do, however, assume that despite these limitations, the thesis contributes to adding to the descriptive, empirical and theoretical basis of the CEFR in proposing a more learner-centered, empirically grounded and theoretically-motivated scale for *Grammatical Range*. The study also provides interesting findings with regard to psychometric testing and rater experience. The development of a new, direct approach to eliciting the rater focus is an innovation in the research on the factor “rater” in psychometric rating procedures.

6. References

- Abel, A., Nicolas, L., Wisniewski, K., Boyd, A., & J. Hana (2014). A trilingual learner corpus illustrating European reference levels. *Ricognizioni. Rivista di Lingue e Letterature e Culture Moderne*, 2(1), 111–126.
- Alderson, J.C. Figueras, N. Kuijper, H. Nold, G. Takala, S. & C. Tardieu (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: An International Journal* 3 (1), 3-30.
- Alderson, J. (2007). The CEFR and the Need for More Research. *The Modern Language Journal*, 91, 659-663.
- Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Baddeley, A. (1986). *Working Memory*. Oxford: Oxford Psychology Series.
- Barnett, J. (1980). National/Functional Approaches. *Annual Review of Applied Linguistics* 1, 43-57.
- Bartning, I. & S. Schlyter (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies* 14, 281-299.
- Bartning, I.; Martin, M. & I. Vedder (2010). (Eds.). *Communicative Proficiency and Linguistic Development. Intersections between SLA and Language Testing Research*. Eurosla Monograph Series.
- Bickerton, D. (1981). *Roots of Language*. Ann Arbor MI: Karoma Press.
- Berwick, R.C. & A.S. Weinberg (1984). *The grammatical basis of linguistic performance*. Cambridge: MIT Press.
- Bond, T.G. & C.M. Fox (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences*. (3rd Ed.). Mahwah: Erlbaum.
- Bongaerts, T. & P. Jordens (1985). Ontwikkelingen in tussentaal. Tweede taalontwikkelingen als herstructureringsproces. *Tijdschrift voor Taal- en Tekstwetenschap* 5, 231-245.
- Bresnan, J., & Kaplan, R.M. (1982). Introduction: Grammars as mental representations of language. In: J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp. xvii-lii). Cambridge, Mass.: MIT Press.
- Bresnan, J. & J. M. Kanerva (1989). Locative inversion in Chichewa: A case study of factorization in grammar. *Linguistic Inquiry* 20, 1-50.
- Bresnan, J. (2001). *Lexical Functional Syntax*. Oxford: Blackwell.

- Brindley, G. (1986). *The assessment of second language proficiency: issues and approaches*. Adelaide: National Curriculum Resource Centre.
- Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum*. NCELTR Research Series. Sydney: Macquarie University.
- Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing*, Singapore: Regional Language Centre.
- Brindley, L. F. (1998). Describing Language Development? Rating Scales and SLA. In: Bachman, L. F. and A.D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, 112-140.
- Brown, R. (1973). *A first language: The early stages*. Cambridge, Mass.: Harvard University Press.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific performance test. *Language Testing* 12, 1-15.
- Brown, A. (2000). An investigation of the rating process in the IELTS Speaking Module. In R. Tullloh (Ed.), *Research reports* (1999, Vol. 3). Canberra: IELTS Australia, pp. 49-85.
- Brown, A; Iwashita, N. & T. McNamara (2005). *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks*. Melbourne: ETS TOEFL Monograph Series.
- Buttery, P. & A. Caines (2012). *Reclassifying subcategorization frames for experimental analysis and stimulus generation*. www.cl.cam.ac.uk/~alk23/lrec06-lexicon.pdf (last access 09.09.17).
- Bond, T. G. & C.M. Fox (2015). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*, 3rd Ed. New York: Routledge.
- Byram, M. (1997). *Teaching and Assessing Intercultural Communicative Competence*. Clevedon: Multilingual Matters.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1), 1–47. <http://dx.doi.org/10.1093/applin/i.1.1>
- Canale, M. (1983). On some dimensions of language proficiency. In J. W. J. Oller (Ed.), *Issues in language testing research*. Rowley, MA: Newbury House, pp. 333–342.
- Carlsen, C. (2010). Discourse connectives across CEFR-levels: A corpus-based study. In I. Bartning, M. Martin, & I. Vedder (Eds.). *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 191–210). Amsterdam: European Second Language Association. Retrieved from <http://eurosla.org/monographs/EM01/EM01tot.pdf> (last access: 30.04.2018).

- Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 16-35.
- Chan, Y. & Walmsley, R.-P. (1997). Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical Therapy*, 77, 1755-1762.
- Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, 37(6), 849–880.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
- Chomsky, N. (1980). *Rules and Representations*. Oxford: Basil Blackwell.
- Clahsen, H. (1980). Psycholinguistic Aspects of L2 Acquisition: Word Order Phenomena in Foreign Worker's Interlanguage. In *Second Language Developments: Trends and Issues*, S.W. Felix (Ed.), 57-79. Tübingen: Narr.
- Clahsen, H.; Meisel, J. & M. Pienemann (1983). *Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr.
- Clahsen, H. (1984). The Acquisition of German Word Order. A Test Case for Cognitive Approaches to Second Language Development. In *Second Languages*. R. Andersen (Ed.), 219-242. Cambridge, Mass.: Newbury House.
- CoE (1954). *European Cultural Convention*. European Treaty Series (18). Strasbourg: Council of Europe.
- CoE (1973). *Systems Development in adult language learning*. Strasbourg: Council of Europe.
- CoE (1982). *Recommendation No. R (82) 18 of the Committee of Ministers to Member States Concerning Modern Languages*. Available from: <https://rm.coe.int/16804fa45e> (last access: 07.07.2018).
- CoE (1990). *Threshold Level 1990 (Modern Languages)*. Strasbourg: Council of Europe.
- CoE (1996). *Modern Languages: Learning, Teaching Assessment: A Common European Framework of Reference*. Draft 2 of a Framework proposal. Strasbourg: Council of Europe.
- CoE (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- CoE (2006). *Special Eurobarometer 243. Europeans and Their Languages*. http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_243_en.pdf (last access: 30.04.2018).

- CoE (2009). *Relating Language Examinations with the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*. Strasbourg: Language Policy Division.
- CoE (2011). *Manual for Language Test Development and Examining. For Use with the CEFR*. Strasbourg: ALTE/Language Policy Division.
- CoE (2016). *CEFR Illustrative Descriptors. Extended Version 2016. Pilot version for consultation*. Strasbourg: Language Policy Division.
- Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly*, 29, 762–765.
- Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics* 5(4), 161-170.
- Crossley, S. A. & D. S. McNamara (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35, 115–135.
- Crystal, D., Fletcher, P. & M. Garman (1976). *The grammatical analysis of language disability*. London: Edward Arnold.
- Crystal, D. & P. Fletcher (1979). Profile Analysis of Language Disability. In: Fillmore, C.; Kempler, J. & W. Wang (eds). *Individual differences in language ability and language behaviour*. New York: Academic Press, pp. 167-188.
- Crystal, D., Fletcher, P. & Garman, M. (1989). *The grammatical analysis of language disability*. London: Cole & Whurr.
- Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. *Review of Educational Research*, 49, 222-251.
- Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 70–89). Cambridge, England: Cambridge University Press.
- Cumming, A., Kantor, R., & D. E. Powers (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph No. MS-22). Princeton, NJ: ETS.
- Cumming, A., Kantor, R., & D. E. Powers (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, (86), 67-96.
- Dalgaard, P. (2008). *Introductory Statistics with R*. Luxemburg: Springer Science & Business Media, 99-100.
- Dalrymple, M. (1999). *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. Cambridge, MA: The MIT Press.

- Dalrymple, M. (2001). *Syntax and Semantics. Lexical Functional Grammar. Vol 34*. San Diego: Academic Press.
- Davies, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33(1)*, 117-135.
- De Bot, K. (1992). A bilingual production model: Levelt's 'speaking' model adapted. *Applied Linguistics 13(1)*, 1-24.
- de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34(1)*, 5-34.
- Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32(4)*, 521–541.
- Deygers, B.; Van Gorp, K. & T. Demeester (2018). The B2 Level and the Dream of a Common Standard. *Language Assessment Quarterly*, DOI: 10.1080/15434303.2017.1421955
- DiBiase, B. (2002). Focusing Strategies in Second Language Development: a classroom-based Study of Italian L2 in Primary School. In DiBiase, B. (Ed.), *Developing a Second Language: Acquisition, Processing and Pedagogy of Arabic, Chinese, English, Italian, Japanese, Swedish*. Melbourne; Language Australia, pp. 95-120.
- DiBiase, B. & S. Kawaguchi (2002). Exploring the typological plausibility of Processability Theory: Language development in Italian second language and Japanese second language. *Second Language Research, 18(3)*, 272–300.
- DiBiase, B. (2008). Focus on Form and Development in L2 Learning. In: Keßler, J.-U. (Ed.), *Processability Approaches to Second Language Development and Second Language Learning*. Cambridge: Cambridge Scholar Publishing, pp. 197-220.
- Díez-Bedmar, M. B. (2015). Article Use and Criterial Features on Spanish EFL Writing: a Pilot Study from CEFR A2 to B2 Levels. In: Callies, M. & S. Götz (Eds.). *Learner Corpora in Language Testing and Assessment*. New York/Amsterdam: John Benjamins, pp.163-190.
- Ding, T. (2012). The Comparative Effectiveness of Recasts and Prompts in Second Language Classrooms. *Journal of Cambridge Studies 7(2)*, 83-97.
- Dulay, H.C. & M.K. Burt (1974). Errors and Strategies in Child Second Language Acquisition. *TESOL Quarterly 8(2)*, 129-136.
- Dulay, H. C., Burt, M. K., & Krashen, S. (1982). *Language two*. New York: Oxford University Press.
- du Plessis, J.; Solin, D.; Travis, L. & L. White (1987). UG or not UG, that is the question: a reply to Clahsen and Muysken. *Second Language Research 3*, 56-75.

- Dvořáková, B. (2012). *Communicative Competence in Second Language Acquisition*. https://theses.cz/id/0ptzxx/Communicative_Competence_in_Second_Language_Acquisition.pdf. (last access: 16.10.2017)
- East, M. (2016). Assessing Foreign Language Students' Spoken Proficiency. *Educational Linguistics* 26. DOI 10.1007/978-981-10-0303-5_2
- Eckerth, J. (2008). Task-based Language Learning and Teaching – Old Wine in New Bottles? In Eckerth, J. & S. Siekmann (Eds.), *Task-based Language Learning and Teaching. Theoretical, Methodological, and Pedagogical Perspectives*. Frankfurt & New York: Peter Lang, pp. 13-46.
- Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly* 2(3), 197-221.
- Edorsy, M.U. (2004). Exploring Variability in Judging Writing Ability in a Second Language: A Study of Four Experienced Raters of ESL Compositions. *ETS TOEFL Research Reports*, 70, 1-72.
- Ellis, R. (1985). Sources of Variability in Interlanguage. *Applied Linguistics* 6, 118-131.
- Ellis, R. (1989). Are Classroom and Naturalistic Acquisition the Same? A Study of the Classroom Acquisition of German Word Order Rules. *Studies in Second Language Acquisition*, 11(3), 305-328.
- Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Ellis, R. (2003). *Task based language learning and teaching*. Oxford: Oxford University Press.
- Ellis, R. (2009). Task-based Language Teaching: Sorting out the Misunderstandings. *International Journal of Applied Linguistics* 19(3), 221-246.
- Fabri, R. (2008). Lexical Functional Grammar. In J.-U. Keßler (Ed.), *Processability Approaches to Second Language Acquisition: Development and Second Language Learning*. Newcastle upon Tyne: Cambridge Scholars, 31-67.
- Faerch, C & Kasper, G (1983). (Eds.), *Strategies in Interlanguage Communication*. London: Longman.
- Falk, Y. (2001). *Lexical Functional Grammar: An Introduction to Parallel Constraint-based Syntax*. Stanford, CA: CSLI.
- Felix, S.W. (1984). Maturational Aspects of Universal Grammar. In A. Davies, C. Cripser & A Howard (Eds.), *Interlanguage*. Edinburgh: Edinburgh University Press.
- Fieller, E. C.; Hartley, H. O.; Pearson, E. S. (1957). Tests for rank correlation coefficients. *Biometrika* 44, 470–481.

- Frawley W. & J. P. Lantolf (1985). Second Language Discourse: A Vygotskyan Perspective. *Applied Linguistics*, 6 (1), 19-44.
- Ghassan, A.S. (2008). The development of verbal structures in L2 Arabic. In J.-U. Keßler (Ed.). *Processability approaches to second language development and second language learning*. Newcastle, UK: Cambridge Scholars Publishing, pp. 267-299.
- Galaczi, E., Ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice*, 18(3), 217–237.
- Garman, M. (1990). *Psycholinguistics*. Cambridge: Cambridge University Press.
- Geng, J. (2010). The Semantic Analysis of the Definite Article' Misuse by Chinese Learners of English. *Asian Social Science* 6(7), 180-184.
- Granfeldt, J.; Nugues, P.; Persson, E.; Persson, L.; Kostadinov, F.; Ågren, M.; & S. Schlyter (2005). *Direkt Profil: A System for Evaluating Texts of Second Language Learners of French Based on Developmental Sequences*. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, 53–60, Ann Arbor. Accessible from: <https://aclanthology.info/pdf/W/W05/W05-0209.pdf>
- Granfeldt, J. & P. Nugues (2007). Evaluating stages of development in second language French: A machine-learning approach. In J. Nivre, H.-K. Kaalep, K. Muischnek & M. Koit (Eds.), *NOALIDA 2007 Conference Proceedings*, 73-80.
- Granfeldt, J. & M. Agren (2014). SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing*, 31(3), 285-305.
- Granget, C. (2009). L'acquisition de la morphologie verbale. In: D. Veronique; C. Carlo, C. Granget, J. Kim & M. Prodeau (Eds.), *Acquisition de la grammaire du français langue étrangère*. Paris: Didier, 164-218.
- Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. Cambridge: Cambridge University Press.
- Grimm, N.; Meyer, M. & L. Volkman (2015). *Teaching English*. Tübingen: Narr.
- Gyllstad, H., Granfeldt, J., Bernardini, P., & Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook*, 14, 1–30.
- Hagenfeld, K. (2013). Psychometric approaches to languages testing and linguistic profiling – a study on the CEFR, Rapid Profile and Autoprofilin. Unpublished Thesis.
- Hagenfeld, K. (2017). Psychometric approaches to languages testing and linguistic profiling – a complementary relationship? In J.U. Keßler, A. Lenzing & M. Liebner (Eds.), *Developing, Modelling and Assessing Second Languages* (pp. 135-162). Amsterdam: John Benjamins.

- Håkansson, G. (2005). Similarities and differences in L1 and L2 development: Opening up the perspective: Including SLI. In M. Pienemann (Ed.), *Crosslinguistic aspects of Processability Theory* (pp.179-197). Amsterdam: John Benjamins.
- Håkansson, G., & Norrby, C. (2007). Processability Theory applied to written and oral Swedish. In F. Mansouri (Ed.), *Second language acquisition research: Theory-construction and testing* (pp.81-94). Newcastle, UK: Cambridge Scholars Press.
- Halliday, M.A.K. (1978). *Language as a Social Semiotic. The Social Interpretation of Language and Meaning*. London: Edward Arnold.
- Harley, B., J. Cummins, M. Swain, and P. Allen (1990). *The Development of Second Language Proficiency*. NY: Cambridge University Press.
- Harsch, C. (2006). *Der Gemeinsame Europäische Referenzrahmen für Sprachen: Leistungen und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI Projekt*. PhD thesis. Accessible from: <https://d-nb.info/980854466/34> (last access: 07.07.2018)
- Harsch, C., & Rupp, A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A testcentered approach. *Language Assessment Quarterly*, 8(1), 1–33.
- Hatch, E. (1978). Apply with caution. *Studies in Second Language Acquisition*, 2(1), 123-143.
- Hawkins, J. A., & Buxter, P. (2010). Criterial features in learner corpora: Theory and Illustrations. *English Profile Journal*, 1, 1–23.
- Hawkins, J. & L. Filipović (2012). *Criterial Features in L2 English*. Cambridge: Cambridge University Press.
- Henning, G. (1992). Dimensionality and Construct Validity of Language Tests. *Language Testing* 9 (1), 20-31.
- Holec, H. (1979). *Autonomy and Foreign Language Learning*. Strasbourg: Council of Europe.
- Huebner, T. (1979). Order-of-Acquisition vs. Dynamic Paradigm: A Comparison of Method in Interlanguage Research. *TESOL Quarterly*, 13(1), 21-28.
- Huebner, T. (1983). *A Longitudinal Analysis of the Acquisition of English*. Ann Arbor MI: Karoma.
- Hulstijn, J.H. (1985). Testing second language proficiency with direct procedures. A response to Ingram. In Hyltenstam K. & M. Pienemann (Eds.), *Modelling and assessing second language development*. Clevedon: Multilingual Matters, 277–282.
- 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.; J.C. Alderson & R. Schoonen (2010). Developmental Stages in Second-Language Acquisition and levels of second language proficiency: Are there links between them? In I. Bartining, M. Martin & I. Vedder (Eds.), *Communicative Proficiency and Linguistic Development. Intersections Between SLA and Language Testing Research*. EUROSLA Monograph Series 1, 11-20.
- Hulstijn, J.H. (2011). Language proficiency in native and nonnative speakers: an agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8(3), 229-249.
- Hulstijn, J.H.; R. Schoonen; N.H. de Jong; M.P. Steinel & A. Florijn (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, 29(2), p.203-231.
- Hymes, D. (1972). On Communicative Competence. In Duranti, A. (Ed.), *Linguistic Anthropology*. Oxford, pp.: 53-73.
- Isaacs, T. & R. I. Thomson (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10, 135–159.
- James, (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. Harlow, UK: Addison Wesley Longman.
- Joe, J. M.; Harmes, C. & C. A. Hickerson (2011). Using verbal reports to explore rater perceptual process in scoring a mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy and Practice*, 18(3), 239-258.
- Johnston, M. (1985). *Syntactic and morphological progressions in learner English*. Canberra: Commonwealth Department of Immigration and Ethnic Affairs.
- Kaplan, R.M., & J. Bresnan (1982). Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (pp.173-281). Cambridge, Mass.: MIT Press.
- Kaplan, R.M. & J. Bresnan (1995). LexicalFunctional Grammar - A Formal System for Grammatical Representation. In Dalrymple, M., R. M. Kaplan, J. T. Maxwell, and A. Zaenen (Eds.), *Formal Issues in Lexical-Functional Grammar*. Stanford, CA: CSLI Publications.
- Kawaguchi, S. (2005). Argument structure and syntactic development in Japanese as a second language. In M. Pienemann (Ed.), *Cross-linguistic aspects of Processability Theory* (pp.253-298). Amsterdam: John Benjamins.
- Keßler, J.-U. (2006). *Englischerwerb im Anfangsunterricht diagnostizieren. Linguistische Profilanalysen am Übergang von der Primarstufe in die Sekundarstufe I*. Tübingen: Narr.

- Keßler, J.-U. (2007). Assessing EFL Development Online: A Feasibility Study of Rapid Profile. In: Mansouri, F. (Ed.), *Second Language Acquisition Research: Theory Construction and Testing* (pp. 118-135). Newcastle upon Tyne: Cambridge Scholars.
- Keßler, J.-U. (2008). Communicative Tasks in Second Language Profiling: Linguistic and Pedagogical Implications. In: Eckerth, J. & S. Siepmann (Eds.), *Research on Task-based Language Learning and Teaching: Theoretical, Methodological and Pedagogical Perspectives* (pp. 291-310). Frankfurt: Peter Lang.
- Keßler, J.-U. & D. Keatinge (Eds.) (2009). *Research in Second Language Acquisition: Empirical Evidence Across Languages*. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Keßler, J.-U.; Liebner, M.; F. Mansouri (2011). Teaching. In: Pienemann, M. & J.-U. Keßler. (Eds.), *Studying Processability Theory* (pp. 149-156). Amsterdam: John Benjamins.
- Keßler, J.-U. & A. Plessner (2011). *Teaching Grammar*. Paderborn: Schöningh.
- Kempen, G., & E. Hoenkamp (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, 11(2), 201-258.
- Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? *Pädagogik* 6, 10-13.
- Krashen, S. (1977). Some issues relating to the monitor model. In H. Brown, C. Yorio & R. Crymes (Eds.). *On TESOL '77* (pp.144-158). Washington D.C.: TESOL.
- Krashen, S., & R. Scarcella (1978). On Routines and Patterns in Language Acquisition and Performance. *Language Learning*, 28(2), 283-300.
- Krashen, S.D. (1981). *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon.
- Krashen, S.D. (1985). *The Input Hypothesis: Issues and Implications*. New York: Longman.
- Lado, R. (1961). *Language Testing: The Construction and Use of Language Tests*. New York: McGrawhill, pp.47-66.
- Lantolf, J. & W. Frawley (1988). Proficiency, understanding the construct. *Studies in Second Language Acquisition*, 10, 181–196.
- Larsen-Freeman, D. & M. H. Long (1991). *An Introduction to Second Language Research*. London: Longman.
- Leclercq, P.; Edmonds, P & H. Hilton (2014). (Eds.), *Measuring L2 Proficiency. Perspectives from SLA*. Bristol: Multilingual Matters.
- Legendre, P. (2005). Species Associations: The Kendall Coefficient of Concordance Revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2), 226–245.

- Lenzing, A. & A. Plesser (2010). *Challenging the scope-precision dilemma in language testing: The common European framework and linguistic profiling*. Paper presented at the 10th International Symposium of Processability Approaches to Language Acquisition (PALA). University of Western Sydney, Australia, 19.-21.09.10.
- Lenzing, A. & J. Roos (2012). Die sprachliche Entwicklung und die Ausdrucksmöglichkeiten von Grundschülerinnen und Grundschülern im Englischunterricht. In: Bär, M.; Bonnet, A.; Decke-Cornill, H.; Grünewald, A. & A. Hu (Eds.), *Globalisierung – Migration – Fremdsprachenunterricht. Dokumentation zum 24. Kongress für Fremdsprachendidaktik der Deutschen Gesellschaft für Fremdsprachenforschung (DGFF)* (pp. 207-220). Hamburg: Baltmannsweiler: Schneider Verlag Hehengehren.
- Lenzing, A. (2013). *The Development of the Grammatical System in Early Second Language Acquisition. The Multiple Constraints Hypothesis*. Amsterdam: John Benjamins.
- Lenzing, A.; Plesser, A.; Hagenfeld, K. & M. Pienemann (2013). Transfer at the Initial State. *Zeitschrift für Anglistik und Amerikanistik. A Quarterly of Language, Literature and Culture*, 61(3), 265-287.
- Lenzing, A. (2016). The development of argument structure in the initial L2 mental grammatical system. In: Keßler, J.-U.; Lenzing, A. & M. Liebner (Eds.), *Developing, Modelling and Assessing Second Languages* (pp. 3-33). Amsterdam: John Benjamins.
- Lenzing, A. (2017). *The Production-Comprehension Interface in Second Language Acquisition: An Integrated Encoding-Decoding Model*. University of Paderborn: Habilitationsschrift.
- Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41-104.
- Levelt, W.J.M. (1989). *Speaking. From intention to articulation*. Cambridge, Mass.: MIT Press.
- Liebner, M. & M. Pienemann (2011). Explaining learner variation. In: Pienemann, M. & J.-U. Keßler (eds.) *Studying Processability. An introductory textbook*. John Benjamins: Amsterdam/New York, 64-74.
- Lightbown, P. & N. Spada (1999). *How Languages are Learned*. Oxford: Oxford University Press.
- Lin, B.J. (2012). *Is Automatic Linguistic Profiling Feasible in a ESL Context?* University of Newcastle: PhD thesis.
- Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, 39, 167-190.
- Little, D. (2007). *The Common European Framework of Reference for Languages and the development of policies for the integration of adult migrants*. Council of Europe:

Strasbourg.
http://www.coe.int/t/dg4/linguistic/Source/Little_CEFRmigrants_EN.doc (last access 20.02.13).

Little, D. (2014). *Learning, teaching, assessment: An exploration of their interdependence in the CEFR*. Presentation at the 11th Annual Conference of EALTA, Warwick, UK. <http://www.ealta.eu.org/conference/2014/programme.html>. (last access 01.12.2017)

Liu, Yan; Wu, Amery D. & B. D. Zumbo (2010). The Impact of Outliers on Cronbach's Coefficient Alpha Estimate of Reliability: Ordinal/Rating Scale Item Responses. *Educational and Psychological Measurement* 70(1), 5–21.

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, 4, 126–141.

Long, M. H. (1985). Input and second language acquisition theory. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition*. Rowley, MA: Newbury House, pp. 377–393.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition*. New York: Academic Press, pp. 413–468.

Long, R. (1991). Focus on Form: A Design Feature in Language Teaching Methodology. In K. de Bot, R. Ginsberg & C. Kramsch (Eds.), *Foreign Language Research in Cross-Cultural Perspective* (pp.: 29-52), *Studies in Bilingualism* 2nd edition. Amsterdam: John Benjamins.

Long, M., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-63). Cambridge: Cambridge University Press.

Long, M.H. (2011). Methodological Principles for Language Teaching. In: Doughty, C. & M.H. Long (Eds.), *Handbook of Language Teaching* (pp.: 15-41). Oxford: Wiley-Blackwell.

Lyster, R. & L. Ranta (1997). Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms. *Studies in Second Language Acquisition*, 19(1), 37-66.

Mackey, A.; Pienemann, M. & I. Thornton (1991). Rapid Profile: A Second Language Screening Procedure. *Language and Language Education* 1(1), 61-82.

Mackey, A. (1992). Targeting morpho-syntax in children's ESL: An empirical study of the use of interactive goal-based tasks. *Working Papers in Educational Linguistics*, 10(1), 67-91.

Mackey, A. (1999). Input, Interaction and Language Development. *Studies in Second Language Acquisition* 21(4), 557-587.

- Mackey, A. (2006). Feedback, Noticing and Instructed Second Language Acquisition. *Applied Linguistics* 27(3), 405-430.
- Maier, E.; Neubauer, L. Ponto, K. Couve de Murville, S. & K. Kersten (2016). Assessing Linguistic Levels of L2 English in Primary School Programms. In: Keßler, J.-U.; Lenzing, A. & M. Liebner (Eds.), *Developing, Modelling and Assessing Second Languages* (pp. 163-192). Amsterdam: John Benjamins.
- Mansouri, F., & Duffy, L. (2005). The pedagogic effectiveness of developmental readiness in ESL grammar instruction. *Australian Review of Applied Linguistics*, 28(1), 81-99.
- Martyniuk, W. (Ed.). *Relating Language Examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*. Cambridge: Cambridge University Press.
- May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing*, 11(1), 29-51.
- McNamara, T. & T. Lumley (1997). The effect of interlocutor and assessment mode variables in overseas assessment of speaking skills in occupational settings. *Language Testing* 14, 140-156.
- McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery* 46, 38-41.
- Meisel, J., H. Clahsen & M. Pienemann (1981a). On determining developmental sequences in natural second language acquisition. *Studies in Second Language Acquisition* 3(2), 109-135.
- Michalska, B. (2010). *Interface between language testing and SLA research: Investigating correlation between proficiency rating scales of the Common European Framework and developmental Stages of Rapid Profile*. Newcastle University: Unpublished Thesis.
- Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium*. Cambridge, UK: Cambridge University Press, pp. 92-114.
- Milanovic, M. & N. Saville (2012). *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*. Cambridge: Cambridge University Press.
- Molenaar, I. W. (1995). Some Background for Item Response Theory and the Rasch Model. In: Fischer, G.H. & I.W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments and Applications* (pp. 3-14). New York: Springer.
- Morrow, K. (2004). Background to the CEF. In Morrow, K. (Ed.), *Insights from the Common European Framework* (pp. 3-11). Oxford: Oxford University Press.

- Nachar, M. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology* 4(1), 13-20.
- Nicholas, H. (1985). Individual differences in interlanguage use. *Australian Review of Applied Linguistics* 8(1), 70–86.
- Nicholas, H. & G. Wigglesworth (2003). Second Language Development: Extending, Elaborating and Refining English Speech. In G. Wigglesworth (Ed.), *The Kaleidoscope of Adult Second Language Learning: Learner, Teacher and Researcher Perspectives* (pp. 134-181). Sydney: National Center for English Language Teaching and Research.
- Nicholas, H. & D. Starks (forthc.). Repositioning and Reframing the Hypothesis Space. In A. Lenzing, H. Nicholas & J. Roos (eds). *Widening contexts for Processability Theory: Theories and issues*. Amsterdam: John Benjamins.
- John M. Norris, Lourdes Ortega (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity, *Applied Linguistics* 30(4), 555–578.
- North, B. (1992). Options for Scales of Proficiency for a European Framework. In Schärer & B. North (Eds.), *Towards a Common European Framework for Reporting Language Competency* (pp. 9-27). Washington: NFLC Occasional Papers.
- North, B. (1996). Description and Assessment of Foreign Language Learning Proficiency in the Swiss Educational System. *Bulletin Suisse de linguistique appliquee* 64, 129-143.
- North, B (1997). Perspectives on language proficiency and aspects of competence. *Language Teaching* 30, 92–100.
- North, B. & G. Schneider (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217-263.
- North, B. (2007). The CEFR Common Reference Levels: validated reference points and local strategies. In Goullier, F. (Ed.), *The Common European Framework of Reference for Languages (CEFR) and the development of language policies: challenges and responsibilities*. Strasbourg: Council of Europe.
- North, B. (2014). The CEFR: Common Framework of Reference. In Milanovic, M. & N. Saville (Eds.), *The CEFR in Practice* (pp. 8-44). Cambridge: Cambridge University Press.
- North, B. & J. Panthier (2016). Updating the CEFR descriptors: the context. *Cambridge English Research Notes* (63), 16-23.
- Özdemir, B. (2004). *Language development in Turkish-German bilingual children and implications for English as a third language*. Unpublished MA Thesis.
- Palotti, G. (2007). An operational definition of the emergence criterion. *Applied Linguistics*, 28(3), 361-382.

- Pienemann, M. (1981). *Der Zweitspracherwerb ausländischer Arbeiterkinder*. Bonn: Bouvier.
- Pienemann, M. (1984). Psychological Constraints on the Teachability of Languages. *Studies in Second Language Acquisition* 6, 186–214.
- Pienemann, M. (1985). Learnability and Syllabus Construction. In Hyltenstam, K. & M. Pienemann. (Eds.), *Modelling and Assessing Second Language Acquisition*. Avon: Multilingual Matters.
- Pienemann, M. & M. Johnston (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 89–94). Adelaide: National Curriculum Resource Centre.
- Pienemann, M., Johnston, M. & G. Brindley (1988). Constructing an Acquisition-based Procedure for Second Language Assessment. *Studies in Second Language Acquisition*, 10, 217-243.
- Pienemann, M. (1989). Is Language Teachable? Psycholinguistic Experiments and Hypotheses. *Applied Linguistics* 10, 52–79.
- Pienemann, M. (1990). *LARC Research Projects 1990*. Sydney NLIA/ Language Acquisition Research Center.
- Pienemann, M. & Louise Jansen (1991). Computational analysis of language acquisition data. In: M. Pennington & V. Stevens (Eds). *Computers in Applied Linguistics: An International Perspective*. Clevedon, Avon: Multilingual Matters. p.201-247.
- Pienemann, M. (1992). *Assessing second language acquisition through Rapid Profile*. University of Sydney: Unpublished Manuscript.
- Pienemann, M., & A. Mackey (1993). An empirical study of children's ESL development and rapid profile. In P. McKay (Ed.), *ESL development: Language and literacy in schools* (pp. 115-259). Canberra: Commonwealth of Australia and National Languages and Literacy Institute of Australia.
- Pienemann, M. (1998). *Language processing and second language development. Processability Theory*. Amsterdam: John Benjamins.
- Pienemann, M., & G. Håkansson (1999). A unified approach towards the development of Swedish as L2: A processability account. *Studies in Second Language Acquisition*, 21(3), 383-420.
- Pienemann, M. (2002). Issues in second language acquisition and language processing. *Second Language Research*, Vol. 18, No 3, pp. 190-192.
- Pienemann, M. (2005a). (Ed.), *Cross-Linguistic Aspects of Processability Theory*. Amsterdam/New York.
- Pienemann, M. (2005b). An Introduction to Processability Theory. In Pienemann, M. (Ed.), *Cross-Linguistic Aspects of Processability Theory*. Amsterdam/New York, 1-60.

- Pienemann, M. (2005c). Discussing PT. In Pienemann, M. (Ed.), *Cross-Linguistic Aspects of Processability Theory*. Amsterdam/New York, 61-84.
- Pienemann, M.; Di Biase & S. Kawaguchi (2005). Processability, Typological Distance and L1 Transfer. In Pienemann, M. (Ed.), *Cross-Linguistic Aspects of Processability Theory*. Amsterdam/New York, 85-117.
- Pienemann, M., Keßler, J.-U., & Roos, E. (Eds.). (2006). *Englischerwerb in der Grundschule. Ein Studien- und Arbeitsbuch*. Paderborn: Schöningh/UTB.
- Pienemann, M. (2007). An Introduction to Processability Theory. In Mansouri, F. (Ed.), *Second Language Acquisition Research – Theory-Construction and Testing* (pp. 13-37). Newcastle upon Tyne: Cambridge.
- Pienemann, M. & J.-U. Keßler (2007). Measuring Bilingualism. In P. Auer & Li Wei (Eds.), *Handbook of Applied Linguistics*, Vol. 5: Multilingualism. Berlin/New York: deGruyter.
- Pienemann, M. (2008). A brief introduction to Processability Theory. In: J.-U. Keßler (Ed.), *Processability approaches to second language development and second language learning* (pp.9-29). Newcastle, UK: Cambridge Scholars Publishing.
- Pienemann, M., & J.-U. Keßler (Eds.). (2011). *Studying Processability Theory: Introductory Textbook*. Amsterdam: John Benjamins.
- Pienemann, M. & A. Lenzing (2015). Processability Theory. In: B. VanPatten and J. Williams (Eds.), *Theories in Second Language Acquisition. An Introduction* (pp. 159-179). Second Edition. Routledge: New York.
- Pienemann, M. & F. Lanze (2017). *Automatic Profiling Expert System APES*. Paper presented 16th International Symposium of Processability Approaches to Language Acquisition (PALA), Pädagogische Hochschule Ludwigsburg, Germany, 04.-05.09.2017.
- Pollitt, A., & N. L. Murray (1993). What raters really pay attention to? In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74-91). Cambridge: Cambridge University Press.
- Porcher, L. (1980). *Reflections of Language Needs in Schools*. Straßbourg: Council of Europe.
- Preston, J. (2009). *Examining the reliability of processability theory-based procedure for use in Japanese SLA assessment*. PDF retrieved from: https://s3.amazonaws.com/academia.edu.documents/33034815/processability_procedural_grammar.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1532187381&Signature=IRUateK9nOKjcptO33rJTekKknQ%3D&response-contentdisposition=inline%3B%20filename%3DExamining_the_reliability_of_processabil.pdf. Last Access: 21.07.2018.

- Prodeau, M. (2009). Du nom au syntagme nominal complexe. In D. Veronique; C. Carlo, C. Granget, J. Kim & M. Prodeau (Eds.), *Acquisition de la grammaire du français langue étrangère* (pp. 75-114). Paris: Didier.
- Prodeau, M; Lopez, S. & D. Veronique (2012). Acquisition of French as a Second Language: Do developmental stages correlate with CEFR levels? *Apples – Journal of Applied Linguistics*, 6(1), 47-68.
- Qiriaz, V. & B. North (in prep). *The CEFR Illustrative Scales – the proposed extended version*. Straßbourg: Council of Europe.
- Rasch, G. (1992). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.
- Richterich, R. & J.J. Chancerel. (1978). *Identifying the needs of adults learning a foreign language*. Oxford: Prentice Hall.
- Richterich, R. (Ed.). (1983). *Case Studies in Identifying Language Needs*. Oxford: Pergamon.
- Roos, J. (2007). *Spracherwerb & Sprachproduktion: Lernziele und Lernergebnisse im Englischunterricht der Grundschule*. Giessener Beiträge zur Fremdsprachendidaktik. Tübingen: Gunter Narr Verlag.
- Roos, Jana (2014). CLIL: Approaching Content Through Communicative Interaction. In: Kupetz, Rita & Becker, Carmen (Eds). *CLIL by Interaction*. Frankfurt am Main: Peter Lang, 101-119.
- Roos, J. (2016). Acquisition as a Gradual Process. Second Language Development in the EFL Classroom. In Keßler, J.-U.; Lenzing, A. & M. Liebner (Eds.). *Developing, Modelling and Assessing Second Languages* (pp. 122-134). Amsterdam: John Benjamins.
- Salamoura, A. & N. Saville (2010). Exemplifying the CEFR: Criterial Features of Written Learner English from the English Profile Programme. In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative Proficiency and Linguistic Development* (pp 101-132). EUROSLA Monograph Series 1.
- Schärer, R. & B. North (1992) (Eds.), *Towards a Common European Framework for Reporting Language Competency*. Washington: National Foreign Language Center Occasional Papers.
- Schärer, R. (1992). A European Language Portfolio. In Schärer, R. & B. North (Eds.), *Towards a Common European Framework for Reporting Language Competency* (pp.1-3). Washington: National Foreign Language Center Occasional Papers.
- Schwartz, B. (1974). Rapport sur la visite à de CET experiments 20-24. *CCCEP* 75(6).
- Selinker, L. (1972). *Interlanguage*. *International Review of Applied Linguistics in Language Teaching*, 10(3).

- Sharwood Smith M. A. (1993). Input Enhancement in Instructed SLA Theoretical Bases. *Studies in Second Language Acquisition*, 15, 165-179.
- Spada, N., & Lightbown, P. M. (2008). Form-focused instruction: Isolated or integrated? *TESOL Quarterly* 42(2), 181-207.
- Steininger, I. (2015). *Modellierung literarischer Kompetenz*. PhD Thesis.
- Takala, S. (2010). Relating Examinations to the Common European Framework. In: Beck, B. & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messungen. DESI Studie* (pp. 306-313). Weinheim/Basel: Beltz.
- Tarone, E. (1983). On the Variability of Interlanguage Systems. *Applied Linguistics* 4, 142-163.
- Thewissen, J. (2013). Capturing L2 Accuracy Developmental Patterns: Insights from an Error-Tagged EFL Learner Corpus. *The Modern Language Journal*, 97(1), 77-101.
- Trim, J. L. M. (1978). *Some possible lines of development of an overall structure for a European unit/credit scheme for foreign language learning by adults*. Strasbourg: Council of Europe.
- Trim, J.L.M; Richterich, R.; van Ek, J.A. & D.A. Wilkins (1980). *Systems Development in Adult Language Learning*. Oxford: Pergamon.
- Trim, J.L.M. (2007). *MODERN LANGUAGES IN THE COUNCIL OF EUROPE 1954-1997. International co-operation in support of lifelong language learning for effective communication, mutual cultural enrichment and democratic citizenship in Europe*. Straßbourg: Council of Europe.
- Trim, J.L.M. (2010). Some earlier developments in the description of levels of language proficiency. In Green, A. (Ed.), *Language Functions Revisited. Theoretical and empirical bases for language construct definitions across ability range* (pp.:xxi-xli). Cambridge: UCLES/CUP.
- Trim, J. L. M. (2012). The Common European Framework of Reference for Languages and its background: a case study of cultural politics and educational influences. In Byram, M. & L. Parmenter (Eds.), *The Common European Framework of Reference: Globalisation of Policy* (pp. 14-34). Bristol: Multilingual Matters.
- Van den Branden, K. (2006). (Ed.), *Task-based Language Teaching. From Theory to Practice*. Cambridge: Cambridge University Press.
- Van Ek, J.A. (1975). *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.
- Van Ek, J.A. & J. Trim (1998). *Threshold 1990* (revised and corrected edition), Cambridge: Cambridge University Press.

- VanPatten B. (Ed.). (2004). Second language acquisition research. Processing instruction: Theory, research, and commentary. Mahwah, NJ, US: Lawrence Erlbaum.
- Veronique, D.; Carlo, C.; Granget, C.; Kim, J.; & M. Prodeau (2009). *Acquisition de la grammaire du français langue étrangère*. Paris: Didier.
- Weinert, F.E. (2001). Concept of Competence: A Conceptual Clarification. In D.S. Rychen & L.H. Salganik (Eds.), *Defining and Selecting Key Competencies* (pp. 45-65). Seattle, WA: Hogrefe & Huber.
- Weir, C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281-300.
- Weir, C.; Hawkey, R.; Green, T. & S. Devy (2009). The cognitive processes underlying the academic reading construct as measured by IELTS. *British Council/IDP Australia IELTS Research Reports*, 9(4), 157-189.
- Westhoff, G. (2007). Challenges and Opportunities of the CEFR for Reimagining Foreign Language Pedagogy. *The Modern Language Journal* (91).
- White, L. (1991). Second Language Competence versus Second Language Performance: UG or Processing Strategies? In L. Eubank (Ed.), *Point – Counterpoint. Universal Grammar in the Second Language* (pp. 167-189). Amsterdam: John Benjamins.
- Wiesniewski, K. (2017a). Empirical Learner Language and the Common European Framework of Reference. *Language Learning*, 67(S1), 232-253.
- Wiesniewski, K. (2017b). The Empirical Validity of the Common European Framework of Reference Scales. An Exemplary Study for the Vocabulary and Fluency Scales in a Language Testing Context. *Applied Linguistics*, 1-28. DOI: 10.1093/applin/amw057,
- Wilkins, D.A. (1973). *An Investigation into Linguistic and Situational Content of the Common Core in a Unit-Credit Scheme*. Strasbourg: Council of Europe.
- Wilkins, D.A. (1976). *Notional syllabuses*. Oxford, Oxford University Press.
- Williams, C. (2007). *A Preliminary Study into the Verbal Subcategorization Frame: Usage in the CLC*. Unpublished Manuscript, RCEAL, Cambridge University, UK.
- Willis, J. (1996). *A Framework for Task-Based Learning*. Harlow: Longman.
- Winke, P. and Gass, S. (2013). The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. *TESOL Quarterly*, 47, 762–789.
- Wode, H. (1976). Developmental sequences in naturalistic L2 acquisition. *Working Papers on Bilingualism* 11, 1-13.
- Zhang, Y. (2005). Processing and formal instruction in the L2 acquisition of five Chinese grammatical morphemes. In M. Pienemann (Eds.), *Cross-linguistic aspects of Processability Theory* (pp.155-177) Amsterdam: John Benjamins.

Zhang, Y. (2007). Testing the topic hypothesis: The L2 acquisition of Chinese syntax. In F. Mansouri, (Ed.), *Second language acquisition research. Theory construction and testing* (pp.137-163). Newcastle, UK: Cambridge Scholars Press.

Zobl, H. (1986). Word order typology, lexical government, and the prediction of multiple, graded effects in L2 word order. *Language Learning* 36, 159-183.

Websites

CoE (2017, February 14) *Education and Languages, Language Policy*. Retrieved from: http://www.coe.int/t/dg4/linguistic/historique_en.asp

CoE (2017, February 14). *Language Policy Unit: Events*. Retrieved from: http://www.coe.int/t/dg4/linguistic/conference_bis_en.asp#P40_1517

CoE (2018, July 07). *General Overview*. Retrieved from: www.coe.int.

CoE (2018, July 07). *History of the CEFR*. Retrieved from: http://www.coe.int/t/dg4/linguistic/historique_en.asp.

CoE (2018, July 07). CoE's Lifelong Learning Initiatives. Retrieved from: <http://pjp-eu.coe.int>.

CoE (2018, July 07). *CoE's resolutions and recommendation*. Retrieved from: http://www.coe.int/t/dg4/linguistic/20thsessioncracow2000_EN.asp#TopOfPage and http://www.coe.int/t/dg4/linguistic/Conventions_EN.asp#TopOfPage.

7. Appendices

This appendix consists of two parts. The first part includes exemplary transcriptions of the original and edited sound files. The second part encompasses the notes of one researcher who participated in the pilot study on the perception of grammatical inaccuracy.

7.1 Exemplary Transcriptions

I present two exemplary original and edited transcriptions here. In each case, the first transcription is the original one and the second transcription is the corresponding edited one.

Ko01-original
P= Participant, I= Interviewer

Instructions task 1

P: ehm there I can see the family Simpson of a tv eh show and yes it's very famous and there you can see that the family is swimming and eh normally Homer the dad from the family is drinking beer and so he now drinks beer also and ehm the boy Bart may eh {jaa} eh jump into the pool /

I: mhm /

P: and {joa} they look like very fun /

I: mhm what can you say about these two /

P: ehm the girl Lisa is eh (#) {eincremen}/

I: eh puts sunscreen /

P: eh {ja} puts sunscreen on her body /

I: mhm /

P: and eh the mother Marge and the girl Maggie is (/) eh are swimming /

I: mhm /

P: and Marge eh help Maggie because eh she is very young /

I: mhm ok /

Instructions task 1 picture two

P: ehm there I can see that the family is eh driving a bike /

I: mhm /

P: behind there and ehm Maggie sit in front of the bike /

I: mhm /

P: in a special seat /

I: mhm /

P: and ehm is holding an ice cream for Homer /

I: mhm /

P: and Lisa is too eh small and eh cannot (#) {treten} /

I: oh reach the paddles /

P: eh reach the paddles /

I: mhm /

P: and eh have some bubbles /

I: mhm /

P: and Marge are also driving /

I: mhm /

P: eh the bike /

I: {jupp} that's great thank you /

Instructions task two

P: eh at six o'clock a.m. the eh bell ring and he must stand up /

I: mhm /

P: he eh looks tired /

I: mhm /

P: and at eh half past seven a.m. /

I: mhm /

P: he sit on the table and eat breakfast /

I: mhm

P: then he go to eh to eh wash his teeth and go (*dusching*) (/) eh to dusch /
 I: take a shower /
 P: take a shower /
 I: mhm /
 P: ehm at seven eh thirty he eh wear his jacket and go to the bus station /
 I: mhm /
 P: and the bus arrives at eight o' clock /
 I: mhm /
 P: eh then he is eh eight eh eh ja eight thirty at the office that's his job and he sit on the table /
 I: mhm /
 P: at the {Büro} eh {Büro}/
 I: office mhm =
 P: = {ja} his office at nine a.m. eh until four p.m. eh he works /
 I: mhm /
 P: and eh then I th (/) eh then he calls someone /
 I: mhm /
 P: and I think at eh five p.m. he is at home /
 I: mhm /
 P: eh so it look like and he's sitting on the couch /
 I: mhm /
 P: and ring someone /
 I: mhm /
 P: and at six p.m. he is sitting for the computer and/ =
 I: = mhm /
 P: play hockey or /
 I: ice hockey right / =
 P: = ja /
 I: thank you /

Instruction task 3

P: ehm what color are the scarf from the girl in the picture /
 I: it's orange /
 P: ok he is purple ehm the hair is red or is it (#)
 I: it's red too /
 P: ok /
 I: mhm /
 P: ehm (#) on the (*back*) from the girl eh is there eh a text or something /
 I: no that is missing in my picture too /
 P: ok ehm in the background on the house is there something (#) like eh {also Kermin}/
 I: mhm that is in my picture too /
 P: ok /
 I: but you can ask me about the color of my roofs /
 P: eh what it means roofs /
 I: ehm {Dach} /
 P: {achso} ehm the roofs are in my picture red and (#) in your {also} or /
 I: they are grey in my picture / =
 P: = ok /
 P: ehm the girl have two eh three {was heisst Sommersprossen oder}/
 I: I'm not sure right now ehm {ja} she has two in my picture right /

P: ok eehm are there two (#) {Blätter}/
 I: leaves /
 P: eh leaves on the picture or are they missing /
 I: one is missing in my picture /
 P: ok ehm in the background in my picture are two trees /
 I: mhm /
 P: ehm in your {also} or /
 I: there's one tree missing in my picture / =
 P: =ok /
 P: ehm is it eh {wichtig oder}/
 I: yes yes /
 P: ok in my picture stands Australia and sixteen c eh you also or not /
 I: one of the two is missing in my picture so ask which one /
 P: ehm I think the eh sixty t /
 I: mh try to make a question out of that /
 P: are they the number and the letter in your picture /
 I: that is missing in my picture /
 P: ok /
 I: see it looks slightly different /

Instructions task four

P: ehm Is it a boy or is it a girl /
 I: it's a boy /
 P: em it's a real person or not /
 I: it's a comic person /
 P: ok ehm is it a older comic person or is it new /
 I: it's a little older {ja}/
 P: kay ehmm is it eh a famous /
 I: yes /
 P: mmh (##) is it eh a comic (*figur*) that plays in real life or is it {nen biss} (/) a little bit fantasy /
 I: it's fantasy it's fantasy / =
 P: = mhkay /
 P: ehm (#) is it yellow /
 I: it is /
 P: mmh is it Spongebob /
 ((picture was shown to her))
 I: yay / ((giggling)) =
 P: = ((giggling))
 I: ok good job that's actually all I needed ehm {Super Danke nochmal} /

Ke01 – edited version
P= Participant, I= Interviewer

Instructions task 1

P: ehm there I can see the family Simpson of a tv eh show and yes it's very famous and there you can see that the family is swim and eh normally Homer the dad from the family is drink beer and so he now drink beer also and ehm the boy Bart may eh {jaa} eh jump into the pool /
 I: mhm /

P: and {joa} they look like very fun /
I: mhm what can you say about these two /
P:ehm the girl Lisa is eh (#) {eincremen}/
I: eh puts sunscreen /
P: eh {ja} put sunscreen on her body /
I: mhm /
P: and eh the mother Marge and the girl Maggie is (/) eh are swim /
I: mhm /
P: and Marge eh help Maggie because eh she is very young /
I: mhm ok /

Instructions task 1 picture two

P: ehm there I can see that the family is eh drive a bike /
I: mhm /
P: it's for three person and eh Bart is hold eh on their skateboard /
I: mhm /
P: behind there and ehm Maggie sit in front of the bike /
I: mhm /
P: in a special seat /
I: mhm /
P: and ehm is hold an ice cream for Homer /
I: mhm /
P: and Lisa is too eh small and eh cannot (#) {treten} /
I: oh reach the paddles /
P: eh reach the paddles /
I: mhm /
P: and eh have some bubble /
I: mhm /
P: and Marge are also drive /
I: mhm /
P: eh the bike /
I: {jupp} that's great thank you /

Instructions task two

P: eh at six o'clock a.m. the eh bell ring and he must stand up /
I: mhm /
P: he eh look tired /
I: mhm /
P: and at eh half past seven a.m. /
I: mhm /
P: he sit on the table and eat breakfast /
I: mhm /
P: then he go to eh to eh wash his teeth and go (*dusching*) (/) eh to dusch /
I: take a shower /
P: take a shower /
I: mhm /
P: ehm at seven eh thirty he eh wear his jacket and go to the bus station /
I: mhm /
P: and the bus arrive at eight o' clock /

I: mhm /
P: eh then he is eh eight eh eh ja eight thirty at the office that's his job and he sit on the table /
I: mhm /
P: at the {Büro} eh {Büro}/
I: office mhm =
P: = {ja} his office at nine a.m. eh until four p.m. eh he work /
I: mhm /
P: and eh then I th (/) eh then he call someone /
I: mhm /
P: and I think at eh five p.m. he is at home /
I: mhm /
P: eh so it look like and he's sit on the couch /
I: mhm /
P: and ring someone /
I: mhm /
P: and at six p.m. he is sit for the computer and/ =
I: = mhm /
P: play hockey or /
I: ice hockey right / =
P: = ja /
I: thank you /

Instruction task 3

P: ehm what color are the scarf from the girl in the picture /
I: it's orange /
P: ok he is purple ehm the hair is red or is it (#)
I: it's red too /
P: ok /
I: mhm /
P: ehm (#) on the (*back*) from the girl eh is there eh a text or something /
I: no that is missing in my picture too /
P: ok ehm in the background on the house is there something (#) like eh {also Kermin}/
I: mhm that is in my picture too /
P: ok /
I: but you can ask me about the color of my roofs /
P: eh what it means roofs /
I: ehm {Dach} /
P: {achso} ehm the roof are in my picture red and (#) in your {also} or /
I: they are grey in my picture / =
P: = ok /
P: ehm the girl have two eh three {was heisst Sommersprossen oder}/
I: I'm not sure right now ehm {ja} she has two in my picture right /
P: ok eehm are there two (#) {Blätter}/
I: leaves /
P: eh leave on the picture or are they miss /
I: one is missing in my picture /
P: ok ehm in the background in my picture are two tree /
I: mhm /
P: ehm in your {also} or /

I: there's one tree missing in my picture / =
P: =ok /
P: ehm is it eh {wichtig oder}/
I: yes yes /
P: ok in my picture stand Australia and sixteen c eh you also or not /
I: one of the two is missing in my picture so ask which one /
P: ehm I think the eh sixty t /
I: mh try to make a question out of that /
P: are they the number and the letter in your picture /
I: that is missing in my picture /
P: ok /
I: see it looks slightly different /

Instructions task four

P: ehm Is it a boy or is it a girl /
I: it's a boy /
P: em it's a real person or not /
I: it's a comic person /
P: ok ehm is it a older comic person or is it new /
I: it's a little older {ja}/
P: kay ehmm is it eh a famous /
I: yes /
P: mmh (##) is it eh a comic (*figur*) that play in real life or is it {nen biss} (/) a little bit fantasy /
I: it's fantasy it's fantasy / =
P: = mhkay /
P: ehm (#) is it yellow /
I: it is /
P: mmh is it Spongebob /
((picture was shown to her))
I: yay / ((giggling)) =
P: = ((giggling))
I: ok good job that's actually all I needed ehm {Super Danke nochmal} /

Ko02 – original version
C= interviewer, A=informant

C OK/ so er the first thing we'll do this morning is look at some pictures
A mmm
C and I'm going to ask you to tell me a story.. about the pictures/ here we have ah some pictures from a store..with=
A =a store ?
C a shopkeeper/
A oh
C and we have some things that he does..everyday/ and I'd like you to tell me the story of what he does.. in a day/
A [gap] first hes= he clean er her shop his shop er before open...mmm..and then he look (her) goods or things

C mhm

A in the...er book= book= er in the shopcase/

C mhm

A and he checks the price.. of their= of his goods

C yes

A and then he..he wants to be a cashier and the customers pay..er..he= he= her bought the something

C mhm

A and then erm..the lady.. show the= what= what she bought to the cashier and then maybe ask= ask something he wants looking for =

C =OK

A = and the shopkeeper erm point= point her to the..what he= she looking for

C good/ alright I have one more story for you

A yes

C this time we'll be in the library

A (yeah)

C this is the University of Sydney library/ and this is the librarian

A oh/ first er..this maybe students come to the librarian and he..add his name to the card= card (librarian) and..she er.. look about the books in the librarian..and put the books in the bookcase..erm...he..er..looking the books what's er books er the people borrow from librarian

C mhm

A and then her lady=..this lady..er..ask something about the.. er books she looking for..and he look er= in the..computer..about er the books (ha)= in the librarian/ he check=

C =mhm

A check in/ and then er this man er.. ask the lady about the information about..in= in this librarian= library/ maybe he= he don't know about this library

C good/ you're a good storyteller/

LARC Track 2

C next we're going to tell some stories./ I'll show you a picture, and I want you to try to figure out what happened/so maybe you'd like to take a minute to look at the pictures and then you can ask me some questions/ [gap] OK?/

A (is) he a businessman?/

C yes he is/

A mmm..[gap] only yes ans= yes-no answer or no =

C = er I can tell you other things as well=

A = oh yeah

C you can ask me anything/

A oh yeah/ mmm [gap] is he from..= where is he?/

C he's in the hospital/

A hospital?/

C yeah/
 A aar..in this room the patient?/
 C yes/
 A he wants look his wife?/
 C yes/ good guess/ it's his wife=
 A = and his wife=
 C mhm/
 A he wife= his wife born a baby?/
 C yes/
 A err [gap] er and then..where= where is it= where is he going?/
 C oh/ I don't know where he's going/
 A mmm.. from hospital?/
 C yes/ from the same room/
 A oh from the same room/
 C mhm/ this time
 A mmm
 C he's not very happy/
 A yeah/ maybe..er his baby die?/
 C no/ the baby's OK
 A oh./ how about his wife?/
 C she's ok too=
 A she's ok too?/
 C mmm/
 A [gap] why (is) not happy?/
 C something happend in the room/
 A something happened?/
 C yeah/
 A mmm [gap] he= he not enough to see his wife?/
 C maybe maybe/ but em I'll give you a hint/ he wanted to call
 the baby Tom/
 A Tom?/
 C Tom, it's a name=
 A =oh
 C it's an English name
 A Oh... so...yeah, yeah, I know/ he want er give her= his baby
 name Tom but her wife er disagree with h..him/ so maybe quarrel=
 C =mhm
 A =in this room/ so...er...he..= he not happy./
 C that's right yes/ she wanted to call the baby Mike
 A (oh Mike)
 C and the baby is called Mike/
 A oh
 C [laugh] so she won the argument/
 A (oh yeah)
 C OK, good/ now I have another story which is just a little
 harder..because there are more pictures/ so I'll give you a minute
 to take a look at the pictures=
 A =mhm
 C =and then you can ask me any questions you like...to find out

what happen in the story/
A [gap] what is she= what is he doing?/
C oh, he's writing down a message from a telephone call/
A oh/
C he's writing down...three million dollars/
A three million dollars/
C mmm/
A is he er operator?/
C No./ It's just his telephone./
A is he go to the doctor?/
C yes/
A [gap] he has got a headache?/
C kind of a headache, he feels ill =mhm/
A =yeah (oh)...he... has got a message 'bout er...his money...three million?/
C right, he= the person on the the phone says you have to pay...three million dollars/
A oh...yeah, yeah/ the person call him...he must pay three million dollars so he..s= or su=..surprised.. and maybe worried... and then er...he go to the= his friends/ maybe to borrow...money from...his friend?/
C No, this is the man who wants the money./
A Oh/ yeah/ and he= he didn't have a lot of money so he cannot pay and the man..er come to...= to him to ask his money...and...and then she...open er his briefcase but she don't have an= a lot of money?/
C no, he has enough money/
A enough?/
C yes/
A oh/(gap) in this...briefcase?/
C right [three million] dollars./
A [this money]/ three million dollars!/
C mmm
A [gap] yeah mmm...yeah, I know/ he= he is a...drug...er...= if somebody er use the drug he feels sick and then she don't have money/ so he bo..= borrow from somebody else...er to buy a drug/ and...er..he..= he didn't know he used too= too much er drug/ so she spent lot of money/
C mhm/
A and then..mmm...one day somebody call..hi= call him...to ask his money/ about three millions/ and...she...er...= this man...er...come to him to ask his money and he give hi= his money to this man to pay er...he...from= to pay his money/ and then...er...the problem...er...clear and he shake hand/
after that he go to the doctor...to want to be the (health/help?) and maybe...she don't want use the drug [again]/
C [mhm], good/ very good/that's a hard one isn't it!/

Ke02 – edited version
C= Interviewer, A= Informant

- C OK/ so er the first thing we'll do this morning is look at some pictures
- A mmm
- C and I'm going to ask you to tell me a story.. about the pictures/ here we have ah some pictures from a store..with=
=a store ?
- A a shopkeeper/
- A oh
- C and we have some things that he does..everyday/ and I'd like you to tell me the story of what he does.. in a day/
- A [gap] first he= he clean er her shop his shop er before open...mmm..and then he look (her) good or thing
- C mhm
- A in the...er book= book= er in the shopcase/
- C mhm
- A and he check the price.. of their= of his good
- C yes
- A and then he..he want to be a cashier and the customer pay..er..he= he= her bought the something
- C mhm
- A and then erm..the lady.. show the= what= what she bought to the cashier and then maybe ask= ask something he want looking for =
- C =OK
- A = and the shopkeeper erm point= point her to the..what he= she look for
- C good/ alright I have one more story for you
- A yes
- C this time we'll be in the library
- A (yeah)
- C this is the University of Sydney library/ and this is the librarian
- A oh/ first er..this maybe student come to the librarian and he..add his name to the card= card (librarian) and..she er.. look about the book in the librarian..and put the book in the bookcase..erm...he..er..look the book what's er book er the people borrow from librarian
- C mhm
- A and then her lady=..this lady..er..ask something about the.. er book she look for..and he look er= in the..computer..about er the book (ha)= in the librarian/ he check=
- C =mhm
- A check in/ and then er this man er.. ask the lady about the information about..in= in this librarian= library/ maybe he= he don't know about this library
- C good/ you're a good storyteller/

C next we're going to tell some stories./ I'll show you a picture,
and I want you to try to figure out what happened/so maybe you'd
like to take a minute to look at the pictures and then you can ask
me some questions/ [gap] OK?/

A (is) he a businessman?/

C yes he is/

A mmm..[gap] only yes ans= yes-no answer or no =

C = er I can tell you other thing as well=

A = oh yeah

C you can ask me anything/

A oh yeah/ mmm [gap] is he from..= where is he?/

C he's in the hospital/

A hospital?/

C yeah/

A aar..in this room the patient?/

C yes/

A he want look his wife?/

C yes/ good guess/ it's his wife=

A = and his wife=

C mhm/

A he wife= his wife born a baby?/

C yes/

A err [gap] er and then..where= where is it= where is he going?/

C oh/ I don't know where he's going/

A mmm.. from hospital?/

C yes/ from the same room/

A oh from the same room/

C mhm/ this time

A mmm

C he's not very happy/

A yeah/ maybe..er his baby die?/

C no/ the baby's OK

A oh./ how about his wife?/

C she's ok too=

A she's ok too?/

C mmm/

A [gap] why (is) not happy?/

C something happend in the room/

A something happened?/

C yeah/

A mmm [gap] he= he not enough to see his wife?/

C maybe maybe/ but em I'll give you a hint/ he wanted to call
the baby Tom/

A Tom?/

C Tom, it's a name=

A =oh

C it's an English name

A Oh... so...yeah, yeah, I know/ he want er give her= his baby
name Tom but her wife er disagree with h..him/ so maybe quarrel=

C =mhm

A =in this room/ so...er...he..= he not happy./

C that's right yes/ she wanted to call the baby Mike

A (oh Mike)

C and the baby is called Mike/

A oh

C [laugh] so she won the argument/

A (oh yeah)

C OK, good/ now I have another story which is just a little harder..because there are more pictures/ so I'll give you a minute to take a look at the pictures=

A =mhm

C =and then you can ask me any questions you like...to find out what happend in the story/

A [gap] what is she= what is he doing?/

C oh, he's writing down a message from a telephone call/

A oh/

C he's writing down...three million dollars/

A three million dollar/

C mmm/

A is he er operator?/

C No./ It's just his telephone./

A is he go to the doctor?/

C yes/

A [gap] he has got a headache?/

C kind of a headache, he feels ill =mhm/

A =yeah (oh)...he... has got a message 'bout er...his money...three million?/

C right, he= the person on the the phone says you have to pay...three million dollars/

A oh...yeah, yeah/ the person call him...he must pay three million dollar so he..s= or su=..surprise.. and maybe worry... and then er...he go to the= his friend/ maybe to borrow...money from...his friend?/

C No, this is the man who wants the money./

A Oh/ yeah/ and he= he didn't have a lot of money so he cannot pay and the man..er come to...= to him to ask his money...and...and then she...open er his briefcase but she don't have an= a lot of money?/

C no, he has enough money/

A enough?/

C yes/

A oh/(gap) in this...briefcase?/

C right [three million] dollars./

A [this money]/ three million dollar!/

C mmm

A [gap] yeah mmm...yeah, I know/ he= he is a...drug...er...= if somebody er use the drug he feel sick and then she don't have money/ so he bo.= borrow from somebody else...er to buy a drug/

and...er..he..= he didn't know he use too= too much er drug/ so she spent lot of money/

C mhm/

A and then..mmm...one day somebody call..hi= call him...to ask his money/ about three million/ and...she...er...= this man...er...come to him to ask his money and he give hi= his money to this man to pay er...he...from= to pay his money/ and then...er...the problem...er...clear and he shake hand/ after that he go to the doctor...to want to be the (health/help?) and maybe...she don't want use the drug [again]/

C [mhm], good/ very good/that's a hard one isn't it!/

7.2 Exemplary Protocol Pilot Study

On the following pages pictures of the pdf-file of one researcher, who participated in the pilot study on the perception of grammatical inaccuracy, are presented.

University of Paderborn
English and American Studies Department
Course: Research Seminar
Instructor: Prof. Dr. Pienemann
Winter Term 2013/2014

Julia Schönlau
jschoen@mail.uni-paderborn.de
Lehramt GyGe: Sport, Englisch
Matrikelnummer: 6572313

The Perception of Inaccuracy

Which inaccurate grammatical features do raters attend to? Which of those features are particularly strong in rater perception. Is it 3rd person -sg-s? Morphological features, wrong syntax or lexis?

Methodology:

Think-aloud protocols; 2 samples (Tatjana B2; Trial inaccurate)

Participants:

6 raters/students of linguistics/English language teacher trainees

Results:

1st student of linguistics: student of English and German, 9th semester

• Sample 1: Tatjana B2

- 1 ○ 1:35 → "upstays" stimmt nicht. Das ist ein falsches Wort. |
- 7 ○ 2:20 → "Ja, das Wort fehlt hier" [having breakfast]
- 2:28 → "an hour", nicht "a hour"
- 2:42 → "coming time - das Wort fehlt wieder"
- 5 ○ 3:15 → "die Satzstellung stimmt nicht ganz"
- 3:25 → "th" ist falsch ausgesprochen"
- 6:36 → "Achso, sie meint "work"! Okay!
- 7: 45 → "manchmal hat sie Schwierigkeiten mit der Uhrzeit!
- 8:26 → "ja, das Wort fehlt hier wieder" [aufräumen]
- 10 ○ 9:35 → "sie spricht die Wörter manchmal seltsam aus -bei awake z.B."
- 10:08 → "*didn't spoke* ist falsch - falsche Form"
- 12:13 → "genau, da fehlt wieder das Wort für Mülleimer"
- 12:29 → "das ist die falsche Satzstellung mit *also*"
- 14:19 → "da ist wieder ein Fehler mit *also*"
- 15 ○ 15:30 → "da fehlt ihr wieder das Wort"
- 15:53 → "*difficult* muss das eigentlich heißen"
- 17:52 → "ja, sie benutzt das "th" irgendwie falsch"
- 18:16 → "Einwohner- da fehlt ihr wieder das Wort"
- 19:05 → "more long time"
- 20 ○ 19:12 → "th" wieder

- o 19:44 → "hier wieder *more smaller*"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

th →
words missing

- o "Ein paar Sachen, die sie falsch macht sind auf jeden Fall, dass ihr ganz oft irgendwelche Wörter fehlen. Also sie muss lange nach Wörter suchen. Das *th* spricht sie häufig falsch aus – gerade wenn es irgendwie am Ende oder in der Mitte von einem Wort ist. Und die Steigerungsformen – sie sagt immer sowas wie *more better* oder *more longer*."

• Sample 2: Trial inaccurate


- o 0:29 → "think ist falsch ausgesprochen – also das *th*"
- o 1:14 → "think wieder"
- o 1:18 → "he sleep – das "s" fehlt"
- o 1:30 → er macht die ganzen 3rd person singular s Wörter falsch
- o 1:44 → "th und das s wieder"
- o 1:48 → "are – also das wäre eigentlich *is*"
- o 2:03 → "look up in the air – ohne 3rd person singular s"
- o 2:52 → "eat something. Third person"
- o 2:59 → "it look like... ohne s"
- o 3:13 → "da fehlt ihm das Wort" (Riesenrad)
- o 3:50 → "looks"
- o 4:00 → "das Wort fehlt ihm auch"
- o 4:18 → "Kino und Tochter – das kann man auch so nicht sagen"
- o 4:27 → "thinks"
- o 4:39 → "angetan – ihm fehlen oft die Wörter"
- o 4:59 → "he is drunk"
- o 6:10 → da macht er die Frage falsch → *Why does the boy have...* wäre richtig"
- o 6:40 → "das *th* wieder in *that*"
- o 9:16 → "falscher Satzbau bei der Frage. "Where have you seen a skateboard for the first time"
- o 9:42 → "Why are you invent. Falsche Zeit- did you"
- o 9:56 → "interested in wäre richtig -> die Präposition ist falsch"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

3rd - ps - s
words missing
th

- o „Bei ihm würde ich sagen ist das, was er am meisten falsch macht, das *third person singular 's'*. Das benutzt er eigentlich bei fast keinem Verb, wo es nötig ist. Außerdem spricht er auch das *th* falsch aus – das hat jetzt nichts mit Grammatik zu tun, aber es ist trotzdem nicht richtig. Und die Fragen kann er nicht richtig bilden. Er sagt zum Beispiel das „do“ am Anfang der Frage meist nicht, sondern nur „have you...“. Das sind so die meisten Fehler. Und vielleicht noch, dass er die Wörter auf Englisch oft nicht weiß und dann auf Deutsch sagt.“

2nd student of linguistics: BA student of English and history, 9th semester

• **Sample 1: T:  ha B2**

- 2:21 → "Ja, da weiß sie nicht, was *frühstücken* heißt"
- 2:29 → "AN hour later"
- 2:39 → "at his walk?? Das kann irgendwie nicht sein."
- 3:36 → "bed. Die Aussprache ist am Ende zu hard"
- 4:20 → "Ach, "his work". Falsch ausgesprochen"
- 5:20 → "Ja, sie benutzt nicht *gets up*. Sie benutzt eine ganz andere Zeitform - eine mit -ing"
- 6:25 → "She's at work"
- 6:44 → Sie spricht das 'th' falsch aus
- 7:38 → das ist die falsche Uhrzeit
- 8:23 → das Wort *aufräumen* fehlt ihr hier.
- 9:19 → "bed → sie sagt das 'd' am Ende wie ein 't'"
- 10:09 → "*didn't spoke* ist auf jeden Fall nicht richtig - *didn't speak*"
- 10:21 → "*understand* - falsche Betonung"
- 11:34 → "*bicycle* - auch wieder die falsche Betonung"
- 12:08 → "yellow - auch falsch"
- 12:14 → "und ihr fehlt das Wort für *Mülleimer*"
- 13:56 → „heißt das nicht *on the bench*?"
- 14:11 → "*sweater* genau. Sie macht eben oft Betonungsfehler."
- 14:49 → "*them*, nicht *they*"
- 15:33 → "ja genau, da fehlt ihr eben wieder das Wort"
- 16:53 → "What do you usually eat there? Falsche Zeit benutzt."
- 17:50 → "Sie spricht das 'th' irgendwie falsch aus"
- 18:16 → "Genau, da fehlt ihr wieder das Wort für *Einwohner*"
- 19:04 → "*for a longer time*. Die Steigerungsformen sind machmal auch nicht so korrekt"
- 19:39 → "*more smaller* ist wieder eine falsche Steigerungsform"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

- th
- „Ich würde sagen, dass diese Person oft Wörter falsch betont oder falsch ausspricht. Sie macht zum Beispiel oft Fehler in der Aussprache des 'th', wie z.B. in 'with'. Außerdem fehlen ihr manchmal die Wörter in Englisch, sodass die Moderatorin nachhelfen muss. Trotzdem macht die Person insgesamt einen recht guten Eindruck in Englisch.“

• **Sample 2: Trial Inaccurate**

- 0:24 → "th in *think* ist falsch ausgesprochen"
- 0:49 → "th ist wieder falsch"
- 0:52 → "entweder *I see him sleeping* oder *I see that Homer sleeps*"
- 1:09 → "*gets a shower*"

- 1:12 → "sleeps → wieder 3rd person sg s"
- 1:36 → "burns"
- 2:20 → "the oder a chef"
- 2:25 → "sleeps"
- 2:50 → "looks - wieder das s"
- 3:08 → "das fehlt ihm das Wort für *Riesenrad*"
- 3:48 → "is not bored"
- 4:02 → "Das Wort *Kino* fehlt"
- 4:29 → "das Wort *angetan* fehlt"
- 4:44 → "sit - das macht er wirklich immer falsch"
- 5:45 → "Where is the money. Das "is" fehlt bei der Frage"
- 6:46 → "das th"
- 9:08 → "die Satzstellung ist falsch"
- 9:41 → "Why did you invent ... - falsche Zeit"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

3rd ps. s
th

- Diese Person macht die meisten Fehler mit dem 3rd person singular "s". Das "s" benutzt er eigentlich gar nicht. Und dann könnte man noch etwas an seiner Aussprache verbessern, da er z.B. das "th" nicht richtig aussprechen kann. Und die Person hat einige Schwierigkeiten in der Fragebildung

3rd student of linguistics: student of English and PE, 7th semester

• **Sample 1: Tapes B2**

- 2:21 → "she doesn't know the right word"
- 2:28 → "an hour later"
- 2:48 → "coming time is the wrong word I guess"
- 3:17 → "you always think about - she uses the wrong word order"
- 3:31 → "just better not more better"
- 4:18 → "work"
- 4:30 → "and the th was not pronounced correctly"
- 5:21 → "she doesn't use the simple present. The interviewer asks questions in simple past and she answers in present progressive"
- 6:36 → "at her work"
- 6:69 → "she pronounces the th incorrectly"
- 7:34 → "to three"
- 10:07 → "I didn't speak English"
- 11: 35 → "bicycle. Wrong pronunciation"
- 12:12 → "the word is missing"
- 12:27 → "I can also see... wrong word order"
- 15:32 → "the word is missing again"
- 15:53 → "I think she means difficult"
- 16:54 → "what do you usually eat there"
- 17:58 → "the pronunciation of the th is not correct"
- 18:16 → "inhabitants misses"
- 19:06 → "for a longer time"
- 19:44 → "more smaller"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

Word
missig
th

- The most striking thing is that she sometimes doesn't have the English word. She also pronounces several words incorrectly - especially words with th like brother or father. At the end she uses wrong comparative forms.

• **Sample 2: Trial Inaccurate**

- 0:29 → "wrong pronunciation of think"
- 0:52 → "the"
- 0:55 → "I can see him sleeping or I can see that he sleeps"
- 1:09 → "he gets a shower"
- 1:27 → "brushes"
- 1:41 → "the bottle burns. He cannot use the third person singular correctly"
- 1:46 → "he IS very uninterested"
- 2:23 → "his boss. Chef is the wrong word"

- 2:53 → "eats and looks. He doesn't use the third person singular s"
- 3:14 → "the word Ferris wheel is missing"
- 3:33 → "Homer's family not the family of Homer"
- 3:52 → "He is not bored"
- 4:00 → "the"
- 4:04 → "cinema"
- 4:07 → "whole family"
- 4:26 → "thinks - he always pronounces the th incorrectly"
- 4:42 → "angetan- he just uses the German word"
- 4:39 → "sits"
- 4:57 → "he's drunk"
- 7:14 → "the"
- 9:11 → "Where have you seen a skateboard for the first time? - wrong word order"
- 9:42 → "why did you invent so many things - wrong time"
- 9:54 → "interested in - wrong preposition"
- 10:05 → "Do you have a wife? - there's no do-fronting"
- 10:24 → "I don't think so"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

3rd - ps - s
th

- For this person I would say that he cannot use the third person singular s correctly. He also cannot use the do-fronting for questions and he has a very bad pronunciation when it comes to the th. These are the main points.

4th student of linguistics: student of English and Spanish, 7th semester

• Sample 1: Ta a B2

- 2:22 → "having breakfast"
- 2:28 → "an hour"
- 2:35 → "walk?"
- 3:31 → "more better"
- 4:20 → "he left his walk? Was meint sie?"
- 6:35 → "achso, work!!"
- 8:19 → "if we look at"
- 9:39 → "awake ist falsch ausgesprochen"
- 10:08 → "didn't speak"
- 11:33 → "bicycle"
- 12:04 → "yellow"
- 12:14 → "Ja, Mülleimer"
- 12:27 → "I can also see - andersrum"
- 13:56 → "on the bench not on the bank"
- 14:10 → "sweater"
- 14:50 → "do you have them to?"
- 15: 23 → "who are swimming"
- 15:52 → "difficult"
- 17:43 → "...to know something about your society or family"
- 18:17 → "Das Wort inhabitants fehlt ihr"
- 19:05 → "more long time - longer time"
- 19:38 → "more smaller"
- 19:52 → "we can make or eat pizza"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

words
missig

- Ich würde sagen, dass sie häufig die englischen Vokabeln nicht direkt gewusst und musste dann nachfragen. Sie braucht oft sehr lange, um einen ganzen Satz zu bilden. Außerdem betont sie einige Wörter seltsam, was aber auch an ihrer Herkunft liegen könnte.

• Sample 2: Trial Inaccurate

- 0:26 → "I think"
- 0:50 → "in the first picture"
- 0:59 → "sleeping"
- 1:00 → "think- th wieder"
- 1:04 → "in the second picture"
- 1:08 → "getting"
- 1:16 → "he also sleeps"
- 1:31 → "he brushes his teeth"
- 1:48 → "Homer is very uninterested"

- 2:03 → "looks up in the air"
- 2:21 → "chef ist Koch"
- 2:52 → "eats something"
- 2:55 → "it looks like"
- 3:21 → "Wort für Riesenrad fällt ihm nicht ein"
- 3:34 → "and"
- 3:40 → "they"
- 3:46 → "looks"
- 3:53 → "he is not bored"
- 4:13 → "he eats"
- 4:18 → "Kino"
- 4:39 → "angetan"
- 4:48 → "Homer sits"
- 4:56 → "he's drunk"
- 5:52 → "Where is the money?"
- 6:07 → "Why does the boy have"
- 6:19 → "I think"
- 9:13 → "Where have you seen a skateboard for the first time"
- 9:41 → "Why did you invent"
- 9:53 → "interested in"
- 10:05 → "Do you have a wife"
- 10:11 → "Do you have some children"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

3rd ps-s


- Ich glaube, dass es vor allem viele Fehler bei der 3. Person Singular macht, weil er das 's' fast nie benutzt. Er kann außerdem nicht fließend sprechen, da ihm immer wieder englische Vokabeln fehlen und er danach fragen muss. Auch die Fragen kann er nicht fehlerfrei bilden.

5th student of linguistics: student of English and PE, 6th semester

- Sample 1: T...na B2
 - 1:35 → "coffee"
 - 1:53 → "breakfast"
 - 2:23 → "breakfast"
 - 2:34 → "at his walk?"
 - 3:33 → "more better? Better"
 - 4:03 → "five to nine"
 - 4:26 → "he left his walk? WORK! Ah, okay!"
 - 5:23 → "half past 8"
 - 6:06 → "breakfast"
 - 6:27 → "secretary"
 - 6:35 → "work"
 - 8:18 → "look at"
 - 10:08 → "I didn't speak"
 - 11:08 → "kindergarden"
 - 11:34 → "bicycle?"
 - 12:08 → "yellow"
 - 12:29 → "I can also see"
 - 13:02 → "there is one rabbit?"
 - 13:56 → "on the bench"
 - 14:02 → "yellow"
 - 14:07 → "skirt"
 - 14:10 → "sweater - Pullover sagen die nicht"
 - 15:19 → "both of the children"
 - 16:05 → "much more difficult"
 - 16:48 → "would be"
 - 19:01 → "would you like"
 - 19:05 → "for a longer time"
 - 19:38 → "more smaller"
 - 19:52 → "we can eat pizza"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

- Mir ist aufgefallen das sie (asiatischer Hintergrund oder so) die Vokale viel deutlicher ausspricht als wir das machen, also da stechen die Konsonanten viel mehr raus. Sie hat ein paar mal v und w verwechselt bei "we" und so. Sie fragt ganz oft Vokabeln nach - also sie sagt das deutsche Wort und fragt nach dem Englischen. Und sie verwendet ziemlich viele Satzfüller wie „äh“. Sie fängt oft einen Satz an ohne darüber nachzudenken, verbessert sich dann oder „verschlimmbessert“ sich.

- Sample 2:  Inaccurate
 - 0:25 → "think - th"
 - 0:50 → "*in* the first picture"
 - 0:59 → "asleep"
 - 1:00 → "I think"
 - 1:04 → "*in* the second picture"
 - 1:08 → "getting a shower"
 - 1:12 → "getting"
 - 1:15 → "sleeps"
 - 1:23 → "shaving"
 - 1:31 → brushes"
 - 1:38 → " a bottle"
 - 1:43 → "the bottle burns or is on fire?"
 - 1:53 → "*in* the picture"
 - 2:03 → "looks up"
 - 2:21 → "his boss?"
 - 2:27 → "*in* the sixth picture"
 - 2:42 → "*in* the seventh picture"
 - 2:53 → "Marge is eating something"
 - 2:56 → "looks like"
 - 3:10 → "I think"
 - 3:21 → "ferris wheel?"
 - 3:26 → "*in* the next picture"
 - 3:36 → "and"
 - 3:40 → "they are driving"
 - 3:47 → "looks"
 - 3:49 → "looks"
 - 3:53 → "he's not bored"
 - 4:02 → "*in* the next picture"
 - 4:09 → "his *whole* family"
 - 4:13 → "he eats"
 - 4:14 → "daughter"
 - 4:27 → "thinks of some beer"
 - 4:39 → "impressed - ach ja, *fond?*"
 - 4:45 → "*in* the next picture"
 - 4:48 → "sits"
 - 4:56 → "he's drunk"
 - 5:52 → "Where is the money?"
 - 6:07 → "Why does the boy have"
 - 6:19 → "I think"
 - 6:29 → "ice-cream"
 - 6:52 → robs"
 - 7:06 → "I think"
 - 7:15 → "on picture five"

- 7:19 → "destroys it"
- 9:13 → "Where have you seen a skateboard for the first time"
- 9:41 → "Why do you invent so many tricks"
- 9:53 → "interested in"
- 10:05 → "Do you have a wife"
- 10:11 → "Do you have some children"
- 10:24 → "so"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

kurzer Satz
3rd ps -)

- Der hat eine ganz kurze Satzbildung. Der kann definitiv das 3. Person Singular s nicht! Verlaufsform fällt ihm auch eher schwer und die Satzstellung passt irgendwie auch noch nicht. Ja, ziemlich Deutsch und ziemlich wenig Englisch.

6th student of linguistics: student of German, English, PE; 11th semester

• Sample 1: T...na B2

- 2:29 → "an hour - nicht a hour"
- 2:21 → "having breakfast - die Vokabel weiß sie nicht sofort"
- 3:30 → "better"
- 4:20 → "his work!"
- 6:20 → "work"
- 8:01 → "shop"
- 9:00 → "dinner"
- 9:33 → "awake"
- 10:07 → "didn't spoke stimmt nicht - didn't speak"
- 10:14 → "my knowledge not knowledges"
- 11:55 → "on my picture I can see"
- 12:06 → "yellow"
- 12:13 → "Mülleimer"
- 12:28 → "I can also see"
- 14:02 → "yellow"
- 15:30 → "da weiß sie das Wort nicht"
- 16:50 → "The first question *would* be"
- 17:47 → "das 'th' "
- 18:17 → "inhabitants"
- 19:04 → "for a longer time"
- 19:40 → "*smaller than yours*"
- 19:52 → "we can *make* pizza"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

Beachtung

- Meiner Meinung nach fällt bei ihr vor allem auf, dass sie viele Wörter nicht richtig betont, wie z.B. *yellow* oder *work*. Man weiß aber trotzdem meistens was sie meint. Manchmal benutzt sie die falsche Satzstellung, vor allem bei Sätze mit *also*.

• Sample 2: Trial Inaccurate

- 0:25 → "think"
- 1:00 → "I see him sleeping"
- 1:13 → "think"
- 1:15 → "he sleeps"
- 1:39 → "burns"
- 1:48 → "he is very uninterested"
- 2:04 → "he looks up"
- 2:17 → "a chef oder nee - boss"
- 2:53 → "eat something"
- 3:06 → "thinks"
- 3:21 → "deutsches Wort für ferris wheel"

- 3:46 → "looks"
- 3:53 → "he is not bored"
- 4:04 → "Kino - cinema"
- 4:07 → "whole"
- 4:19 → "the"
- 4:25 → "thinks"
- 4:33 → "deutsches Wort angetan"
- 4:48 → "sits"
- 4:44 → "He is drunk"
- 5:52 → "Where is the money?"
- 7:06 → "think"
- 7:17 → "or destroys it?"
- 9:13 → "Where have you seen a skateboard for the first time"
- 9:41 → "Why do you invent"
- 9:56 → "interested in marketing things"
- 10:25 → "I don't think so"

Welche Dinge fallen dir bei diesem Beispiel besonders auf? Wo liegen die meisten Fehler?

3rd ps-s

- Er ist relativ unsicher in Englisch und macht viele Fehler - vor allem bei der dritten Person Singular. Er bildet generell sehr einfache Sätze. Bei der Bildung von Fragen hat er teilweise große Schwierigkeiten und benutzt oft die gleichen Strukturen.

Summary:

Sample 1: Most of the raters say that this person has some problems in finding the correct English terms. They also state that she pronounces several words in the wrong way.

Sample 2: Most of the raters detect several mistakes while hearing the speech sample but when they have to sum up the main mistakes they usually mention the third person singular first, even though there are several other mistakes. Many of them also mention a wrong pronunciation of several words.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbstständig angefertigt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe. Die vorliegende Arbeit wurde bisher weder im In- noch Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere außerdem, dass ich zu keiner Zeit ein Promotionsverfahren an einer anderen Hochschule oder Fakultät beantragt habe.

Paderborn, den
