Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstraße 19
72074 Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

M.A. Thesis in Computational Linguistics

Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects

Zarah Leonie Weiß

zweiss@sfs.uni-tuebingen.de

May 30th, 2017

First Examiner and Supervisor:	Prof. Dr. W. Detmar Meurers
Second Examiner:	Prof. Dr. Harald R. Baayen

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln einschließlich des WWW und anderer elektronischer Quellen angefertigt habe. Alle Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht.

(Zarah Leonie Weiß)

Abstract

This thesis analyses German L2 writing proficiency by means 398 complexity measures from the domains of i) language use; ii) human language processing; iii) discourse and encoding of meaning; and three sub domains of theoretical linguistics, namely iv) syntax; v) lexicon and semantics; and vi) morphology; with an additional focus on task effects on language complexity in learner corpora. To account for these, available task information from the *Merlin* and the *Falko Georgetown* corpus as well as retrospectively annotated cognitive and functional task factors are included in all analyses.

A descriptive cross-corpus investigation of over 100 features shows, how complexity measures exhibit diverging sensitivity to heterogeneous task backgrounds across complexity domains and linguistic constructs. Furthermore, a series of ordinal Generalized Additive Models (GAMs) illustrates, how interactions between cognitive or functional task factors with selected complexity measures significantly improve model fit; and give highly interesting insights into task effects; as well as valuable prompts for further research in more controlled settings.

On *Merlin*, classification experiments all ordinal GAMs reach F1 scores of at least 71% by using only 13 complexity measures. The best classification results are obtained for a model including also performance effects, which outperforms previously reported results with F1 scores of 85%.

KEYNOTES: *L2 Complexity; Generalized Additive Regression Models; Learner Corpora; L2 German; L2 Writing Proficiency; Second Language Acquisition; Task-Effects*

Zusammenfassung

Die vorliegende Masterarbeit erfasst Kompetenz von Deutsch als Zweitsprache durch die Analyse sprachlicher Komplexität unter Berücksichtigung kognitiver und funktionaler Aufgabeneinflüsse. Hierfür werden 398 Merkmale sprachlicher Komplexität aus den Bereichen Sprachnutzung, Diskurs und Bedeutungskodierung, kognitive Sprachverarbeitung, und theoretische Linguistik (Lexikon, Syntax, Morphologie) analysiert. In diese Analysen werden direkt verfügbare sowie nachträglich annotierte Informationen über Aufgaben und Aufgabeneffekte aus dem *Merlin* Korpus und dem *Falko Georgetown* Korpus mit einbezogen.

In einer deskriptiven, korpus-übergreifenden Studie von mehr als 100 Komplexitätsmerkmalen aus allen oben genannten Bereichen, wird zunächst gezeigt, dass Komplexitätsmerkmale aus verschiedenen Bereichen unterschiedlich anfällig für heterogene Aufgabenhintergründe sind. Weiterhin wird in einer Reihe ordinaler GAMs nachgewiesen, wie Interaktionen kognitiver oder funktionaler Aufgabeneffekte mit ausgwählten Komplexitätsmerkmalen Modellierungsergebnisse signifikant verbessert. Die Regressionsanalysen eröffnen zudem höchst interessante Einsichten in Ausgabeneffekte und geben wertvolle Anstöße für weitere Untersuchungen in kontrollierteren Datensätzen.

Klassifikationsexperimente in den *Merlin* Daten zeigen zudem, dass alle ordinale GAMs unter Einsatz von nur 13 Komplexitätsmerkmalen F1 Werte von mindestens 71% erzielen. Das beste Modell, das im Rahmen dieser These gebildet wurde, erreicht unter zusätzlicher Berücksichtigung von Leistungseffekten sogar F1 Werte von 85%.

Keynotes: *Komplexität, Genauigkeit, Flüssigkeit; Generalisierte Additive Regressionsmodelle; Lernerkorpusforschung, Deutsch als Zweitsprache, Zweitsprach-Schreibkompetenz, Zweitspracherwerb, Aufgabeneinflüsse*

Acknowledgements

I want to sincerely thank my supervisor Detmar Meurers for his advice and input, his continued support and tested patience, and for giving me a free hand in pursuing this topic way beyond the scope of a Master thesis.

Furthermore, I want to express my deep gratitude towards my second examiner, Harald Baayen, who had always an open door for my endless questions, and on whom I could rely for advice and help, whenever I struggled with my data.

I also want to thank Marije Michel, for the review and discussion of my retrospective annotations of task factors; Simón Ruiz for his repeated help, when I struggled with SLA nomenclature and literature; and my private mathematician, Maximilian Weiß, for his review of the *Additive Regression Modeling* chapter.

Last but not least, I want to express my profound gratitude towards my parents Heidemarie Weiß and Carsten Thielmann, as well as to my brother and my sister in law Maximilian and Tatjana Weiß. Without your unfailing support and continued encouragement, this thesis would not have been possible, either.

Contents

List of Tables

List of Figures

Part I.

Introduction

1. Introduction

This thesis analyses German L2 writing proficiency by means of a broad collection of complexity measures. It has an additional focus on task-effects on language complexity in learner corpora. To account for these, already included task information as well as retrospectively annotated cognitive and functional task factors are introduced to the descriptive and inferential statistical analyses, that are conducted on the cross-sectional *Merlin* corpus and the partially longitudinal *Falko Georgetown* corpus.

Complexity is – together with accuracy and fluency – a crucial dimension of language performance in Second Language Acquisition (SLA), and commonly analyzed in order to make inferences about various aspects of language performance, such as language proficiency, development, readability, or general aspects of cognitive ability (Crossley, Kyle & McNamara 2016; Hancke, Vajjala & Meurers 2012; Kyle 2016; Lu & Ai 2015). Despite the abundance of research committed to measuring language performance in terms of complexity, and the steady increase in quantity and sophistication of measures, findings have lead to inconsistent results across learner corpora. In recent years, a lack of theoretical underpinnings for commonly employed complexity measures and their interpretation in empirical studies has been blamed for this (Bulté & Housen 2014; Housen, Vedder & Kuiken 2012; Pallotti 2009). Furthermore, research broadened towards the investigation of factors, which might cause differences in the reported observations. Examples for this are analyses of heterogeneous learner profiles, due to varying first language (L1) backgrounds (Crossley & McNamara 2011), or differences in learning strategies (Jarvis et al. 2003). A more recent line of investigation targets task-effects on language performance, following insights from Task-Based Language Teaching (TBLT) and Second Language Instruction (SLI) research (Tracy-Ventura & Myles 2015; Yoon & Polio 2016). Task factors are rarely controlled for in learner corpora, and by extension, rarely accounted for in learner corpus-based studies (Tracy-Ventura & Myles 2015).

Following this line of research, this thesis adopts an approach by Alexopoulou et al. 2017, who investigate how task factors influence the observations, that are

made for complexity, accuracy, and fluency (CAF) on learner corpora with diverse task backgrounds. For this, they perform a retrospective analysis of cognitive and functional task factors from task-based frameworks by Skehan 1996 and Robinson 2001 on selected tasks from the EF-Cambridge Open Language Database (EFCAM-DAT). While Alexopoulou et al. 2017 analyze a small collection of features from all three CAF dimensions, this thesis focuses exclusively on complexity. It has a dual focus on both, i) measures of complexity themselves; and ii) observable task-effects on them. In particular, the following research questions are investigated:

1. How do measures of complexity model German L2 proficiency?
2. To which extend is this influenced by cognitive or functional task-effects?
3. Does a retrospective analysis of German learner corpora with diverse task backgrounds improve complexity-based L2 proficiency modeling?

To address these questions, a collection of 398 complexity measures from the domains of language use; human language processing; discourse and encoding of meaning; and three sub domains of theoretical linguistics, namely syntax; lexicon; and morphology is investigated. These complexity measures are automatically extracted using an augmented and updated version of the system used in Galasso 2014; Hancke 2013; Hancke, Vajjala & Meurers 2012; Weiß 2015. The system is employed on the German section of *Merlin* corpus and the L2 section of the *Falko Georgetown* corpus: Both corpora exhibit diverse task backgrounds, that may potentially interfere with proficiency modeling due to their heavy correlation with test and course levels.

Then, two strands of analyses are conducted: First, a descriptive cross-corpus analysis of complexity across proficiency levels is performed. It includes more than 100 complexity measures, which were selected from the set of nearly 398 measures based on theoretical considerations. In particular, they were chosen to represent at least one commonly assessed complexity concept from each domain listed above, for example lexical variation or periphrastic grammatical constructions. This allows to get an overview over differences between proficiency levels for a large basis of varied measures, to address the first research question. It also takes a first step towards addressing the second research question, by comparing measures in diverse and heterogeneous task backgrounds across corpora.

Second, exploratory ordinal regression analyses were conducted on each corpus using GAMs. These studies address all three research questions, by modeling proficiency with a data-driven selection of complexity measures and comparing, how

introducing interactions of complexity measures with task factors effect the models. A detailed discussion of each study further investigates, to which extend the results seem to reliably depict task-effects, and to which extend idiosyncratic properties of the uncontrolled corpora limit the interpretability of findings. Together, all studies clearly confirm the influence of task factors on complexity in German learner corpora, which has so far predominantly been discussed for English. Depending on complexity domain and concrete measure, complexity features exhibit diverging degrees of sensitivity to task-effects and heterogeneous task backgrounds, though. The analyses also clearly show the limitations of using task factors obtained form a retrospective analysis of corpora, which are not controlled for task factors, and give various prompts for future research in more controlled settings.

The remainder of this thesis is structured as follows: The next three chapters provide more theoretical background for this thesis. This starts with a brief overview over related approaches in Chapter 2 with a special focus on popular systems for complexity analyses, recent studies on complexity, approaches that factor in heterogeneous learner profiles, and, finally, studies that incorporate task-effects in their research on linguistic complexity. Throughout this literature review, the main focus was set on studies investigating L2 proficiency.

Chapter 3 continues with a proper theoretical introduction of complexity, as a it is conceptualized in SLA together with accuracy and fluency as a dimension of language performance. After briefly outlining this framework as a whole, some common criticism is discussed with a special focus on issues regarding the operationalization of complexity. Then, popular complexity taxonomies are discussed. Chapter 4 elaborates on how task factors may influence language performance in SLA. For this, task factors and their background in TBLT research are briefly discussed and related to Learner Corpus Research (LCR). Then, two common framework of cognitive task factors are introduced as well as functional task factors including a brief discussion on how they are assumed to effect complexity.

Afterwards, the next three chapters are dedicated to the two German learner corpora, that are analyzed throughout this thesis: Chapter 5 presents the German section of the *Merlin* corpus and Chapter 6 the L2 section of the *Falko Georgetown* corpus. Both chapters briefly describe the respective corpora, before elaborating on the tasks represented in them and the data sets designed for the following analyses. Then, Chapter 7 discusses the manual annotation of cognitive and functional task factors on both corpora. After outlining the annotation procedure, the resulting task factors and their distribution across both corpora are discussed.

The next part of this thesis is dedicated to the automatic analysis of linguistic complexity. For this, a brief discussion of all measures elicited in this thesis is presented in Chapter 8. It separately discusses measures of language use, discourse and the encoding of meaning, human language processing, as well as measures of lexical, syntactic, and morphological complexity. Then, Chapter 9 outlines the system used for feature extraction. For this, the general pipeline and the underlying resources are outlined, before elaborating on which linguistic units are assumed for the analyses. The chapter closes with a brief comment on the evaluation of feature extraction performance.

Chapter 10 presents a descriptive visual analysis of about 100 complexity measures and compares them across corpora. These measures were selected based on theoretical considerations and include measures from all research domains discussed in Chapter 8. The chapter also includes a link to plots for all 398 complexity measures that were extracted throughout this thesis, but could not be discussed here for reasons of space.

After this, some background on GAMs is provided, which are used for the following statistical analyses. For this, the chapter first gives a brief background on linear regression in general, before elaborating on generalized linear and additive regression modeling. Then, a more detailed discussion of regression splines and spline penalties is given, before closing with a discussion of additive mixed models and ordinal regression.

The second to last part of the thesis presents two empirical regression studies that were conducted on *Merlin* and *Falko Georgetown* to investigate task-effects on complexity measures for increasing L2 proficiency. Chapter 12 outlines the general exploratory model design followed throughout all regression studies. Then, Chapters 13 and 14 present the conducted studies, focusing especially on model fit and discussion, as well as on classification performance for *Merlin*. The thesis concludes with some final remarks on the overall findings and future work in Chapter 15.

Part II.

Theoretical Background

2. Related Work

There is already a vast amount of research on the operationalization of empirical measures of complexity and how to use them to model language proficiency, text readability, writing quality, and related aspects of language performance in a broader sense. This chapter starts with a brief overview over larger surveys of research findings and freely available, automatic complexity analysis systems in Section 2.1. The remainder of the chapter targets a sample of recent empirical studies on complexity: Section 2.2 elaborates on prototypical studies on complexity for the assessment of L2 and partially L1 proficiency and development. Section 2.3 focuses especially on studies of heterogeneous learner profiles. The chapter closes with a discussion of recent approaches to complexity-mediated proficiency assessment under the influence of task effects in Section 2.4.

2.1. Surveys & Systems

There is a long history of studies that use linguistic complexity to assess some form of language performance. For readability assessment, these go back to at least the 1920s (Thorndike 1921; Vogel & Washburne 1928), and to the 1930s for student's writing assessment (Frogner 1933; LaBrant 1933) and continued throughout the 1960s and 1970s (Hunt 1965, 1970; Larsen-Freeman 1978) until today. While these earlier attempts relied mostly on few superficial global measures of complexity that may be extracted from texts with no or minimal automatic support, over the past decade progress in the development of NLP tools facilitated the systmatic investigation of more sophisticated measures from various linguistic domains, such as syntax, lexicon, morphology, semantics, etc., as well as discourse and cohesion, human language processing, and language use.

There are several research syntheses in the field of complexity analyses, which are dedicated to maintain a clear overall picture of the ongoing research discussion given the increasingly growing set of measures and – partially diverging – findings. Probably the most influential survey on this issue is Wolfe-Quintero, Inagaki & Kim

1998. They review over 100 CAF measures in L2 writing from 39 studies on English, French, Swedish, German and Russian as second or foreign languages, to research how writing development is evaluated and which measures appear most promising. They focus on the development of syntax, lexicon, and morphology. For their study, they analyze variation and correlation and overall effects of measures for proficiency level assessment as an approximation of language development, because most of the investigated studies relied on independent proficiency studies. They assume CAF measures to be good indicators of language development, if they progress linearly and consistent with proficiency levels. Based on this, they recommend clause type, article, and passive ratios, as well as, words, clauses, dependent clauses, verb phrases, lexical and sophisticated word types, error free clauses, and errors per t-units or clauses. There have been several follow-up syntheses of complexity research, which may be consulted for a more recent overview over the topic, for example Ortega 2003, who summarizes 25 cross-sectional and longitudinal studies on college-level English L2 writings; or – more recently – Housen, Vedder & Kuiken 2012, who reviewed 40 studies with special regard to the divergent operationalizations and underlying theoretical conceptions of complexity.

When reviewing literature on language performance and linguistic complexity, a fundamental shift may be observed from manually encoded complexity measures towards increasingly automated approaches relying on NLP tools. Today, there is a variety of automatic complexity analysis systems available to facilitate complexity analyses for (inter-)language systems for various research purposes, such as proficiency assessment, essay grading, readability assessment, plagiarism detection, or authorship attribution. For English, the Coh-Metrix Web Interface is one of the most influential systems. It assesses over 600 measures of discourse, and lexical and grammatical sophistication of English (Crossley & McNamara 2012: 120) and has been used in a vast number of studies on English L2 and L1 writing quality and development. Other popular systems for English are the Syntactic and Lexical Complexity analyzer by Lu 2010, the Linguistic Analysis tool by Kyle 2016, and the Common Text Analysis Platform (CTAP) by Chen & Meurers 2016. Also, there is a collection of more than 200 measures for readability assessment by Vajjala 2015. There are also some systems for German, such as the DeLite readability checker by von der Brück & Hartrumpf 2007; von der Brück, Hartrumpf & Helbig 2008, which is based on 48 sophisticated measures of morphological, lexical, syntactic, and semantic complexity as well as of discourse and cohesion and human language processing. The system used in this thesis has not yet been made available for other

researchers, but will be integrated to CTAP as soon as time permits.[1]

2.2. Recent Complexity Studies

Most complexity studies are corpus-based and measure language development and proficiency in terms of syntactic and lexical complexity. This holds in particular for longitudinal studies, which are typically based on learner corpora of adult English L2 writings, which were elicited in the context of university writing courses. Knoch, Roushad & Storch 2014, for example, analyze English as a Second Language (ESL) writing in an immersion context by eliciting writings from 101 English L2 students at an English-medium university in a longitudinal test-re-test design study with a duration of 12 months. Students did not participate in writing courses during the time of the study, though. ibid. only find significant increases in text length in words, which they relate to fluency. Error free clauses and t-units (accuracy), words per t-unit, words per clause, or clauses per t-unit (grammatical complexity), as well as lexical sophistication, Malvern et al. 2004' diversity index, and the amount of academic words (lexical complexity) did not increase significantly. In a subsequent study, Knoch et al. 2015 extend the study to an observation window of 3 years with 31 undergraduate ESL students, which confirms their previous findings. Interestingly, this is not in line with Ortega 2003' investigation of findings from 25 studies on college-level English L2 writings, who attested that taken all studies together substantial increases in mean t-unit lengths could be observed after 12 months. Knoch, Roushad & Storch 2014; Knoch et al. 2015 relate this lack of significant development partially to a lack of practice and instruction reported by the students in interviews and questionnaires, as well as to the essays' topic restrictions and limited size. However, in particular the diverging findings for t-unit length might as well be interpreted as evidence for Biber, Gray & Poonpon 2011's claim that t-units are unreliable measures of language proficiency, because they have lead to inconsistent observations across studies in the past (ibid.: 13), see also Norris & Ortega 2009 for cirticism on t-units.[2] On a more general note, it should be

[1] Note that unlike other complexity analysis systems, CTAP was designed to be an expandable platform, thus allowing researchers to share their measures and to re-use measures from others (Chen & Meurers 2016).

[2] In fact, there might be a general issue with t-unit-based measures due to diverging definitions of t-units: While most studies refer to Hunt for a definition of t-units, Foster, Tonkyn & Wigglesworth 2000 find overall four distinct definitions of t-units by Hunt. These differ for example with respect to whether adjacent, solitary linguistic material may be included in t-units, see Biber, Gray & Poonpon 2011; Foster, Tonkyn & Wigglesworth 2000 for details.

pointed out that Knoch, Roushad & Storch 2014; Knoch et al. 2015 only measure a limited number of global measures to measure CAF. This limits the explanatory power of their findings, because they might fail to capture the development of local structures.

Crossley & McNamara 2014 assess syntactic complexity on a more local scale. For this, they extracted 11 clausal and phrasal measures using Coh-Metrix (Graesser et al. 2004; McNamara et al. 2014). With these, they analyze 57 advanced adult English L2 writings elicited in a semester long intensive writing course, to investigate L2 development of the syntactic domain and how measures of language development relate to expert writing quality ratings. They find a significant increase in measures of noun modification, words preceding the main verb, and negation, and a decrease in the redundancy of syntactic patterns, verb phrase frequency, and the number of clauses. At the same time, they find that essay ratings correlate most with clausal complexity, showing a divide between language development and proficiency ratings.

Bulté & Housen 2014, too, assess structure-level complexity features in English L2 writings from 45 adult ESL learners participating in an academic English program over 4 months. They analyze 10 manually extracted measures of syntactic complexity and 3 automatically extracted measures of lexical complexity with respect to their progress-sensitivity as features of L2 development and proficiency. For syntactic complexity they find significant increases of nominal modification and clausal coordination. Yet, they cannot confirm increases in subordination. As they point out, this mirrors Crossley & McNamara 2014's results, but is not in line with findings by Norris & Ortega 2009, who report subordination to complexify for intermediate learners. Interestingly, they do not find significant changes for lexical complexity in their data. This might be due to the limited amount of measures they applied, as they only measure three transformations of the type token ratio: the diversity index (Malvern et al. 2004), the Guiraud index (Guiraud 1960), and the advanced Guirad index (Daller, Hout & Treffers-Daller 2003). For a more detailed discussion of these indices, please see Jarvis et al. 2003.

A linguisically more diverse study is conducted by McNamara, Crossley & McCarthy 2009, who assess English L1 writing quality in terms of coreference, connectives, syntactic complexity, lexical diversity, and word characteristics measured by Coh-Metrix. They find no significant correlation for the use of connectives or co-referentiality. However, they find more proficient writers to use a higher average number of higher level constituents per word and on average more words preceding

the main verb. In terms of lexical complexity, more proficient learners had higher Measure of Textual Lexical Diversity (MTLD) scores (McCarthy & Jarvis 2010) and used less frequent words.

Crossley & McNamara 2009 investigate differences between English L1 and L2 writers based on lexical and cohesion measures. For this, they again assess i) noun, argument, and stem overlap between sentences as measure of co-reference; ii) latent semantic analysis; iii) word frequency, familiarity, concreteness, imagability, meaningfulness, and age of acquisition; iv) hypernymy and polysemy indices; v) spatiality; and vi) ratios for causal connectives with Coh-Metrix. They find that L1 writers exhibit higher ratios for argument overlaps and causal verbs. Also, they use more abstract and hierarchically connected words, more lexical connections, and more polysemous words, which are also more familiar. However, L2 writers use more words that are acquired late.

Crossley et al. 2011 use similar measures of lexical sophistication, syntactic complexity, cohesion to assess English L1 development of adolescents and young adults as a function of grade levels (9th, 11th, college first year) and compare it with ratings of writing quality: They measure i) cohesion in terms of ratios of causal verbs and particles, ratios of connectives, ii) coreference in terms of noun, argument, stem, and content word overlap, LSA, coreferecne chains of pronouns and nouns; iii) semantic complexity in terms of polysemy and hyperonymy ratios; iv) lexical complexity in terms of MTLD, lexical frequency measures, and familiarity, imagability, and concreteness; v) syntactic complexity in terms of the mean number of words preceding the main verb, the ratio of high-level constituents per word, and the ratio of modifiers per noun. They also assess the uniformity of syntactic constructions and text length measures in terms of number of sentences, paragraphs, and word length. They find that older writers produced more, less frequent, less diverse, and more abstract words. They also produced significantly more complex noun phrases. However, younger students used more connectives and more content word overlap. These findings are very similar to those from Crossley & McNamara 2009 for the comparison of L1 and L2 writers. Yet, younger students use more polysemous words: this is a difference to ibid., where L2 writers used less polysemous words.

While Coh-Metrix includes some measures of syntactic patterns, but target phrasal complexity predominantly in terms of modifier ratios, Paquot 2017 specifically studies phraseological complexity in advanced English L2 writers with French L1 background. Her approach is based on pattern occurrences of phraseologi-

cal units as defined by Gries 2008 and works in analogy to the most common approaches to assess lexical complexity in terms of lexical diversity and lexical sophistication: Paquot 2017 measures the root of the ratio of unique phraseological patterns to all phraseological patterns in a text as a phraseological version of a root type token ratio. Also, she assesses phraseological sophistication in terms of i) the ratio of phraseological collocations that occur in an academic collocation word list to all collocations; and ii) by point-wise mutual information of each collocation follwoing Bestgen & Granger 2014 who found that mutual information scores for bigrams are correlated with human CEFR ratings. Note that this operationalization of phraseological sophistication assesses language use and stylistic differences rather than the theoretical linguistic notion of a complex phrase, which is why phraseological complexity is used different from phrasal complexity throughout this thesis.[3] Only adjectival and adverbial modifiers as well as direct objects were considered as phraseological patterns in her study. Also, she assessed several measures of lexical diversity and sophistication for comparison using Lu 2010's L2 Syntactic Complexity Analyzer. Her results show that neither the measures for lexical diversity nor the measures for phraseological diversity are able to identify significant differences across intermediate to high proficiency levels (B2 to C2), however, phraseological sophisication, unlike lexical sophistication, shows significant increases in point-wise mutual information of the collocations with increasing proficiency for adjacent and non-adjacent levels. In particular, she finds that adjective modifiers are suited best to distinguish between B2 and C2 levels, adverbial modifiers are suited best to distinguish B2 and C1/C2 levels and direct objects are suited best to distinguish C2 from lower levels.

In summary, the following general trends are often assumed to hold for increasing proficiency: On low and intermediate proficiency levels, learners tend to write increasingly longer sentences, which might either indicate globally increased sentence complexity or higher fluency, and use more subordination (Lu 2011; Norris & Ortega 2009; Ortega 2003). However, there is also evidence that increases in clausal complexity might rather measure characteristics of human ratings than actual language development (Crossley & McNamara 2014). This fits into findings that high clausal complexity is actually a characteristic of spoken language and that it is a common misconception to equate it with academic language (Biber, Gray & Poonpon 2011: 10). Hence, while it has been established that higher profi-

[3] Please see Chapter 8 for details on the difference between measures of language use and measures of theoretical linguistic concepts.

ciency ratings are associated with more clausal complexity, there is also evidence that this is rather a measure of rater conceptions than of L2 development. On higher proficiency levels, grammatical complexification mostly targets the phrasal domain (Taguchi, Crawford & Wetzel 2013), especially with respect to noun phrases (Crossley & McNamara 2014; Crossley et al. 2011; McNamara et al. 2014), and measures of lexical and clausal complexity often fail to account for differences between advanced-intermediate and advanced learners (Biber, Gray & Poonpon 2011; Ortega 2012; Paquot 2017). Other complexity measures, such as lexical and clausal complexity, were found to be less informative when distinguishing between these proficiency levels (Biber, Gray & Poonpon 2011; Ortega 2012; Paquot 2017). Note however, that increasing phrasal complexity is also an established characteristic of academic writing (Biber, Gray & Poonpon 2011; Biber, Gray & Staples 2014) and Paquot 2017's study targets academic L2 writing, while studies that do not report similar issues with assessing differences between high-intermediate to advanced learners are often conducted on argumentative essays and narratives (ibid.: 19). Hence, to my knowledge it stands to systematically assess to which extend genre and other functional task factors may influence this development. Finally, lexical complexity has been shown to increase along with L2 development in terms of lexical abstractness and familiarity for more proficient learners. The same holds for the complexity of the semantic word nets invoked in learner texts (Crossley & McNamara 2012; Crossley et al. 2014). However, these do not necessarily correlate with higher proficiency scores (Crossley & McNamara 2014). As for discourse and cohesion measures, studies conducted with Coh-Metrix indicate that at higher proficiency levels, more proficient learners employ less cohesion markers such as connectives, i.e. produce cohesion gaps (Crossley & McNamara 2012; Crossley et al. 2014; McNamara, Crossley & McCarthy 2009). However, they do exhibit partially increased uses of longer coreference chains, depending on their learner profile (cf. below) (Jarvis et al. 2003).

2.3. Studies on Heterogeneous Learner Profiles

Note that all most of these studies assume linear relationships between language performance and complexity features, i.e. they assume writers to form homogeneous proficiency groups with respect to these measures. Although several studies admit this to be a vast simplification (Crossley & McNamara 2011; Wolfe-Quintero, Inagaki & Kim 1998), this assumption is still commonly applied and usually leads

to moderate correlations. However, some studies do include the assumption of varying learner profiles in their analyses. Crossley & McNamara 2011, for example, investigate features of English L2 writing that remain stable across different L1 backgrounds, i.e. they analyze inter-group homogeneity of English L2 writers and find that hyperonymy, polysemy, lexical diversity, and stem overlap show homogeneous values across L1 backgrounds (Czech, Finnish, German, Spanish) while being significantly different from English L1 writings.

Jarvis et al. 2003 follow another approach and identify different learner profiles in high-rated essays via clustering and analyze to which extend clusters illustrate varying writing strategies across highly proficient writers: They cluster overall 338 texts from highly proficient English L2 learners with various L1 background based on 20 measures by Biber 1986, 1988; Biber, Conrad & Reppen 1998 including word length, type token ratio, measures for connectives, pronouns, tense and verb mode, articles and demonstratives. The resulting clusters show different writing strategies across L1 backgrounds. These profiles were predominantly defined by feature combinations: for example, two identified profiles related to learners who showed an increased nominal writing style, which was combined with the use of relatively few pronouns, while two other profiles were defined by high pronoun and low noun usage. Jarvis et al. 2003 point out that both, nominal writing style and the use of pronouns are associated with high proficiency, but seem to be used as complementary writing strategies with learners.

Overall, research findings in complexity analysis show some trends in how the inter-language system changes with increasing proficiency, but they also result in diverging findings. Although several additional variables have been investigated to explain apparently heterogeneous learner profiles that would explain the diverging findings, such as L1 backgrounds, complexification strategies (e.g. relying on nominalization rather than pronoun use), or mode (written vs. spoken), these aspects are not yet systematically included in complexity analyses. Another approach towards the analysis of factors that effect complexity in L2 and L1 performance is presented in the next Section, which elaborates on so called task effects.

2.4. Studies on Task-Effects

Learner corpus research and complexity studies based on secondary-usages of learner corpora have rarely investigated influences of task factors on CAF, see Section 4.1 for details. Recently, though, studies in these fields increasingly address

task effects (cf. Alexopoulou et al. 2017; Polio & Park 2016; Tracy-Ventura & Myles 2015). Also, studies from the domain of TBLT have focused on task effects on CAF since the 1990s in order to facilitate linguistically informed syllabus design and task comparison in language teaching need to take into account task factors (Robinson 2001: 27): For example, Foster & Skehan 1996 analyze functional task effects in terms of task theme, by investigating CAF differences between personal information exchange, narratives, and decision-making tasks. They find less subordinations in personal writing, which they confirm in a follow up study (Skehan & Foster 1997), where they find subordination to occur most often for their decision-making task. However, they also identify this task as the cognitively most demanding one, which makes the origin of the effect less clear. Furthermore, these studies unfortunately did not investigate other complexity measures and may, therefore, only give a very limited account on task effect on complexity.

More recent approaches are based on broader feature collections, though: Yoon & Polio 2016, who analyze how genre differences (argumentative vs. narrative writing) effect English L2 writings from 37 university students over four months in a longitudinal setting, measure 12 features of complexity extracted with Lu 2010's Syntactic Complexity Analyzer. These are sentential complexity in terms of i) lengths of sentential units; ii) coordination and subordination at the clausal level; phrasal complexity in terms of iii) ratios of complex noun phrases, verb phrases, and phrase coordination; and lexical complexity in terms of iv) word frequencies; v) word length; and vi) vocd-D as measure of lexical diversity. Also, they assessed accuracy and fluency. With these measures, they found that while there was a limited development of complexity over time, argumentative texts showed higher phrasal complexity than narrative texts: argumentative writings had significantly more verbs, complex noun phrases, and phrasal coordination. However, they did not find convincing effects on the clausal level. For lexical complexity, they found longer words in argumentative texts, but they also were lexically less diverse. Interestingly, these genre differences did not sufficiently replicate on the L1 writings they used for comparison. This indicates that genre effects L1 and L2 writing differently.

Another study targeting task effects on L1 and L2 was conducted by Tracy-Ventura & Myles 2015, who analyses past tense morphology in spoken L1 and L2 Spanish across three task themes: i) a guided interview; ii) a picture-based narrative; and iii) a description of historical figures. They found, for example, that learners reliably produced imperfect tense only for narrative tasks, but not for the others.

From their results, they argue for the need of variable task background in learner corpora, to facilitate the study of language acquisition, since different context clearly elicit different constructions, which leads to the issue of underrepresentation in corpora.

Vasylets, Gilabert & Manchón 2017 analyze the effects of task complexity and mode on L2 performance for 78 university students, who were English learners with a Spanish and Catalan L1 background. They were asked to perform argumentative instruction-giving tasks once in a simple and once in a complex version; half of the participants solved the tasks orally, the other half solved them in writing. Task complexity was assessed in terms of reasoning demands, which is a resource-directing measure from Robinson's Cognitive Hypothesis (CH), see Chapter 4. Linguistic complexity was assessed in terms of 4 measures: they measure structural complexity in terms of analaysis-of-speech units (AS-units) per tokens and syntactic nodes per AS-unit as a measure of subordination.[4] They also measure phrasal complexity in terms of the modifier / noun phrase ratio from Coh-Metrix. Lexical diversity is measured with the D-value and lexical sophistication in terms of a lexical frequency profile and propositional complexity is assessed in terms of idea units and extended idea units (see ibid.: 407f for the respective definitions). They also assess accuracy and time on task. They find that aside from significant differences between task modes, across task modes learners produced on average longer AS-units and more idea units when solving tasks with higher reasoning demands. Also, they used more sophisticated language as assessed by the lexical frequency profile. However, they found no cognitive task effects on noun phrase complexity, subordination, and lexical diversity. They also found some effects to be only present in written mode, such as an increase in extended idea units, i.e. differences in types of ideas, or an enhancement of accuracy.

Alexopoulou et al. 2017 investigate how functional and cognitive task factors effect accuracy and fluency in the EF-Cambridge Open Language Database. Task themes are narrative, descriptive, and professional texts; cognitive task factors are measured in terms of Skehan's Limited Attentional Capacity Model (LACM) and Robinson's CH. The complexity measures they investigate are i) syntactic complexity in terms of the ratios of words per sentence and clause and the ratio of subordinate clauses per t-unit, wh-phrases per sentence, and complex noun phrases per clause;

[4] Analysis-of-speech units are syntactic units for the analysis of spoken language similar to t-units: "An AS-unit is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either." (Foster, Tonkyn & Wigglesworth 2000: 365).

ii) lexical complexity in terms of MTLD; iii) cohesion and coreference in terms of global and local argument overlap; and iv) morphological complexity in terms of ratios of past tense verb forms, ratios for gerundial verb forms, and ratios of third person singular simple present verb inflections; using the complexity measures by Vajjala & Meurers 2012. Overall, they report increases for their complexity measures for less cognitively demanding tasks, but only within task themes. Across task themes, functional task effects are found to be stronger than cognitive task effects: They find more lexical diversity and higher mean t-unit and clause lengths for professional texts, but more subordination per t-unit for narratives. Also, they find more argument overlap (global and local) and more wh-phrases for narratives than for other task themes and more complex noun phrases and more gerunds for professional texts. Overall, their findings confirm the predictions made based on functional and cognitive task factors.

3. Language Complexity in Second Language Acquisition

Measures of complexity have a long tradition as predictors for more general constructs of language performance, such as L2 and L1 proficiency, development, writing quality, or text readability (Bulté & Housen 2014; Ellis & Barkhuizen 2005; Ortega 2012; Wolfe-Quintero, Inagaki & Kim 1998). This chapter discusses the theoretical background of L2 and L1 complexity as a dimension of the CAF framework from SLA research. After briefly introducing the full framework in Section 3.1, Section 3.2 presents some commonly discussed criticism on CAF with a special focus on language complexity. This is followed by a comprehensive overview of complexity taxonomies in Section 3.3.

3.1. CAF Framework

As a concept, L2 complexity origins from research on SLA, where language performance is assessed in terms of complexity, accuracy, and fluency (CAF) (Housen, Vedder & Kuiken 2012). *Fluency* refers to native-like production speed and *accuracy* to native-like production error rate (Housen, Vedder & Kuiken 2012; Pallotti 2009; Wolfe-Quintero, Inagaki & Kim 1998). *Complexity*, though, is usually not defined with respect to an explicit L1 norm. One popular definition understands complexity in terms of the elaborateness and variedness of a system's components and their inter-relatedness.[1] This definition is proposed by Rescher 1998 for complexity in general and by Ellis & Barkhuizen 2005 specifically for L2 contexts and is widely used throughout CAF literature (Bulté & Housen 2014; Housen, Vedder & Kuiken 2012; Pallotti 2009, 2015; Paquot 2017).[2] However, it should be noted that the

[1] Although most studies refer to only the notions of elaborateness and variedness, but tend to exclude inter-relatedness of components.

[2] It should be pointed out, that in the following 'complexity' is discussed from the perspective of SLA. Beyond this narrow focus, complexity is a very common term across research domains, for example in philosophy, information theory, linguistics, mathematics, literature studies, etc. There are, therefore, several traditions of defining complexity, which are not of further interest for this

definition of complexity exhibits some variance across studies (cf. below). In fact, complexity is commonly argued to be an inherently vague concept (Hennig 2017: 7; Rescher 1998: 8). Furthermore, it should be clear, that despite the lack of explicit reference to a norm in most definitions, complexity is not an absolute but a relative property. Put differently, a construction in itself is not more or less complex, but requires a reference norm compared to which it may be considered to be more complex. Thus, constructions are alyways more or less complex in comparison to other constructions, which might either be explicitly referenced or relying on an implicit standard of reference.[3] To paraphrase Fischer 2017: 21: some construct **a** cannot be inherently complex by itself, but should rather be thought of as being more complex than another construct **b** for some recipient **c** with respect to some sub-system **d**, if some operationalizational criterion is met. Given this notion, it seems plausible that complexity should rather be measured in terms of sets of diverse local measures of complexity, rather than relying on a few global measures or attempting to generate a single complexity score (ibid.: 19), because a construct compared to another might be more complex in terms of some characteristics while being less complex in terms of others.

All three dimensions of CAF are commonly operationalized in order to assess some language performance (Polio 2012: 146f). For each of the constructs, there is a core set of measures that is typically associated with them across studies, which has been continuously augmented, especially since advances in NLP facilitate the implementation of increasingly elaborate features: Complexity is most commonly measured as *lexical complexity* in terms of type token ratios, lexical types per token,

thesis. However, there are some common definitions from information theory, that are applied to varying degrees across fields, and should, therefore, briefly be mentioned as a side note: i) the *Kolmogorov complexity*, which measures the complexity of x and y in terms of the length of their shortest possible description, and ii) Gell-Mann's *effective complexity*, which compares the length of the description of the regularities and structures of x and y. However, these definitions play a minor role in the context of measuring aspects of language performance and will, therefore, not be discussed in further detail. For a more comprehensive overview over information theoretic complexity, please see Rescher 1998.

[3]Note that this is not a unique property of complexity due to its aforementioned inherent vagueness. The requirement of a reference norm is caused by the semantics of gradable predicates in general and applies to *complex* the same way it applies to adjectives like *expensive* or *loud*. In fact, I would argue that many of the issues related to the operationalization and interpretation of complexity (and difficulty) in CAF research is strongly tied to the semantics of relative and judge-free adjectives. Readers with an inclination for semantics might find some approaches towards the definition of the semantics of these types of adjectives highly rewarding with respect to their understanding of complexity and the issues around its definition, operationalization, and measurement, see Kennedy & McNally 2005 for more general considerations to the semantics of gradable predicates and Pearson 2013 in particular for judge-free adjectives.

or sophisticated tokens per token, and as *grammatical* or *syntactic complexity* in terms of tokens, and coordinated or dependent clauses per sentence or t-unit, or modifiers per noun (Housen, Vedder & Kuiken 2012; Wolfe-Quintero, Inagaki & Kim 1998). Accuracy, on the other hand, is often measured in terms of error-free t-units or overall error counts (Wolfe-Quintero, Inagaki & Kim 1998). Fluency is primarily assessed via production unit rates in time, where production units may be words, sentences or utterances.[4] Unfortunately, information on production time is not necessarily available, especially for written productions, which makes this type of measure less accessible. Hence, ibid.: 14 also consider frequency or length of sentences, t-units, utterances, clauses, and phrases as relevant production units for the assessment of fluency, and thus consider measures such as words per t-unit or noun phrase as indices of fluency, see also Knoch, Roushad & Storch 2014: 3 and Yoon & Polio 2016 for a similar view. Interestingly, these measures are also commonly associated with complexity., though This is due to the fact that, while conceptually distinct, observationally CAF are not fully independent: If complexity is at least partially bound to an increase in linguistic units, writing more complex constructions requires an increase in fluency, too. In order to differentiate complexity and fluency, Wolfe-Quintero, Inagaki & Kim 1998 introduce granularity of measure as an argument: since length measures are insensitive to the means used for an increase in length (e.g. coordination, subordination, phrasal modification), they do not measure complexification of the syntactic structure so much as production rate. This criterion is not applied consistently across other studies, though, so that researchers have to take care when comparing research findings to account for potentially diverging categorizations of measures.

3.2. CAF Criticism

Despite having shown to be a useful concept for research on SLA and First Language Acquisition (FLA), CAF has been criticized on theoretical and methodological grounds in recent years, due to issues of separability of its components in actual language productions, see for example Bulté & Housen 2014; Fischer 2017; Hennig 2017; Housen & Kuiken 2009; Housen, Vedder & Kuiken 2012; Pallotti 2009, 2015.

[4] Please note that this relates to speed fluency, while Tavakoli & Skehan 2005 also consider breakdown fluency and repair fluency as relevant subcategories. Since this section focuses on complexity, elaborating on these subtypes of fluency would be beyond the scope of this section. For a more comprehensive overview, please see Pallotti 2009; Tavakoli & Skehan 2005; Wolfe-Quintero, Inagaki & Kim 1998.

It is often argued, that the separate dimensions of CAF are not independently anchored or derived from cognitive fundamentals or theoretical definitions and thus lend themselves to varying interpretations and operationalizations (Housen & Kuiken 2009). This is problematic, because it makes observations and experiments hard to compare, interpret, and replicate. Especially the definition of complexity has been criticized to be too vaguely defined and to show high degrees of variation across studies (Bulté & Housen 2014; Housen, Vedder & Kuiken 2012; Pallotti 2009, 2015; Paquot 2017: 592).[5] In fact, empirical studies often fail to provide an explicit underlying definition of complexity, but instead contour complexity in terms of observational characteristics as more, longer, deeper, more varied, or less frequent linguistic constructions (Bulté & Housen 2014: 44f.), or more sophisticated language (Crossley & McNamara 2009, 2014; Wolfe-Quintero, Inagaki & Kim 1998).

This is illustrated in Figure 3.1 from Bulté & Housen 2014's research review on complexity analyses in SLA. It shows concepts they found to be often closely linked to or even equated with the concept of complexity in SLA and FLA research, which – as they point out – is highly problematic, because these constructs are to some extend related, yet conceptually clearly distinct. In particular, ibid. point out, that the attributions *late acquisition* and *more advanced, mature, or developed* exhibit a progressive, temporal component, which is frequently assumed implicitly or explicitly when analyzing complexity, despite not being included in any common definition of the term. This link also implies continuous growth of complexity, which – as Pallotti 2009: 594 and Bulté & Housen 2014: 45 point out – should rather be subject to investigation than an *a priori* assumption. This holds especially, since it is not undisputed from a theoretical point of view, whether complexity

[5]It is beyond the scope of this thesis to evaluate, whether complexity is in fact the least well defined CAF concept, because issues with the definition and operationalization of complexity are of primary interest in this section due to the focus of this thesis on complexity analyses. However, it should be pointed out, that while reviewing literature on complexity as a dimension of CAF, several issues related to accuracy and fluency became apparent, too: The question how fluency should be operationalized especially for written productions has already briefly been mentioned. Furthermore, most operationalizations of fluency, that are simple to employ in written contexts, such as number of words on some scale, are heavily confounded with other aspects such as complexity (cf. discussion above), but also by task factors such as planning time, cf. Yoon & Polio 2016: 5, i.e. they suffer from the same lack of clear interpretability as many complexity measures. Also, while accuracy is claimed to be the most well-defined dimension of CAF (Pallotti 2009: 591, Norris & Ortega 2003: 737), it has been criticized not to measure interlanguage development at all, because it assesses the degree of normative conformity of a production, cf. comparative fallacy (Bley-Vroman 1983), see also Pallotti 2009: 591f, Wolfe-Quintero, Inagaki & Kim 1998: 33. Furthermore, its operationalization, too, is not entirely clear, as the task of error categorization and weighting is in itself highly demaning (Pallotti 2009: 592). For a more comprehensive discussion of these issues, please see Bley-Vroman 1983; Pallotti 2009.

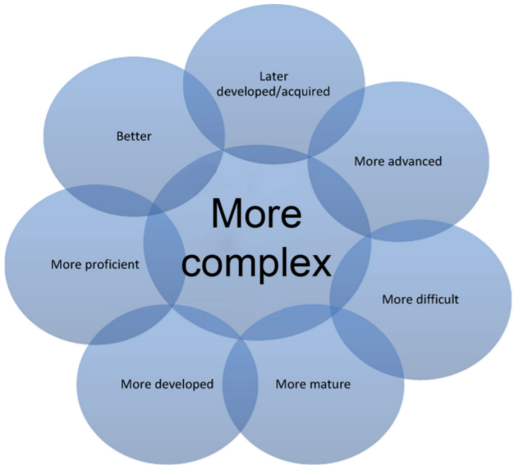


Figure 3.1.: L2 complexity and related constructs (Bulté & Housen 2014: 46)

in fact grows continuously with progressing development: While there is the philosophical assumption of ever increasing system complexity (cf. Rescher 1998), this assumption is not shared for L2 development in SLA research. In fact, u-shaped development or learning is a well established concept in SLA research, which is typically associated with the restructuring of the inter-language grammar (Patten & Benati 2010: 141).[6] *More proficient* combines this temporal-progressive component with the two other concepts shown in the graphic: *better* and *more difficult*. Difficulty is in fact often considered to be an aspect of complexity in a broader sense and will be further discussed in the next subsection. However, complex constructions are not inevitable side effects of proficiency or 'better' productions. Assuming that increased complexity necessarily equates with increased proficiency runs into a similar fallacy as the previous equation with development, insofar as this correlation, too, is rather a hypothesis to be tested than something to be incorporated in the implicit or explicit definition of complexity. This holds especially since, again, there are theories making diverging predictions as well as studies indicating that this simple equation does not hold unanimously. Sekhan's LACM, for example, which is elaborated on in Section 4.2, assumes that the CAF dimensions compete for limited cognitive resources, so that complexity is not only influenced by a learners proficiency, but also by the cognitive resources required by accuracy and

[6]Patten & Benati 2010: 142 uses the acquisition of irregular tenses in English as an example for this: Initially, irregular tenses are produced correctly, but learners start to overgeneralize regular past tense markings when learning and internalizing regular tenses. Only later in their acquisition do they start to form correct irregular past tense forms again.

fluency. Thus, a proficient learner could exhibit decreased proficiency compared to previous performance when pushing fluency. Furthermore, for most informal contexts overly complex language is inappropriate and may be seen as an indicator of a lack of socio-linguistic awareness or register competence (Ortega 2003: 494).[7] This is supported by Pallotti & Ferrari 2008, who compared native and non-native speaker's syntactic complexity. They found native speakers to adjust the complexity of their productions to what was required to efficiently solve their task at hand, while non-native speakers failed to adjust their language accordingly. In reference to findings like these, Pallotti 2009 suggests i) to use L1 productions as a reference standard for the extend of complexity that should be exhibited by proficient L2 learners (ibid.: 598), and ii) to assess *adequacy* as a fourth dimension of language performance, in order to measure to which extend a performance realizes its goal efficiently (ibid.: 596). This aspect is strongly related to the underlying task, thus relating non-linear relations in CAF not only to cognitive constraints but also to semantic and pragmatic task demands (ibid.: 599). Please see Chapter 4 for a more detailed discussion of adequacy and task demands.

3.3. Complexity Taxonomies

Several attempts have been made in recent years to clarify the concept of complexity by administering more rigorous definitions (Bulté & Housen 2014; Housen, Vedder & Kuiken 2012; Pallotti 2009, 2015). Bulté & Housen 2014; Housen, Vedder & Kuiken 2012, for example, address complexity following the general non-linguistic definition by Rescher 1998 as the elaborateness, variedness, and inter-relatedness of a system's components and accompany it with a detailed taxonomy of complexity constructs shown in Figure 3.2. In the following, this taxonomy serves as an orientation to discuss the most common distinctions made when describing complexity in SLA research and related fields from a theoretical angle.

The most crucial distinction, which is commonly made when discussing language complexity, distinguishes between *difficulty* (or *relative complexity*) and *absolute complexity*. Unfortunately, despite their common usage, researchers do not always relate the same concepts to those terms: Housen, Vedder & Kuiken 2012 understand difficulty as all anaspects of complexity, which are experienced individually, while

[7] Pallotti 2009 argues that this is the case for accuracy and fluency as well, pointing to a study by Sanell 2007, who found a decrease in accuracy when highly proficient speakers of French start to use colloquial speech.

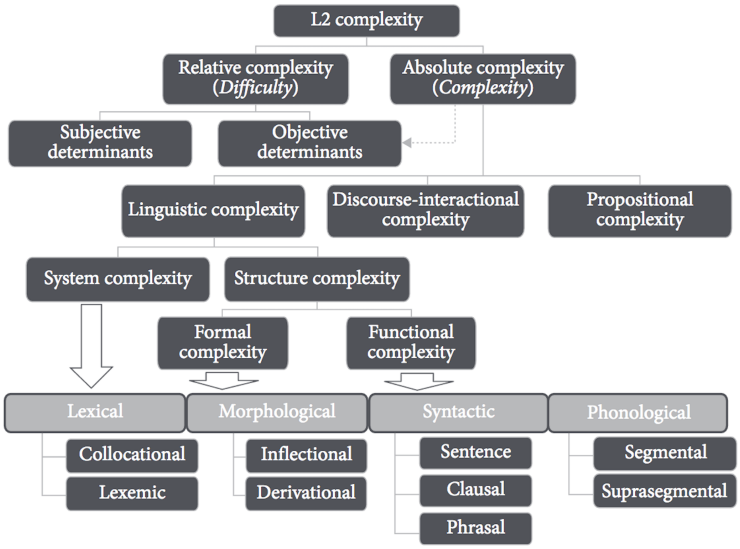


Figure 3.2.: Taxonomic model of L2 complexity, cf. Housen, Vedder & Kuiken 2012

complexity refers to more general aspects of language performance. This is what they understand to be commonly measured in complexity analyses and it is also the most widely agreed upon definition of absolute complexity (Dahl 2004; Fischer 2017; Miestamo 2008), which goes back to a similar distincton made by Robinson 2001 for task complexity, cf. Section 4.3.[8] Pallotti 2009: 593, however, uses complexity and difficulty to distinguish language performance properties and properties of task demands: Following Skehan rather than Robinson, she defines difficulty as an objective property of tasks and excludes subjective task difficulty from her taxonomy. Thus, when speaking about difficulty, it is advisable to specify the sense in which the term is used to avoid misconceptions. Housen, Vedder & Kuiken 2012 further distinguish between linguistic, discourse-interactional, and propositional complexity, which is a discrimination also explicitly made in previous work, especially based on Coh-Metrix, cf. Crossley & McNamara 2009, 2014.

Additional to these general distinctions of types of complexity found in the upper half of the figure, in the lower half of the taxonomy, Housen, Vedder & Kuiken 2012 further identify aspects of linguistic complexity with regard to their scope and dimension of impact independent of the language domain they apply to: They differentiate between the complexity of a system as a whole (*system complexity*) and

[8]This distinction is sometimes presented using a different terminology, though: Dahl 2004 refers to difficulty as *subject related* complexity and to absolute complexity as *object related* complexity, for example.

the complexity of a system's components (*structure complexity*). This distinction is also made by Pallotti 2015: 119f, but she uses the term *text complexity* instead of *structure complexity*.[9] Finally, the distinction between formal and functional complexity addresses an issue also discussed in Pallotti 2009: 593, who questions to which extend the notion of complexity applied to different linguistic domains is in fact comparable. She points out, that across linguistic domains, complexity is more commonly equated with i) highly compositional structures, ii) an increase in alternative realizations, and iii) agglutination of components. The first aspect relates to what Housen, Vedder & Kuiken 2012 define as *formal complexity*, which measures the quantity of discrete components of an index, such as the derivation steps used to derive a target structure. The third aspect relates to what they define as *functional complexity*, which measures functional ambiguities or syncretisms. The difference between the two is illustrated in Example 1.

(1) 'She carrie-**s**$_1$ her dog-**s**$_2$.'

While the English plural marker -s_1 and the English third person singular present tense marker -s_2 have equal formal complexity, the latter is functionally more complex. Increases in alternative realizations are not covered in ibid.'s taxonomy – which does not claim to be exhaustive – and it might be worthwhile to consider augmenting it with a third sub-type of structure complexity addressing this aspect. Note that Pallotti 2015 also advocates to differentiate between mandatory and voluntary complexity. Note that ibid. actually uses the terms *grammatical* and *stylistic complexity*, in order to differentiate between complex structures that are required for grammaticality and complex structures that constitute stylistic choices. However, the terms *mandatory* and *voluntary complexity* seem more adequate, since the underlying distinction may not only be apply to syntactic complexity but also to semantic or

[9] Also, it should be noted that she questions, whether system complexity is in fact measurable for any language system, but especially for L2 systems, since they cannot be fully assessed, which is a necessary requirement for her definition of complexity, though. She thus restricts operationalizations to text complexity (Pallotti 2015: 120). Bulté & Housen 2014; Housen, Vedder & Kuiken 2012, though, do not share these concerns. In this context it should be noted, that Pallotti 2015 in fact excludes all measures, which make assumptions not supported by theoretical definitions or are ambiguous in terms of whether they are sensitive to more than changes in complexity. In particular, she argues for using only unambiguous measures of structural and formal aspects of the linguistic system (ibid.: 118). Based on this, only the following measures remain: inflectional morphology for L1 assessment, syntactic complexity in terms of the number of constituents and lexical diversity, but not lexical density or frequency, because functional and lexical words are not structurally more complex (ibid.), whereas the other lexical measures do assess structural differences in the linguistic system. While such a restrictive approach seems conceptually interesting, it seems debatable whether it is more fruitful than cautiously interpreting findings from a less restricted set of measures.

pragmatic complexity. To my knowledge, this difference between voluntarily and mandatorily introduced complexity has not yet been used in empirical settings.

On a final note it should be pointed out, that all types of complexity discussed here apply throughout linguistic domains. However, as Pallotti 2009 remarks, they are not equally frequently operationalized across domains. For example, she states that there is a prevalence of equating complexity in the lexical domain with an increased number of alternative realizations, i.e. increased variability, whereas the syntactic domain often assesses elaborateness of language. In fact, all studies reviewed in this thesis as well as the studies conducted here do suffer from an under-representation of measures of variedness for all except the lexical domain. This is problematic, because the validity of generalizations about the development of complexity in a linguistic subsystems is dependent on the representativeness of measures employed to assess this linguistic system. Ideally, measures should assess the linguistic subsystem under investigation globally or or locally (i.e. the entire system as well as specific structures). Structure-specific constructions should assess formal or functional complexity in terms of elaborateness, variedness, and inter-relatedness irrespective of the language domain. The bandwidth of measures employed in a study in terms of these concepts directly points to the generalizability of its findings, because missing concepts of complexity in a study may be prone to only show an incomplete picture of CAF development. The knowledge of elaborate complexity taxonomies may help researchers to evaluate findings on complexity in terms of their representativeness and should guide the development of new complexity measures as well as the comparison of diverging findings across studies.

4. Task Factors in Second Language Acquisition

In task-based elicitation contexts, task performance as well as language performance is heavily influenced by task-effect (Alexopoulou et al. 2017; Robinson 2001; Skehan 1996; Tracy-Ventura & Myles 2015; Yoon & Polio 2016). Research on how certain task factors effect performance has mostly been conducted in TBLT research, where the primary focus is on task performance and language performance as a central aspect of task performance. However, as discussed in Section 2.4, the analysis of task factors also becomes increasingly common in LCR and SLA analyses, that focus specifically on language performance and CAF. This chapter introduces task factors for this CAF and learner corpus focused context. Note that for this, tasks are understood as "holistic activit[ies], which engage[] language use in order to achieve some non-linguistic outcome while meeting a linguistic challenge, with the overall aim of promoting language learning" Samuda & Bygate 2008: 69. This is not the only definition of task, that has been established in TBLT, but it is the one promoted by Alexopoulou et al. 2017, because it lends itself nicely to the application to learner corpora, which may be regarded as "collection[s] of instructed writing activities" (ibid.: 4). This makes it especially suited for the purposes of this thesis, too.

The remainder of this chapter is structured as follows: Section 4.1 briefly outlines the origins of task-based frameworks in the context of TBLT and reviews how this relates to LCR. Then, Sections 4.2 and 4.3 introduce two task-based frameworks, that focus on cognitive task factors: A central element in these frameworks is the assumption, that cognitive task factors have a major impact on language performance due to both, the communicative requirements they impose and the cognitive resources they consume. The chapter closes with Section 4.4, which discusses another productive strand on task-effects, that does not reason about cognitive demands and their effects on learners, but on the functional effects of tasks on language performance. Note that this type of task factor is higly related to the issue of *adequacy* discussed in Section 3.2, i.e. to the assumption that the complexity of

language productions is bound to the requirements of a task's communicative goal as well as socio-cultural practices and roles (Bruner 1986; Ravid & Tolchinsky 2002).

4.1. Task Factors in TBLT and LCR

The investigation of task-effects and their influence on various aspects of language and task performance goes back to the mid 1980s and early 1990s, when the focus in SLI shifted from grammar exercises to task-based teaching (Robinson 1995; Skehan 1996: 100). This turn prompted the demand for means of informed syllabus design, sequencing tasks, and comparing and evaluating task performance (Skehan 1996: 53; Robinson 2001), with a primary focus on language performance as an aspect of task performance in TBLT. This need lead to the development of task-based frameworks, such as Skehan's LACM and Robinson's CH, which are centered around the operationalization of cognitive, functional, sociological, and individual task factors. Today, the importance of varied elicitation tasks for the analysis of learner language has been widely acknowledged in research on TBLT, SLI, and SLA (Ellis & Barkhuizen 2005; Foster & Tavakoli 2009; Tavakoli & Foster 2011; Tracy-Ventura & Myles 2015).

Many learner corpora and empricial studies do not control for task factors, though. Yoon & Polio 2016, for example, criticize that most longitudinal studies do not control for functional task factors such as genre or task, rendering them impractical for the analysis of language development, see also Tracy-Ventura & Myles 2015 for criticism on this issue. ibid. argues that this lack of task awareness in corpus based CAF studies is due to LCR being rooted in corpus linguistics, which does not promote the controlled elicitation conditions, that would be required in order to obtain data from a sufficiently varied and balanced distribution of tasks: Experimental elicitation procedures of data are argued to impair the spontaneity and authenticity of the elicited data, which are two highly emphasized characteristics of corpora in corpus linguistics (Gilquin & Gries 2009; Tracy-Ventura & Myles 2015). These conflicting requirements of representative, varied learner language that counters issues of under-representation and task-effects on the one hand (cf. SLA), and authentic, spontaneous data, that was not collected in an experimental setting on the other hand (cf. corpus linguistics) lead to an ongoing discussion in LCR. For more details on this, please see Tracy-Ventura & Myles 2015: 5f.

In the meantime, most empirical studies on CAF and L2 proficiency or development suffer from a lack of 'task-aware' resources, despite the known impact of

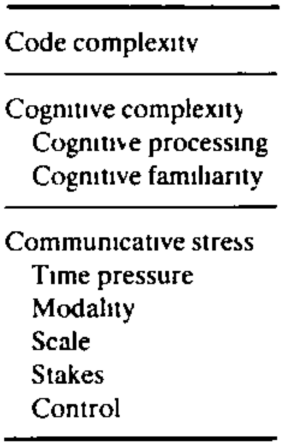


Figure 4.1.: Task Factors identified by Skehan 1996: 52.

task-effects on language performance. However, the readily available task-based frameworks form research on TBLT allow for a re-analysis of existing learner corpora, even if these exhibit unfavorable task distributions across potential response variables, as long as they feature varying tasks, cf. Chapter 7.

4.2. Skehan's Limited Attentional Capacity Model

Skehan's Limited Attentional Capacity Model (LACM) (Skehan 1996, 1998) is a prominent task-based framework in TBLT research: It differentiates tasks on the basis of their *difficulty* assessed in terms of their i) code complexity; ii) cognitive complexity; and iii) communicative stress, see Figure 4.1 from Skehan 1996: 52 for an overview.[1] *Code complexity* targets formal factors of language code (ibid.: 52), such as vocabulary load and variety, but also syntactic difficulty (ibid.: 52). *Cognitive complexity* assesses the processing demands of a task, which may be increased by the requirement of active, on-line reasoning, that could not be prepared in advance (i.e. cognitive processing) and decreased by the applicability of previous schematic knowledge or common structures, that may be re-used (i.e. cognitive familiarity) (ibid.: 52). Cognitive complexity may be reduced by providing supplementary material that reduces the learners' need for ad hoc reasoning or increased by

[1]Note that Skehan's use of the term *difficulty* is interchangeable with his notion of *complexity*: A systematic difference between the two terms in order to establish a clear distinction between subject-dependent and subject-independent factors is introduced by Robinson 2001 and does not apply to Skehan's use of difficulty and complexity in the LACM (ibid.: 29).

introducing elements that are usually not associated with the task at hand (Skehan 1996: 55). *Communicative stress* refers to aspects of the context in which a task is embedded: These include *time pressure* introduced by the time available for problem solving as well as the *modality* of task instruction and solution: spoken instructions or solutions are assumed to be cognitively more demanding than those transported in written mode (ibid.: 52). It also includes *scale* factors such as the number of task participants and the involved relationships as well as the *stakes* of solving the task, i.e. the pressure under which a learner is to solve a task completely and correctly (ibid.: 52). Finally, giving more *control* over task goals and input speed to a learner may ease communicative stress in Skehan's framework (ibid.: 53).

Increasing difficulty along these three dimensions is assumed to consume more attentional resources from the learner's limited resource pools. The resulting reduction of available cognitive resources is assumed to introduce a competition between CAF (Alexopoulou et al. 2017: 4;Skehan 1996: 51;Skehan & Foster 1997). For example, reducing the cognitive difficulty of a task by offering more time for preparation or increasing task familiarity, is predicted to release more attentional resources towards accuracy and complexity (Skehan 1996: 54). For too difficult tasks, the LACM predicts high emphasis on fluency at the expense of both, complexity and accuracy, because learners attribute so many processing resources on conveying the correct meaning, that they are forced to resort to less complex, lexicalized language without paying particular attention on accuracy (ibid.: 53).

Note that Skehan's LACM is specifically targeted at task sequencing and design in class rooms. For this, it entails a pre-task, during-task, and post-task set up, which is not elaborated here due to the specific focus of this thesis on complexity analyses. For a more detailed account of the full framework please see **1998**; ibid.

4.3. Robinson's Cognition Hypothesis

Another prominent task-based framework is Robinson's Cognitive Hypothesis (CH) (Robinson 1995, 2001). As Skehan's LACM it, too, consists of three main components and is centered around the assumption that cognitive processing demands affect how attentional resources are allocated to CAF: i) task complexity; ii) task conditions; and iii) task difficulty. Figure 4.2 shows an overview of these three components. In this framework cognitive factors such as those collected under Skehan's dimension of cognitive complexity are allocated under the component *task complexity*. A core difference to Skehan's LACM is, that Robinson's CH assumes,

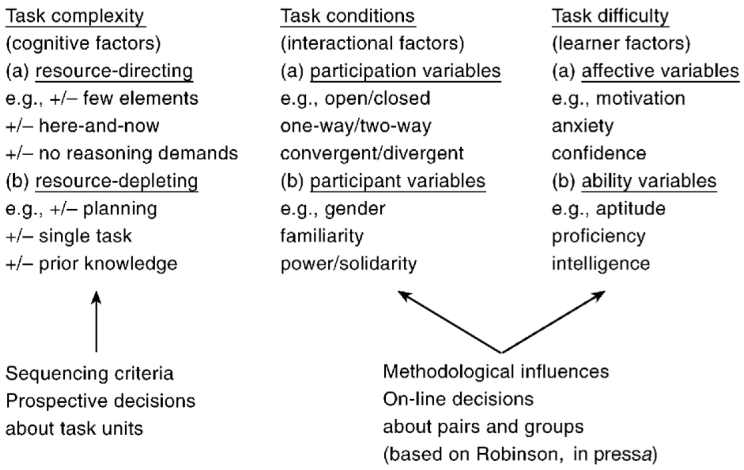


Figure 4.2.: Task complexity, condition, and difficulty (Robinson 2001: 30, Figure 1).

that learners may access multiple pools of attentional resources (Alexopoulou et al. 2017; Robinson 1995). Cognitive task factors, thus, may not only deplete attentional resources from CAF, but also direct them towards the language use, if functional needs require increased focus on language code to convey the intended meaning with sufficient precision. This is predicted to lead to increases in accuracy and complexity, potentially on cost of fluency.

For example, one type of resource-directing factors is temporal and spatial dislocation of references, which is originally a factor discussed in FLA research, cf. Robinson 1995: 102: The discussion of abstract or displaced frames of reference is cognitively more demanding, since the mental abstraction requires more cognitive resources. It also functionally requires the use of more deictic expressions as well as clarifications and descriptions of the dislocated setting in time or place and, thus, is prone to elicit more complex language in terms of referring expressions, temporal structures, and adverbial clauses. In this context, full spatio-temporal dislocation (*there-and-then*) is often set in opposition to full confound with the task setting (*here-and-now*) (Robinson 1995, 2001; Robinson, Ting & Urwin 1995). However, partial dislocation of either space or time are possible conditions, too, i.e. *there-and-now* or *here-and-then*, see Alexopoulou et al. 2017 for an example and Robinson, Ting & Urwin 1995 for a more detailed account on dislocated references in general. Other types of resource-directing task factors are the amount of referenced elements and the amount of abstract reasoning required for a task: more reference elements cognitively require learners to hold more discourse referents in memory. They also

functionally require the use of referring expressions. High reasoning demands are associated with an increased use of connectives and more elaborate clausal syntax.

Resource-depleting task factors are assumed to cause cognitive resource demands, that draw attention from the working memory without redirecting it to language code (cf. Robinson 2001: 30). These include increased planning time, the number of tasks, and prior knowledge. For example, increasing the number of tasks to be solved does not functionally require learners to employ more elaborate or varied language on either of the individual tasks. Instead, it divides cognitive resources and time, which makes it possible to enhance fluency due to time pressure, but not accuracy or complexity.

Aside from cognitive task factors, the CH includes two more components: interactional factors, which are referred to as *task conditions*, and learner factors, which are referred to as *task difficulty*. Task conditions are to some extend similar to Skehan's communicative stress: Participation variables are based on circumstantial factors of the task set up and include aspects like the direction of the information flow (mono-directional or multi-directional) or whether the task is open or closed with respect to its communicative coals. This is clearly similar to Skehan's factors control and modality. It also includes participant variables, which are sociological factors of the task set up such as gender, familiarity of participants, and group dynamics, which relates to scale in Skehan's framework.

Task difficulty is a concept, that is not directly found in Skehan's LACM, which does not differentiate between subjectively and objectively high task demands. It includes more permanent individual factors such as proficiency, aptitude, and intelligence (i.e. ability variables) as well as highly variable, situation-dependent variables, such as motivation, anxiety, and confidence (i.e. affective variables). Parts of these affective variables relate to Skehan's dimension of communicative stress, too, since the perception of time pressure and stakes as well as how much they affect task performance also have a subjective component, that depends on motivation, anxiety, and confidence. When comparing Robinson's task difficulty and task conditions with Skehan's communicative stress, it becomes apparent, that there is a confound between subjective and objective factors in the LACM.[2]

[2]In fact, Robinson 2001 points out to be the first to introduce a distinction between learner-dependent difficulty and learner-independent complexity in a task-based framework (ibid.: 29). However, the distinction has been adopted in later research and is now especially productive in complexity taxonomies, cf. Section 3.3. In fact, the distinction between linguistic complexity and difficulty in the CAF framework is often attributed to the discussion of complexity and difficulty in task-based second language instruction frameworks.

Note, that the CH also includes a broader set of predictions towards how task complexity, task conditions, and task difficulty differ with respect to their susceptibility to task changes, because like the LACM it originally emerged in the context of syllabus design and task sequencing. For example, while ask complexity is assumed to be an invariant property of a certain task design, task conditions and task difficulty are assumed to be subject to methodological decisions by the teacher (Robinson 2001: 30). Yet, these aspects of the framework are of less interest for this thesis and, therefore, not discussed in more detail. Please see ibid. for a more comprehensive discussion of the framework.

4.4. Functional Task Factors

Skehan's LACM and Robinson's CH both make concrete and partially diverging predictions on how task factors should influence CAF. Yet, experiments and studies conducted to test both frameworks lead to mixed results (Yoon & Polio 2016). Another strand of research focuses on purely functional aspects of tasks, without further reasoning on how these affect aspects such as cognitive processing and memory load (Alexopoulou et al. 2017; Foster & Skehan 1996; Lu 2011; Vyatkina 2012). In fact Biber, Gray & Staples 2014; Yoon & Polio 2016 suggest, that functional differences might even have a greater impact on CAF measures than cognitive factors and explain the so far inconclusive findings with regard to the predictions made by LACM and CH with this.

Functional aspects are, for example, a task's communicative goal and sociocultural practices and roles (Bruner 1986; Ravid & Tolchinsky 2002). However, one of the most commonly investigated functional task characteristics is task type (or discourse type), which was found to have a high impact on CAF measures (Biber, Gray & Staples 2014; Foster & Skehan 1996). The following genres are commonly distinguished in SLA literature (Berman & Slobin 1994; Bruner 1986): i) narrative texts, which describe events by focusing on people and their actions within a defined time frame (Berman & Slobin 1994); ii) argumentative texts, which make logical arguments or discuss ideas and beliefs (ibid.); iii) descriptive texts, which list or describe facts (Alexopoulou et al. 2017). Biber, Gray & Staples 2014 argue, that these texts should exhibit different linguistic properties according to their functional differences: for example, narrative texts are expected to exhibit more past tense markers and third person pronouns, while argumentative texts are expected to contain more relative clauses.

Part III.

Data

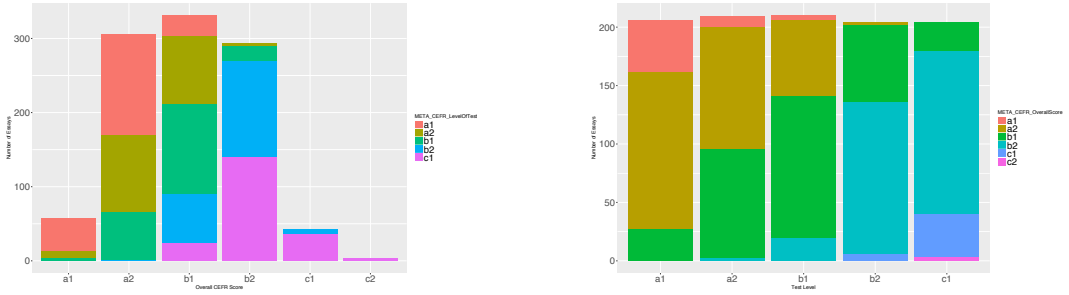
5. German Merlin Data

This chapter introduces the German section of the *Merlin* corpus, which is one of the largest freely available German learner corpora and includes texts from an extraordinarily broad variety of proficiency scores assessed by expert CEFR ratings. The following section first briefly describes the available data in terms of its quantity, elicitation context, and available meta information. Section 5.2 then elaborates on the different tasks, that are included in the German *Merlin* corpus. Section 5.3 closes with a brief description of the data set, which was used for further analyses throughout this thesis.

5.1. Corpus Description

Merlin is a cross-sectional trilingual L2 corpus, which was compiled in the course of the EU LLP project MERLIN: "Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context" (2012-2014) (Abel et al. 2013). It was designed to represent learner writings across CEFR proficiency levels for Czech, German, and Italian L2. Overall, it consists of 2,286 texts. The German section – henceforth *Merlin*, since neither the Czech nor the Italian section are of relevance for this thesis – has a size of 1,033 texts from different learners at varying proficiency levels. The data were elicited from official standardized language certification tests. From these tests, approximately 200 freely written texts were extracted per test level, with test levels ranging from A1 to C1 for Merlin ibid.: 113. These hand-written texts were digitized as close, anonymous transcriptions. Additionally, a normalized version (target hypothesis I) was developed by forming orthographically and grammatically acceptable sentences with minimal changes, to facilitate automatic NLP on the data ibid.: 118. Text length in tokens is highly variable across test levels. Level A1 tests have a mean text length of 48.96 ($\pm$19.85) tokens, A2 tests 68.79 ($\pm$27.63), B1 tests 106.10 ($\pm$40.87), B2 tests 167.01 ($\pm$37.03), and C1 tests 218.74 ($\pm$43.23).

Each text provides general information on learners, such as age, gender, and



(a) Number of essays per holisitic proficiency score grouped by test level.

(b) Number of essays per test level grouped by holisitic proficiency score.

Figure 5.1.: Distribution of holistic proficency scores and test levels in the German section of *Merlin*.

first language as meta information as well as further details on the test situation, such as task id, test level (A1-C1) and a variety of detailed proficiency ratings (A1-C2). Thus, *Merlin* is not only unusually large for a German L2 corpus, it also provides a uniquely broad range of CEFR proficiency ratings. Other German L2 corpora, like *Falko*, represent a considerably more narrow range of learner proficiency (Abel et al. 2013: 119). Texts were rated by human experts trained on the CEFR-based Merlin rating grid by Wisniewski et al. 2013. It contains CEFR scales for grammatical accuracy, vocabulary range, vocabulary control, coherence and cohesion, orthographic control, and socio-linguistic appropriateness (Abel et al. 2013: 113). Furthermore, a holistic CEFR rating scale created based on these separate ratings. It ranges from A1 to C2. These ratings will be referred to as Merlin proficiency scores or ratings throughout the thesis, in contrast to the CEFR level assigned to each test, which will be referred to as test levels. This distinction is important, because it is possible, and indeed quite common in the data, that learners score proficiency levels different from the level of the test taken, either by means of over- or under-performance. This leads to a heavily skewed distribution of holistic proficiency levels, despite the fact that Merlin is balanced for test levels, as illustrated in Figure 5.1a: Levels A1 and C1 account for only 10% of the data, while the other 90% are approximately evenly distributed among levels A2 to B2 and with only four instances, less than 0.1% of the data are attributable to level C2. These proficiency levels are assigned across several test levels, as may be seen in Figure 5.1b. This is most extreme for B2 rated learners, who may be found throughout all test levels.

The figures also, show that learners may score several levels above or below

Success	A1	A2	B1	B2	C	∑
Failed	13	67	90	140	0	310
Passed	44	239	241	153	46	723

Table 5.1.: Distribution of success (passed/failed) across overall CEFR scores on *Merlin* data.

their respective test level, see for example the distribution of B1 scores assigned throughout test levels. Combining test levels and proficiency scores makes it possible to infer whether learners succeeded or failed in their tests, although this is not a pre-annotated variable in the meta data. For this thesis, performance was included as a binary predictor, that indicates whether learners scored at or above their test level (*success*) or below their test level (*fail*). Table 5.1 shows the distribution of passed and failed tests across CEFR scores. Note that with this definition of success, there are no failures at the A1 level, because the *Merlin* data does not contain any texts scoring lower than A1.

5.2. Tasks

Overall 15 approximately evenly frequent represented tasks are included in the German Merlin corpus: Each task belongs to a single test level (A1 to C1) and each test level features three tasks. Table 5.2 illustrates this mapping from task to test level. It also displays the distribution of tasks across overall proficiency scores. The direct translation of task to test level influences this distribution unfavorably, since learners mostly achieve scores at the same or an adjacent level of the test they took. Only 5.8% of the assigned scores fall outside this range (marked in bold font).[1] Thus, proficiency scores may be predicted from tasks due to the idiosyncratic distributional properties of the data set: In a preliminary experiment task explained 49.2% of the variance of predicting proficiency scores in a simple GAM with no other predictors. Furthermore, the concurrence of task, test level and proficiency scores impedes the attribution of findings to a distinct source and the high number of distinct tasks makes them less suited as a categorical predictor in a regression

[1] These outliers themselves exhibit another structure: A1 and A2 level test takers overachieve, i.e. reach scores that are considerably higher than their test level, and B1, B2, and C1 level test takers underachieve, i.e. reach scores that are considerably lower than their test level. This effect contributes to the highly diverse composition of learners with B1 scores, which is normally distributed across all test levels.

Task	Test	∑	A1	A2	B1	B2	C1	C2
Going swimming	A1	56	8	45	**3**	0	0	0
Apartment search	A1	77	11	50	**16**	0	0	0
Child birth	A1	74	25	41	**8**	0	0	0
Ticket offer	A2	66	5	28	31	**2**	0	0
Pet sitting	A2	72	0	32	40	0	0	0
Housing office	A2	70	4	43	22	**1**	0	0
Announce visit	B1	67	**2**	31	29	5	0	0
Happy birthday	B1	70	0	24	38	8	0	0
Happy new year	B1	73	**2**	10	54	7	0	0
Application	B2	69	0	**1**	22	42	4	0
Work complaint	B2	70	0	**1**	20	47	2	0
Information request	B2	65	0	0	24	41	0	0
Housing situation	C1	72	0	0	**7**	52	13	1
Learning German	C1	42	0	0	**1**	26	15	2
Traditions & Assimilation	C1	90	0	0	**16**	62	12	1

Table 5.2.: Mapping of tasks to test levels, task frequency, and their distribution across overall proficiency scores (A1 to C2). Scores, that are more than one level above or below the test level, are highlighted in bold font.

model.

5.3. Data Sets

It has been noted in the previous section, that the distribution of CEFR scores in the Merlin corpus is not balanced, cf. Figure 5.1a. When conducting classification experiments for L2 proficiency, this may cause bias when classifying the under-represented levels. For these cases, it might be advisable to create a balanced data set, that only uses a subset of the data provided by the Merlin corpus. However, this is bound to lead to a massive reduction of either the number of texts or the number of proficiency scores represented in the data.

Since the primary objective in this thesis is to model the data, keeping both, the broad variety of proficiency scores and the large data support was preferred over a balanced distribution of proficiency scores. Hence, all 1,033 texts from the Merlin data were used. The only changes made to the data were merging all C1 and C2 scores into a single C1/C2 level. The resulting data set was used throughout all studies on the Merlin data. No other data set was designed from the corpus.

6. Falko Georgetown L2 Data

This chapter introduces the *Falko Georgetown L2* corpus, which is part of the error annotated collection of German L2 and L1 writings in the *Falko* corpus collection. It is to my knowledge the only freely available German learner corpus with a longitudinal subsection. The following section first briefly describes the available data in terms of its quantity, elicitation context, and available meta information. Section 6.2 then discusses the different tasks, that are included in the *Falko Georgetown L2* data. Section 6.3 closes with a brief description of the *Falko Georgetown L2* corpus-based data sets, that were used for further analyses throughout this thesis.

6.1. Corpus Description

The *Falko Georgetown* corpus consists of German L2 and L1 texts, the latter were not considered throughout this thesis, though.[1] The L2 sub corpus of *Falko Georgetown* was elicited between 2001 and 2004 throughout four consecutive curricular writing courses at Georgetown University by the German Department of Georgetown University. The courses were offered to intermediate to advanced L2 speakers of German, who predominantly studied German as a major or minor and were mostly native speakers of English. Each course was offered as either an intensive one-semester course or a regular course, which lasted two semesters. The German L1 sub corpus consists of comparable German L1 baseline texts written by professional authors, students, and teachers. Overall, *Falko Georgetown L2* corpus consists of 209 texts, which were written as take-home final exams by 123 students. The number of tokens per text increases with advancing course levels, but shows wide standard deviations, like the *Merlin* texts: Mean text length in tokens is 317.2 ($\pm$150.9) for the first course level, 507.6 ($\pm$246.8) for the second course level, 566.0 ($\pm$188.1) for the

[1] Note that although *Falko Georgetown* is part of the error annotated *Falko* corpora, it does not seem to be augmented with a readily usable target hypothesis: All writings collected in the corpus are transcriptions of the original student writings without any corrections. Furthermore, neither the handbook of *Falko* corpora (*Das Falko-Handbuch Korpusaufbau und Annotationen*) nor the documentation of *Falko Georgetown* indicate the existence of a target hypothesis for the data.

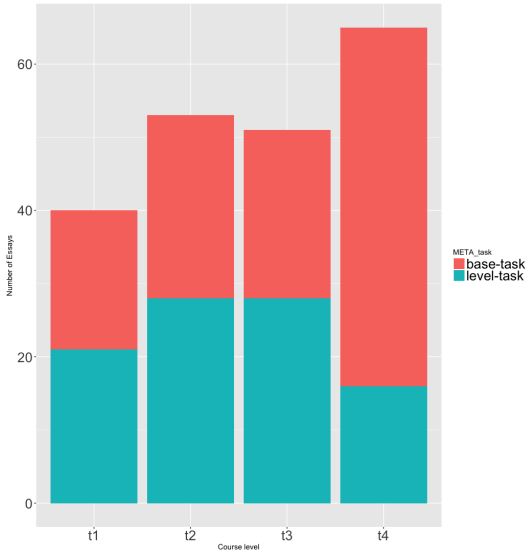


Figure 6.1.: Number of essays per course level grouped by writing occasion.

third course level, and , 615.1 ($\pm$419.7) for the fourth course level.

The corpus contains several meta information annotations including course duration (1 or 2 semesters), course level (1 to 4), hours per course, task, elicitation semester, and writer id. Figure 6.1 shows the distribution of texts across courses. All texts were either collected in course-dependent curricular writing tasks or in a curriculum independent reference task, which was administered across course levels. This is illustrated in the figure by color grouping. The data does not feature any form of external evaluation of the learners' writings, such as CEFR scores or grades. However, the writing courses are designed as consecutive courses targeting learners with increasing exposure to the language. They were, therefore, used as approximations of L2 proficiency, assuming that more advanced courses elicit more advanced writing. It was further assumed, that L2 development of intermediate to advanced learners is progressing towards higher proficiency, i.e. that the equation holds with some validity.[2]

When eliciting the *Falko Georgetown L2* corpus, several students participated in multiple courses. This lead to a longitudinal sub corpus consisting of 116 texts by

[2]It should be noted, though, that this is clearly a simplification, since development is known to not necessarily linearly progress towards higher proficiency, as in the case of u-shaped development, see Wolfe-Quintero, Inagaki & Kim 1998 for a critical discussion. In want of a more suited approximation, it was nevertheless decided to rather use this coarse approximation of proficiency than not analyzing the data, and to keep in mind that potential inconsistencies in the findings might be due to this decision.

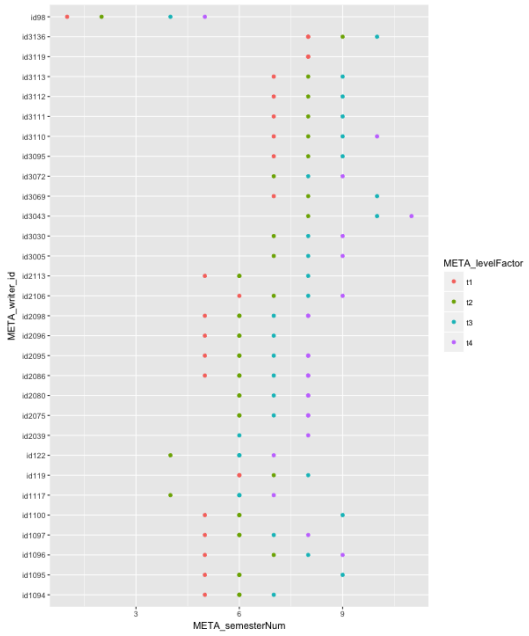


Figure 6.2.: Texts written by learners who contributed multiple writings to *Falko Georgetown L2* data plotted by semester and grouped by course level.

30 students. Since the elicitation of the longitudinal data was not controlled, some students participated in intensive as well as regular courses, though, and not all students participated in the same amount of courses. Hence, the longitudinal sub section of the *Falko Georgetown L2* corpus is less conservative than other longitudinal learner corpora. Figure 6.2 shows all texts from students who contributed more than one text to the corpus sorted by semester (winter term 2001 to winter term 2004). In this graphic, color groupings indicate course level. It shows that most learners participated in 3 courses, mostly levels 1 to 3, and some participated in 4 or 2 courses. Also, some learners discontinued their courses or participated in courses with a regular course duration (i.e. two semesters) rather than taking the intensive one-semester courses. This longitudinal section consists nearly exclusively of texts from curricular writings tasks.

6.2. Tasks

As mentioned before, the *Falko Georgetown L2* corpus includes texts from two different writing occasions: Most texts are the final writing tasks from the curricular language courses. For these, each course level is confounded with a specific task,

Task	∑	Level 1	Level 2	Level 3	Level 4
Write a letter	21	21	0	0	0
Continue a novel	28	0	28	0	0
Write an article	28	0	0	28	0
Write a speech	16	0	0	0	16
Book review	116	19	25	23	49

Table 6.1.: Frequency of tasks across course levels in the *Falko Georgetown L2* corpus.

as can be seen in the overview in Table 6.1. There is a total confound of tasks and test levels for all curriculum dependent tasks. The reference task, though, – a book review – contributes texts to all course levels. Since most longitudinal data is elicited in curriculum dependent tasks, analyses of proficiency on this part of the data are prone to suffer from undetected task-effects. Therefore, like *Merlin*, the *Falko Georgetown L2* data might benefit greatly from the explicit annotation of task factors.

6.3. Data Sets

The longitudinal subsection of the *Falko Georgetown L2* data is a unique characteristic of the corpus, therefore, it was decided to center the analyses around this subset of the data. However, as already mentioned, the data is not conservative insofar as the data set does not contain i) an equal number of data points per learner; ii) texts from the same course levels for each learner; and iii) data points were not controlled to be elicited in consecutive semesters. Hence, two versions of a longitudinal data set were initially designed: One more conservative longitudinal data set was created, which includes only texts written in intensive one-semester courses. It also only counted texts as longitudinal if the same learner produced them without discontinuing participation in the courses for one or more semesters. This lead to a data set consisting of 100 texts by 28 learners. Since this is only very few data, a second, more liberal longitudinal data set was designed, in order to test whether including discontinued participation and participation in two-semester courses made a difference. However, this less strict data set only gained 116 texts and 30 learners. Since including course duration and discontinued participation as variables in a regression model would consume two degrees of freedom, this was not considered to be an improvement. Hence, the more conservative longitudinal data set was used throughout all studies.

Because this is very little data, it was decided to test the validity of results by comparing models throughout various reference data sets. For this, three reference data sets were created: First, the *inverse* of the conservative longitudinal data was used to test how results replicate on data, that has not been included in the previous model. Overall, it includes 109 texts by 92 learners. Since the longitudinal data consists mostly of curricular writing tasks, this data set contains mostly book reviews, but also some non-longitudinal curricular writing tasks. Therefore, a second *book review* data set was designed, which only includes book reviews, and has a size of 116 texts by 110 learners. Finally, the *full* data was used as a data set, because it provides the largest data support available for this data.

7. Analysis of Task Factors

All tasks included in *Merlin* and *Falko Georgetown* were analyzed for cognitive and functional task factors. These additional annotations are primarily intended to make task differences in the corpora interpretable in statistical analyses. For this, task factors need to be variable and approximately uniformly distributed across course levels; a criterion which is not met by any of the tasks themselves in either corpus, cf. Sections 5.2, 6.2. A secondary effect of the analysis is, that the task design across course and test levels becomes analyzable: In both corpora, all tasks are designed for a specific proficiency level, except for *Falko Georgetown*'s curriculum independent reference task. Hence, it may be interesting to see, to which extend task factors mirror the increases in targeted proficiency.

In the next section, the operationalization of task factors is briefly outlined, to make the obtained annotations transparent. Then, Section 7.2 presents the annotation of task factors and the analysis of their distributional properties for the *Merlin* corpus. Section 7.3 does the same for *Falko Georgetown*.

7.1. Operationalization of Task Factors

For both corpora, the detailed task descriptions provided in their supplementary material were used as source of information for the analyses, see Merlin project 2014a,b,c,d,e,f,g,h,i,j,k,l,m,n,o for the *Merlin* data and *Falko Georgetown Dokumentation* 2007 for the *Falko Georgetown L2* data. With respect to the choice of task factors, the present approach follows Alexopoulou et al. 2017, who conduct a similar analysis on the EF-Cambridge Open Language Database (EFCAMDAT) corpus for L2 English: For this, they operationalize two factors of Skehan's LACM, namely i) code complexity; and ii) cognitive complexity, as well as four factors from Robinson's resource-directing task complexity factors, namely i) shared context (i.e. ±tempo-spatial dislocation); ii) reasoning demands; iii) the number of reference elements; and iv) perspective taking. These cognitive task factors were operationalized as follows:

Code complexity was categorized into high, medium, or low, depending on whether instructions provided no, few, or detailed language material to draw from.

Cognitive complexity was categorized into high, medium, or low, depending on the extend to which they required learners to reason about the structure of their writings as opposed to already clearly outlining the required text structure.

Shared context assesses temporal and spatial dislocation of the task setting compared to the writing situation. This information was directly taken from specifications provided in the task instructions.

Reasoning demands were categorized as high, medium, or low, based on the quantity and elaborateness of mental operations required by a task. In particular it was assessed, whether tasks required spatial reasoning, i.e. referencing a location without extra-linguistic support, and whether tasks required reasoning about other people's intentions, beliefs, desires, or relationships.

Referenced elements was categorized into few, many, and open depending on the number of discourse referents minimally required to solve the task that were named in the task descriptions.

Perspective was assessed based on whether learners were requested to include their own perspective (own), the perspective of someone else (other) or the perspective of multiple other people (others).

Additional to these, the following functional task factors were annotated in this thesis: i) genre; ii) target audience; iii) formality; iv) task theme; and v) task type. They were operationalized as follows:

Genre refers to the textual category elicited by a task. This information is encoded straight forward in the task descriptions of both corpora and was not altered for the analyses.

Audience refers to the targeted recipient. This information is encoded straight forward in the task descriptions of both corpora and was not altered except for choosing reasonable hypernyms, if this allowed to group more tasks together.

Formality refers to the tone of the writing. This information is encoded straight forward in the task descriptions of both corpora and was not altered for the analyses.

Task theme refers to the general direction of the topic of a text, such as professional/occupational interests, public social affairs, maintenance of private social relations (i.e. *small talk*), or (by extension) goal-oriented personal matters. The latter is referred to somewhat loosely as *demand* throughout the thesis, in want of a better term.

Task type refers to a popular category in task-based analyses, which is determined by a combination of functional needs and genre. It includes the categories: argumentative, narrative, descriptive/expositional, and instructional.

If additional information on the task or test setting was readily provided in the corpus material, such as writing time or the number of expected words, these were also included in the analysis without modifications.[1]

7.2. Task Factors in German Merlin Data

On *Merlin*, task factors were annotated as shown in Table 7.3 at the end of this chapter. Naturally, test properties such as time and number of expected words show a perfect confound with test level. Also, one may see, that many task factors group together test levels, but do not vary within test levels, such as formality, code complexity, and perspective taking. Genre is relatively stable, too, assuming that letters and emails are nearly interchangeable constructs, which might even be argued to belong to the same genre, due to their functional and formal similarities. Most variable in this regard is task type, which is only invariable for B1 tests. Other task factors vary at least partially, such as audience, task theme, cognitive complexity, shared context, reasoning, and referenced elements.

Overall, the full analysis of task factors shows a highly interesting picture, which makes tasks considerably more interpretable. Of particular interest is, how task factors and test properties change with increasing test level. For example, B1 test takers are given considerably less writing time and are required to include more discourse referents, but in turn code complexity is decreased. Compared to this, B2 test takers are given tasks with high code complexity and increased cognitive complexity, but the amount of referenced elements decreases for two of the three tasks, too. In this context, *work complaint* seems to be considerably

[1] The only exception to this is number of expected words in the *Merlin* data for B1 tests, which are the only tasks where no number of words is given. However, since all B1 level tasks included letters to which test takers should write a response to, the word expectation was set to the average number of words contained in these prompt letters.

more demanding than its peer tasks on the same test level, since it provides no shared context, includes many reference elements, and shows medium cognitive complexity. Unfortunately, a detailed analysis on whether this makes learners, who take B2 tests, perform different across tasks in terms of language complexity, was beyond the scope of this thesis. A more detailed analysis of the implications of differences between task factors within test levels on measures of language complexity is planned for future work, though.

The distributional properties of task factors across proficiency scores are shown in Table 7.2. For each task factor, it shows the distribution of the task factor's levels across proficiency ratings. It may be seen, that most task factors are broadly distributed across proficiency scores. This holds in particular for task type and theme, which are broadly distributed and shows approximately the same number of instances per factor level. This is particularly remarkable, because they both have a relatively large number of factor levels, whereas most other factors are binary or ternary.[2] Most binary measures are represented with all their factor levels throughout proficiency scores, but heavily skewed in the overall number of instances per factor level; this holds, for example, for code complexity or shared context, as well as for referenced elements.

For statistical analyses, none of the identified task factors is ideal, but all measures, that are mostly fully represented across proficiency scores should be in principle analyzable. Most promising, though, seem task type and task theme, since they have very similar numbers of instances for each factor level and are not ideally, but sufficiently represented across proficiency scores. Furthermore, each of them provides four different factor levels, which allows for a more varied analysis than the binary task factors. Aside from these data-related aspects, both measures are commonly investigated functional task factors, and thus, also highly relevant from a theoretical perspective.

7.3. Task Factors in Falko Georgetown L2 Data

On *Falko Georgetown*, task factors were annotated as shown in Table 7.4 at the end of this chapter.[3] It shows, that most tasks address a public audience in a predominantly formal setting, but with very different task themes and types. Also,

[2]The only exception from this are the test properties writing time, which also have 4 to 5 factor levels, but are, unlike task theme and type, completely confounded with test level.

[3]Note, that this table does not include writing time or expected number of words, since neither of these information was contained in the *Falko Georgetown* documentation.

Feature	Level	∑	A1	A2	B1	B2	C
Writing time	30	414	4	67	187	150	6
	45	207	44	136	27	0	0
	50	208	9	103	93	3	0
	60	204	0	0	24	140	40
Expected words	30	207	44	136	27	0	0
	40	208	9	103	93	3	0
	128	210	4	65	121	29	9
	150	204	0	2	66	130	6
	200	204	0	0	24	140	40
Genre	Essay	204	0	0	24	140	40
	Letter	624	38	179	248	153	6
	Email	205	19	127	59	0	0
Formality	Formal	408	0	2	90	270	46
	Informal	625	57	304	241	23	0
Type	Descriptive	337	37	129	64	94	13
	Instructional	215	16	110	87	2	0
	Argumentative	271	0	2	59	177	33
	Narrative	210	4	65	121	20	0
Theme	Demand	341	28	198	112	3	0
	Society	204	0	0	24	140	40
	Profession	204	0	2	66	130	6
	Small talk	284	29	106	129	20	0
Audience	Agency	274	4	45	88	131	6
	Friend	555	53	261	219	22	0
	Public	204	0	0	24	140	40
Code complexity	Low	210	4	65	121	20	0
	High	823	53	241	210	273	46
Cognitive complexity	Low	690	57	304	265	64	0
	Medium	139	0	2	42	89	6
	High	204	0	0	24	140	40
Shared context	here & now	813	44	245	241	239	44
	there & then	220	13	61	90	54	2
Reasoning	Low	344	50	210	78	6	0
	Medium	557	7	96	236	199	19
	High	132	0	0	17	88	27
Referenced elements	Few	549	53	240	166	86	4
	Many	280	4	66	141	67	2
	Open	204	0	0	24	140	40
Perspective	Own	829	57	306	307	153	6
	Own & Others	204	0	0	24	140	40

Table 7.1.: Distribution of task properties provided in or inferred from supplementary material across proficiency scores.

all tasks are cognitively demanding insofar as they require learners to write in a spatio-temporally fully dislocated setting and to handle multiple referents. Code and cognitive complexity as well as reasoning demands differ across tasks, but do not show a systematic increase correlated with higher task level. This indicates, that task instructions were not systematically manipulated to lead to cognitively more demanding tasks with advanced courses.

In order to identify which task factors are suited variables for further analyses, the distribution of task factor levels across course levels was investigated. This is shown in Table 7.2. In order to be interpretable in a statistical analysis, all levels of a task factor have to be represented roughly uniformly across as many course levels as possible. Due to the perfect confound of curricular writing tasks and course levels, and the fact that only one task was included in the curricular-independent writing setting, only three task factors meet this requirement: i) elicitation context; ii) code complexity; and iii) cognitive complexity. *Elicitation context* has the most favorable distribution across course levels, but it only separates one task from the other four tasks. Since these curricular tasks exhibit highly heterogeneous task factor profiles, this seems insufficient for an analysis of specific task effects. However, it does allow to split the data into two equally sized partitions, of homogeneous and heterogeneous task backgrounds. While such a divide is unsuited for the analysis of specific task effects, it does allow to compare whether measures are influenced by differences in task backgrounds or whether they are equivalent, when calculated on homogeneous and heterogeneous task backgrounds.

Code and cognitive complexity are measures of Skehan's LACM, which both measure cognitive task complexity as defined by Robinson 2001. Their influence on language complexity is thus very interesting from a theoretical perspective. Also, they are the only measures, which are variable across most course levels and divide tasks in a roughly balanced opposition of 2:3, rather than 1:4 as elicitation context does. Hence, these two task factors are most suited for further statistical analyses, that target the effect of specific task factors. Since they both have the same theoretical background and both have instances for the high and low condition for all courses but one, they seem roughly equally suited for further analyses.

Feature	Level	∑	Level 1	Level 2	Level 3	Level 4
Elicitation context	Curricular task	93	21	28	28	16
	Reference task	116	19	25	23	49
Audience	Friend	40	40	0	0	0
	Public	169	0	53	51	65
Genre	Letter	21	21	0	0	0
	Novel	28	0	28	0	0
	Article	28	0	0	28	0
	Speech	16	0	0	0	16
	Review	116	19	25	23	49
Formality	Formal	160	19	25	51	65
	Informal	49	21	28	0	0
Theme	Small talk	21	21	0	0	0
	Mystery	28	0	28	0	0
	Society	160	19	25	51	65
Task type	Instructional	21	21	0	0	0
	Narrative	28	0	28	0	0
	Descriptive	28	0	0	28	0
	Argumentative	132	19	25	23	65
Code complexity	Low	72	0	28	28	16
	High	137	40	25	23	49
Cognitive complexity	Low	65	21	28	0	16
	High	144	19	25	51	49
Reasoning	Low	21	21	0	0	0
	Medium	56	0	28	28	0
	High	132	19	25	23	65
Referenced elements	Few	49	21	28	0	0
	Many	28	0	0	28	0
	Open	132	19	25	23	65
Perspective	Own	28	0	28	0	0
	Other	21	21	0	0	0
	Own & others	160	19	25	51	65

Table 7.2.: Distribution of task properties on full *Falko Georgetown L2* corpus (including only variable task factors).

Task	Test Level	Time	Expected Words	Genre	Audience	Formality	Theme	Task Type	Code Complexity	Cognitive Complexity	Shared Context	Reasoning	Referenced Elements	Perspective
Going swimming	A1	45 Min.	30	Email	Friend	Informal	Demand	Descriptive	High	Low	T & T	Low	Few	Own
Apartment search	A1	45 Min.	30	Email	Friend	Informal	Demand	Instructive	High	Low	H & N	Low	Few	Own
Child birth	A1	45 Min.	30	Letter	Friend	Informal	Small talk	Descriptive	High	Low	H & N	Low	Few	Own
Ticket offer	A2	50 Min.	40	Letter	Friend	Informal	Demand	Instructional	High	Low	H & N	Medium	Few	Own
Pet sitting	A2	50 Min.	40	Email	Friend	Informal	Demand	Instructional	High	Low	H & N	Medium	Few	Own
Housing office	A2	50 Min.	40	Letter	Agency	Informal	Demand	Descriptive	High	Low	H & N	Low	Few	Own
Announce visit	B1	30 Min.	128	Letter	Friend	Informal	Small talk	Narrative	Low	Low	H & N	Low	Many	Own
Happy birthday	B1	30 Min.	128	Letter	Friend	Informal	Small talk	Narrative	Low	Low	H & N	Medium	Many	Own
Happy new year	B1	30 Min.	128	Letter	Friend	Informal	Small talk	Narrative	Low	Low	T & T	Medium	Many	Own
Application	B2	30 Min.	150	Letter	Agency	Formal	Profession	Argumentative	High	Medium	H & N	Medium	Few	Own
Work complaint	B2	30 Min.	150	Letter	Agency	Formal	Profession	Argumentative	High	Medium	T & T	Medium	Many	Own
Information request	B2	30 Min.	150	Letter	Agency	Formal	Profession	Descriptive	High	Low	H & N	Medium	Few	Own
Housing situation	C1	60 Min.	200	Essay	Public	Formal	Society	Descriptive	High	High	H & N	Medium	Open	Own & others
Learning German	C1	60 Min.	200	Essay	Public	Formal	Society	Argumentative	High	High	H & N	High	Open	Own & others
Traditions & Assimilation	C1	60 Min.	200	Essay	Public	Formal	Society	Argumentative	High	High	H & N	High	Open	Own & others

Table 7.3.: Properties and task factors annotated for *Merlin* tasks.

Task	Test Level	Audience	Genre	Formality	Theme	Task Type	Code Complexity	Cognitive Complexity	Shared Context	Reasoning	Referenced Elements	Perspective
Write a letter	1	Friend	Letter	Informal	Small talk	Instructional	High	Low	T & T	Low	Few	Own
Continue a novel	2	Public	Novel	Informal	Mystery	Narrative	Low	Low	T & T	Medium	Few	Other
Write an article	3	Public	Article	Formal	Society	Descriptive	Low	High	T & T	Medium	Many	Own & others
Write a speech	4	Public	Speech	Formal	Society	Argumentative	Low	Low	T & T	High	Open	Own & others
Book review	1-4	Public	Review	Formal	Society	Argumentative	High	High	T & T	High	Open	Own & others

Table 7.4.: Properties and task factors annotated for *Falko Georgetown L2* tasks.

Part IV.

Analyzing Linguistic Complexity

8. Feature Collection

All complexity analyses conducted throughout this thesis are based on a pool of 398 measures of complexity measuring lexical, grammatical and morphological aspects as well as textual cohesion, cognitive load, and indices of academic language. All these measures are allocated to one of the following four main categories: i) measures of language use, ii) measures of discourse and encoding of meaning, iii) measures of human language processing, and iv) measures of the linguistic system. This grouping by Meurers 2017 partially relates to the more traditional division of complexity measures into measures of lexical, syntactic, and morphological complexity and measures of textual cohesion. However, it is more precise than the traditional nomenclature in terms of the underlying assumptions made, when defining these complexity measures. For example, it distinguishes measures of lexical frequency, which assess language use, and measures of lexical diversity, which is a characteristic of the linguistic system. Most taxonomies group these measures together as features of linguistic complexity, although the reasons to assume these feature groups assess complexity are fundamentally different. Also, it is more adaptive, as it easily accommodates for example measures of cognitive load.

For the remainder of this chapter, all complexity measures are introduced within their respective feature groups: Section 8.1 elaborates on measures of language use, Section 8.2 discusses measures of discourse and encoding of meaning, and Section 8.3 measures of human language processing. Finally, Section 8.4 elaborates on lexical complexity (Section 8.4.1), sentential and phrasal complexity (Section 8.4.2), and morphological complexity (Section 8.4.3).

8.1. Measures of Language Use

Measures of language use are predominantly employed in psychological and corpus-linguistic research. They include, for example, word frequencies obtained in representative language samples and age of acquisition measures, which are typically

assessed by post-hoc self-assessment of native speakers (Birchenough, Davies & Connelly 2016; Brysbaert et al. 2011; Schröder et al. 2012). There are also some language use approaches to grammatical constructions, such as measures of phraseological sophistication from Paquot 2017, which assess the frequency of phraseological units or point-wise mutual information of linguistic constructions (Bestgen & Granger 2014; Paquot 2017).

The feature collection includes a series of lexical frequency measures, which assess ratios of lexical type and lemma frequencies per lexical types in a document. These ratios are calculated in four different variants: i) a simple variant that is calculated as described above; ii) a variant that works on log frequencies rather than the raw frequencies obtained in variant i); iii) a variant that is based on annotated lexical types or lemmas instead of simple types or lemmas, i.e. that only considers frequencies of a lexical type or lemma given its Part of Speech (POS); and iv) a variant that works with the log frequencies instead of the raw frequencies obtained from variant iii). Also, it includes measures that are based on log lexical type occurrences in frequency bands, that are defined independently for each word data base, see Hancke 2013: 31f for a more elaborate account. Finally, the ratio of lexical types not found in a data base is also measured.

Three different frequency data bases are used for the calculation of frequency measures in the complexity code, each of which provide some unique aspect that the other two data bases lack: First, the *SUBTLEX-DE* frequency data base by Brysbaert et al. 2011, which is based on movie and television series subtitles. Brysbaert & New 2009 found subtitle frequencies to serve as the better predictors of word recognition times for English, than for example book- or newspaper-based frequencies. Brysbaert et al. 2011 report similar results for German. The subtitle data was spell-checked automatically with the Igerman98 spell checker.[1] It contains frequencies for 190,500 types and 209,936,190 tokens. These were extracted from subtitles of 4,610 movies and series from `www.opensubtitles.org` (ibid.: 16). Second, the code also includes frequencies from the *Google Books 2000-2009* corpus, which were included in the *SUBTLEX-DE* data, because due to their considerably larger token basis of 30.1 billion words (ibid.: 16) Google Books provides more informed information on rare words (ibid.: 10,34). Third, the *dlexDB* data base (Heister et al. 2011) was used, which is based on the core corpus of the Digitales Wörterbuch der deutschen Sprache (DWDS) (cf. Geyken 2007).[2] The frequencies are

[1] Cf. `http://freecode.com/projects/igerman98`.
[2] `http://dlexdb.de`.

based on 2,224,542 types and 122,816,010 tokens from novels, news articles, prose, use texts, and transcriptions of spoken language from the 20th century (Heister et al. 2011: 11). *dlexDB* neither has the largest data support, like *Google Books 2000*, nor the most natural occurring data, like *SUBTLEX-DE*. However, it is the only data base that uses linguistically informed frequencies, such as lemma and POS-based frequencies. Hence, it is the only data base for which (annotated) lemma and annotated type ratios may be computed. Also, despite the fact that Brysbaert et al. 2011 report *SUBTLEX-DE* to explain more variance in reaction times than *dlexDB*, the differences in performance were not all significant.

Aside from frequency measures, the code features several measures that approximate of age of acquisition: While age of acquisition is usually measured by means of self-assessment tests, where adult subjects estimate, when they acquired a specific word in their native language, these features are based on an approximation that relies on a normalized L1 corpus of children's writing, the KCT corpus by Lavalley, Berkling & Stüker 2015. The corpus consists of 1701 texts written in free writing tasks from native speakers of German aged 6 to 17. Using this data, the maximal and average minimal age of acquisition of the lexical types in each text is calculated. Note that this corpus is also used to assess the same frequency measures as noted above, but specifically targeting children's L1, thus creating some hybrid measures of age of acquisition and lexical frequency. While these measures may be argued to be only very coarse approximation of language acquisition, the measures have proven to be informative in L2 studies, see for example Study 2 in Chapter 14.

8.2. Measures of Discourse and the Encoding of Meaning

Discourse and the encoding of meaning, too, is predominantly a psychological and psycho-linguistic construct and is commonly measured in terms of Propositional Idea Density (PID) (Brown et al. 2008) or lexical concreteness. This domain also includes measures of textual cohesion, such as the usage of connectives and textual co-reference (Crossley, Kyle & McNamara 2015; Galasso 2014; Graesser et al. 2004). Textual *cohesion* is determined by the amount of linguistic items used to link propositions or idea units. These are assumed to relate to the *coherence* of a text, which refers to the mental representation or understanding a reader derives from a text (Louwerse et al. 2004).

The largest quantity of measures from this research background in the complexity code was implemented by Galasso 2014 and is based in the *Coh-Metrix* system

(Graesser et al. 2004; McNamara et al. 2014), which is probably the most well known tool for the analysis of discourse and textual cohesion. These measures may be divided into two groups: First, *co-referential cohesion* in terms of terms of noun, argument, stem, and content word overlap. Following *Coh-Metrix*, each of these measures is assessed i) assuming a *local* notion of overlap, which only considers adjacent sentence pairs, and ii) assuming a *global* notion of overlap, which considers all possible sentence pair combinations throughout the text. Nouns are considered to overlap, if they share lemma, number, and case.[3] For argument and stem overlap, shared phi features are not required. While these features are binary, the system uses proper counts for content word overlap. Unlike McNamara et al. 2014, Galasso 2014 decided against requiring shared number and case for content word overlap counts, as this proved to be too restrictive for the German morphology.

Second, the system also counts causal and conditional, logical, temporal, additive, as well as adversative and concessive connectives per 1,000 words, thereby considering a subset of connective types used by McNamara et al. 2014. Single- and multi-word connectives are identified using lists of connectives provided by Duden (Gr) 2009.[4] In order to prevent overestimating the number of connectives, the system also requires potential connectives to be tagged as either conjunctions, adverbs, pronominal adverbs or adpositions. Ambiguities between types of connectives are not resolved, instead, connectives like *wenn*, which might be conditional (*if*) or temporal (*when*), are counted once for each list. Additional to these measures of overtly realized connectives by Galasso 2014, the system also contains a small set of measures targeting different types of German conditional clauses by Weiß 2015, which were designed to capture the most common realizations of German conditional clauses, to allow for a variationalistic comparison.

Additional to the *Coh-Metrix* measures, Galasso 2014 also adopted referring expression and transition measures from Barzilay & Lapata 2008; Todirascu et al. 2013, since these introduce cohesive links to entities across sentences, which increases textual coherence. These measures include ratios of various types of pronouns per sentence and text as well as a ratio of proper names per text. The transition measures assess changes in grammatical function of same entity across

[3] Enforcing case overlap is a restriction introduced by Galasso 2014 in order to account for German case inflection. McNamara et al. 2014 does not require this additional assumption, since case is not morphologically realized in English.

[4] While this set of measures was originally introduced to the system by Galasso 2014, the lists of connectives were updated by Maria Chinkina, who also implemented the identification of multi-word connectives.

sentences. Galasso 2014 includes for transitions all noun overlaps ignoring phi features. Following Barzilay & Lapata 2008 she assesses four grammatical functions: subject, object, other complement, and absent.

The complexity code also assesses encoding of meaning in terms of PID as implemented in Brown et al. 2008's Computational Idea Density Rater (CPIDR) system, which measures the ratio of propositions to non-punctuation tokens. ibid.'s PID is an operationalization of Kintsch 1974's *Propositional theory*, which understands the mental representation of a text as a list of propositions, which is constantly augmented during reading. It is important to note, that while ibid. assumes roughly the same notion of proposition as the from the field of semantics commonly known *logical proposition*, his definition of propositions disregards phi-features and common nouns: Following ibid., a proposition is in principle assumed to be i) a verb with a fully saturated argument structure, or ii) a phrasal adjunct. However, there is a series of exceptions to this general rule (ibid.), which are taken into account Brown et al. 2008 as well as in the presented system: Both systems count conjunctions, numbers, determiners, prepositions, adjectives, predeterminers, possessives, adverbs, verbs, and interrogative pronouns as propositions unless they are articles, the non-initial part of a multi-word connectives and interrogatives, non-sentence initial modal verbs, non-attributive cardinals, linking or auxiliary verbs.

8.3. Measures of Human Language Processing

Human language processing views complexity from the perspective of cognitive science, psycho-linguistics and information theory. It evaluates *cognitive complexity* as processing costs of linguistic structures, often measured in processing time, in terms of, for example, surprisal or cognitive load. The complexity code implements cognitive complexity in two fashions: i) in terms of integration costs based on Gibson 2000's Dependency Locality Theory (DLT), and ii) based on distances between positions in the topological field.

The DLT is a prominent theory of sentence processing proposed that was originally proposed by ibid. It is based on two key assumptions:

1. Human sentence processing includes two processes, which draw from attentional resources: i) the *storage* of incomplete discourse structures, i.e. incomplete dependent clauses; and ii) the *integration* of discourse referents into the currently build discourse structure.

2. Human sentence processing is sensitive to the *locality* or distance between two discourse elements that are to be integrated, insofar as larger distances increase their integration cost.

Storage costs consume memory units based on the number of elements, that are minimally required to build a complete sentence, i.e. when a sentence starts with a determiner, at that point the storage cost is 2, because at least a noun and an intransitive verb are required to form a complete sentence (Gibson 2000: 114f). Integration costs are caused by i) *structural integration*, which links arguments and adjuncts to the maximal projection of the head that governs them and antecedents to their referents; and by ii) *discourse integration*, which links discourse referents to the discourse structure. Structural integration is assumed to be sensitive to the distance between discourse elements, because when integrating an element to a structure, it is required to retrieve all intervening discourse elements from memory. Hence, the structural integration cost of a discourse referent into a discourse structure includes the structural integration cost of all interim discourse referents, that were loaded into memory since the reference element to be integrated was last activated (ibid.: 103f). For simplicity, ibid. assumes a linear relationship between the number of intervening discourse referents and the integration cost for a discourse referent. Discourse integration is assumed to be influenced by the accessibility of discourse referents, i.e. whether they are focused or not, the simplified DLT proposed by ibid. does not account for these cost differences. Also, it only considers the heads of nouns and verbs as discourse referents of either discourse objects or discourse events, based on the criterion that discourse referents need to have a spatio-temporal location (ibid.: 103).

Shain et al. 2016 propose three types of modifications to augment integration costs for this simplified DLT based on dependency parses:

DLT-V "Verbs are more expensive. Non-finite verbs receive a cost of 1 (instead of 0) and finite verbs receive a cost of 2 (instead of 1)" (ibid.: 51). This is motivated with the assumptions that discourse events might be cognitively more complex than discourse objects and that embedded discourse events, i.e. non-finite verbs, are costly, too.

DLT-C "Coordination is less expensive. Dependencies out of coordinate structures skip preceding conjuncts in the calculation of distance, and dependencies with intervening coordinate structures assign that structure a weight equal to that of its heaviest conjunct" (ibid.: 51). This is based on the assumption that

the coordination of discourse referents is trivial in terms of cognitive pressing costs. Hence, coordinated referents should only receive one collective count for the coordinated set of referents.

DLT-M "Exclude modifier dependencies. Dependencies to preceding modifiers are ignored" (Shain et al. 2016: 51). This condition is intended "to avoid excessive 'double-counting' of material intervening in long modifier dependencies" (ibid.: 51).

As ibid. point out, these conditions may be combined, which leads to overall eight possible implementation variants of the DLT. In the feature collection, measures of integration cost are included in terms of i) maximal total integration cost; ii) mean total integration cost at finite verbs; and iii) number of adjacent high integration cost areas. The second measure is based on the fact, that total integration costs are highest at finite verbs, because these are the only positions where discourse integration costs are non-zero. The third measure was designed to assess nested discourse structures. The required threshold to qualify discourse costs as 'high' is currently set to 3, because relative clauses with transitive verbs have total integration costs of 3, i.e. the measure will be sensitive to relative clauses occurring in the middle field. However, for future work it would be desirable to identify a threshold that is supported by more evidence. These three features are implemented in all eight variants that result from applying combinations of the conditions that are suggested by ibid.

The feature collection also includes two measures that approximate distances between verbs and arguments in terms of linguistic units, namels syllables, instead of in terms of semantic units, i.e. discourse referents:[5] i) the average number of syllables per middle field for middle fields in clauses where the right sentence bracket contains at least one element; and ii) the average number of syllables between the first argument of a verb and the finite verb, see Weiß 2015 for more details on these measures. Although these measures operate fully within concepts of theoretical linguistics, they are considered to constitute measures of human language processing, because as with the DLT, the key assumption underlying these measures is, that distance between discourse elements increases complexity.

[5]See von der Brück & Hartrumpf 2007: 3 for a similar distinction.

8.4. Measures of the Linguistic System

Measures of the linguistic system are based on assumptions derived from theoretical linguistics. Since the linguistic system consists of a variety of domains that constitute distinct fields in linguistics, naturally, these measures are also further distinguished based on the linguistic domains from which they are derived. The system assesses features from the three most commonly measured domains of i) lexical (and semantic) complexity; ii) syntactic or grammatical complexity; and iii) morphological complexity. The following subsections elaborate on each of these types of measures of the linguistic system. However, note that this is not an exhaustive set of domains, for which complexity measures of the linguistic system may be derived. Other potential domains are, for example, phonological and pragmatic complexity (Bulté & Housen 2014: 44). However, these are less commonly investigated in CAF research and currently not part of the system used throughout this thesis.

8.4.1. Lexical Complexity

Lexical complexity is among the most commonly investigated complexity measures (Bulté & Housen 2014; Wolfe-Quintero, Inagaki & Kim 1998). It is typically associated with vocabulary range (lexical density and variation) and size (lexical sophistication) (Wolfe-Quintero, Inagaki & Kim 1998: 101), but also with lexical relatedness, and frequency. The latter assesses vocabulary use, though, whereas the other types of measures are based on assumptions about the linguistic system (cf. below). So although it is commonly listed as an aspect of lexical complexity, in this thesis lexical frequency is kept apart from these other measures targeting the lexicon, see Section 8.1 for more details. Also, lexical relatedness is also sometimes referred to as semantic complexity (Bulté & Housen 2014; von der Brück, Hartrumpf & Helbig 2008).

In this thesis, lexical complexity is assessed through various implementations of each of these sub-domains: *Lexical diversity* estimates the range and variety of the full vocabulary by assessing the ratio of types and tokens through various measures. For this, the system measures raw type token ratios as well as various transformations that are designed to make the measure less sensitive to text length, such as squared, corrected, or bilogarithmic type token ratio, Uber index, or Yule's K, or considerably more elaborate formulas such as HD-D and MTLD. These measures are based on McCarthy & Jarvis 2007 and are elaborated on in more detail

in Hancke 2013: 28f. *Lexical density and variation* is conceptually a suit of type of type token ratios that target specifically lexical words. As a measure, lexical density is defined as the number of lexical tokens per tokens (Lu 2011) and lexical variation as the number of lexical types per lexical token. Both measures may be further subdivided into the variation or density of specific categories of lexical words, such as verb, noun, adverb, or adjective variation and density. For verb variation, the system also includes several variants of mathematical transformations applied to type token ratios, such as squared or corrected verb variation. For a comprehensive list and description of all measures of lexical density and variation included in the system, please see Hancke 2013: 29f.

While these measures assess the elaborateness and variedness of the exhibited vocabulary, there is also a straight forward implementation of the relatedness of vocabulary: *Lexical relatedness* assesses the amount of semantic relations between words, such as hyperonymy, hyponymie, synonymy, and antonymy, following Crossley & McNamara 2009. The system assesses these measures using GermaNet (see Section 9.2 below). For a more elaborate account on the detailed measures, please see Hancke 2013: 33f.

8.4.2. Syntactic Complexity

Syntactic complexity, which is also often referred to as grammatical complexity, measures the variation and sophistication or elaborateness of grammatical constructions (Foster & Skehan 1996: 303; Wolfe-Quintero, Inagaki & Kim 1998: 69). In recent years, the more elaborate distinction between i) sentential or clausal; and ii) phrasal organization of syntactic or grammatical complexity was introduced (Bulté & Housen 2014; Kyle 2016). Clausal complexity is the more traditional strand of features, which is often associated with measures of clausal subordnation and the use of syndetic and asyndetic constructions, but conceptually, it may also include measures of varying degrees of clausal integration, such as appositions, parentheses, non-integrated sentence constructions. These are, however, less often implemented due to technical limitations, although they are well motivated from a theoretical linguistic perspective. Phrasal organization only gained increasing interest in recent years. It is often assessed in terms of phrasal modification, but conceptually, this may also include measures of valency and argument structures as well as phrasal coordination.

In the presented system, *clausal complexity* is assessed in terms of i) types of subordination and ii) clausal modification. Types of subordination are measured

in terms of clauses, complex t-units, or t-units per graphematic sentences, clauses per t-units, dependent clauses per t-units or graphematic sentences, etc. Sentential modification is assessed with measures of complex noun phrases, verb phrases, or modifiers per clauses, t-units or graphematic sentences and by the numbers of various phrase types per sentences, clauses, or t-units. For elaborate descriptions of these measures, please see Hancke 2013: 39ff.

Phrasal complexity is assessed by means of modifier ratios as well as two sets of measures that specifically target i) verbal modification and verb clusters; and ii) complex noun phrases. Modifier ratios include measures of phrasal coordination as well as the number of words per noun phrases, verb phrases, and prepositional phrases. Measures of verbal modification include measures of verb modification by means of adjectival and adverbal modifiers, participle modifiers, and prepositional modifiers as well as coverage of modifier types. Verb clusters are assessed by means of average verb cluster size and its standard deviation as well as by types of verb cluster types, i.e. the ratios of clusters were a main verb governs another verb, an auxiliary verb governs another verb, or a modal verb governs another verb. Measures of noun modification assess the complete topology of the German complex noun phrase by measuring ratios of determines, possessive noun attributes, prenominal and postnominal attributes, attributive participles, comparative noun modifiers, and clausal noun modifiers, as well as the coverage of noun modifiers.

Also, a set of distinct grammatical constructions is assessed, that is to be located somewhere between the typical definitions of sentential and phrasal complexity and includes measures of the linguistic system, that are commonly associated written *academic language*. These include a noun to verb ratio as well as measures of all periphrastic tenses (future 1 and 2, present and past perfect) and a series of passive measures, including measures of quasi passive constructions. It also includes a measure of non-subject prefields. For more details on these, please see Weiß 2015. Finally, this set of measures also includes a series of *superficial length measures*, that are often discussed to be measures of fluency in systems that measure multiple CAF dimensions, but associated with syntactic complexity in systems focusing on complexity, because they are highly sensitive to syntactic modifications. These measures are the number of sentences, the number of words per text, and the average length of graphematic sentences, t-units, and clauses in words.

8.4.3. Morphological Complexity

Morphological complexity is traditionally less commonly measured than lexical and syntactic complexity, since morphology has a minor role in the English language, on which the majority of studies focuses (Pallotti 2015: 121;Bulté & Housen 2014: 44). However, in recent years, several studies showed that for synthetic languages, such as German, morphological complexity measures are highly informative (François & Fairon 2012; Hancke, Vajjala & Meurers 2012; von der Brück, Hartrumpf & Helbig 2008). For German, morphological complexity may target three aspects of the morphological system: i) inflection; ii) derivation; and iii) composition. *Inflectional complexity* may be assessed by measuring the full inflectional system of German in terms of person, number, mode, tense, case, and finiteness. This is, in fact, a rare situation, since most other aspects of the linguistic system cannot be quantified as exhaustively. *Derivational complexity* is less straight-forward to pin down and the complexity system assesses it in terms of 24 native and foreign nominalization suffixes and an overall ratio for derived nouns. It also assesses the amount of deverbal nominalizations, which is operationalized as nouns ending in *ung* or *en*, if they have neuter case. *compositional complexity* is assessed in terms of the compound nouns to noun ratio and average compound depth. Finally, this set of measures also entails two shallow, global measures, that do not assess any particular structure of the morphological system, but are prone to be highly sensitive to any changes related to the morphological structure. These are word length measures in terms of characters and syllables per word. For a more detailed description of these measures, please see Hancke 2013; Weiß 2015.

9. Complexity Analysis System

This chapter introduces technical aspects of the system, that is used in this thesis to measure overall 398 measure of linguistic elaborateness and variedness of morphological, lexical, phrasal, clausal, sentential, and discourse domains, as well as markers of German academic language, such as deagentivation strategies, and empirical correlates of cognitive processing load, such as indices of the DLT, see Chapter 8 for a comprehensive overview. To my knowledge, this is the most comprehensive collection of such measures for German, which is currently used in research. This system is based on work by Galasso 2014; Hancke 2013; Hancke, Vajjala & Meurers 2012; Weiß 2015 and has been continuously developed, expanded and restructured augmented by several programmers, including the author of this thesis.

The next section introduces the full pipeline for feature extraction. Section 9.2 provides a list of all libraries and data bases on which the system relies. Then, Section 9.3 briefly elaborates on the definitions of linguistic units employed throughout the system. The chapter closes with an evaluation of the system's performance on standard and non-standard data in Section 9.4.

9.1. Pipeline

The complexity pipeline written in Java (1.8). It performs a three-stepped sequential analysis of input data: i) An elaborate NLP tool chain (cf. Section 9.2) annotates linguistic information on plain text data, that is required as input. ii) Overall 324 linguistic constructions are extracted and counted based in these linguistic annotations and additional language resources (cf. Section 9.2). iii) 398 complexity formluae and ratios are calculated based on the frequencies of the obtained linguistic constructions.[1] All measures are written to a matrix of documents and complexity

[1]The nomenclature of complexity features differs widely among studies: While Wolfe-Quintero, Inagaki & Kim 1998 refer to them as *measures*, Housen, Vedder & Kuiken 2012 prefer the term *index*. This thesis uses the terms *complexity measures* and *complexity features* interchangeably to refer to frequency-based *indices* or *ratios*: Ratios are calculated by simply dividing of two frequencies.

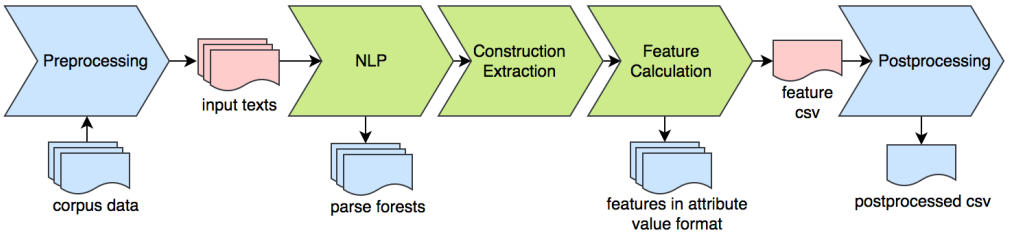


Figure 9.1.: System pipeline from plain text corpus to feature analysis.

measures, which is saved in comma separated value (CSV) format for further statistical analysis. This process is visualized in Figure 9.1. Input and output may require further pre- or respectively post-processing, such as initial normalization of the data or augmentation of the output file with additional meta information. As these steps may be highly corpus dependent, they are not part of the core pipeline.

9.2. Resources

The automatic linguistic annotation of the plain text input is the most vital step of the analysis and dictates quantity and quality of the complexity measures, that may be extracted. The analysis involves the following steps: tokenization and sentence segmentation using OpenNLP (1.6.0)[2], POS tagging, lemmatization, morphological analysis and dependency parsing based on the Mate tools (3.6.0) by Bohnet & Nivre 2012, constituency parsing with the Stanford PCFG parser (3.6.0) using Mate POS tags, and topological field parsing using the Berkeley parser (1.7.0) Petrov & Klein 2007. The respective default models for German contained in these packages are used for all these processing steps, except for the last one. For topological field parsing, the system relies on the model trained by Ramon Ziai.[3] Table 9.1 gives an overview over all NLP components involved in the analysis.

This is, for example, necessary to normalize for length differences (e.g. ratio of nouns per token), to obtain the percentage of a certain variant for a given variable (e.g. ratio of modal verbs per verbs), or to simply relate different types of lexical material to each other (e.g. noun to verb ratio). Formulas transform frequencies of linguistic constructions beyond mere division, cf. Wolfe-Quintero, Inagaki & Kim 1998: 10. Examples for this are many of the various types of type token ratios, which are designed in order to account best for text length. So, while the classical type token ratio is a proper ratio, obtained by dividing the frequency of types by the frequency of tokens, the squared type token ratio is an index following the formula $\sqrt{\frac{freq.\ types}{freq.\ tokens}}$.

[2] `https://opennlp.apache.org`.

[3] The model was obtained via personal correspondence.

Task	Component	Version	Model
Tokenization and sentence segmentation	OpenNLP	1.6.0	default
POS tagging Lemmatization Morphological analysis Dependeny parsing	Mate tools	3.6.0	default
Compound splitting	JWordSplitter	3.4.0	default
Constituency parsing	Stanford PCFG parser	3.6.0	default
Topological field parsing	Berkeley parser	1.7.0	cf. Ramon Ziai

Table 9.1.: NLP components used in the complexity analysis pipeline.

Based on these annotations, frequencies of linguistic constructions such as t-units, noun modifiers, or accusative case markings are assessed, as well as word frequencies. This step relies on additional resources, such as Tregex (3.6.0) (Levy & Andrew 2006) for tree pattern matching, lexical information from GermaNet (9.0.1) (Henrich & Hinrichs 2010), word frequencies form dlexDB (Heister et al. 2011) and SUBTLEX-DE & Google Books 2000 (Brysbaert et al. 2011),[4] compound splitting with JWordSplitter 3.4.0,[5] and lists of multi-word connectives from Duden (Gr) 2009. These frequencies are either normalized in ratios or computed to complexity ratios and formulae.

9.3. Operationalization of Linguistic Units

A clear definition of the linguistic units used to measure complexity is a crucial aspect of allowing for comparability across studies (Bulté & Housen 2014; Foster, Tonkyn & Wigglesworth 2000; Hancke 2013; Housen, Vedder & Kuiken 2012), because while some units may be clear, others are open to varying operationalizations. This present system relies on the definition of linguistic units that was established in Hancke 2013: 12ff. However, the respective units are briefly repeated here, too, since the definition of linguistic units is so crucial for the interpretation of results.

Parts of speech are assigned based on STTS POS tags, cf. Projekt Tiger 2003: 121 or Hancke 2013: 12. Since this procedure is straight-forward, it is not elaborated on here. The following linguistic units are open to diverging operationalizations,

[4] See Section 8.1 for more information on word frequencies.
[5] `http://www.danielnaber.de/jwordsplitter/`.

though, and require more elaborate discussion:

Clauses are all maximal projections of finite verbs and elliptical constructions with sentential status (i.e. all sub-trees tagged with *S*), as well as *to* infinitives that have a sentential status (*satzwertige zu Infinitive*). For details, please see Hancke 2013: 12ff.

Complex t-units are t-units that include subordinate clauses.

Conjunctional clauses are all dependent clauses that are introduced by a subordinating conjunction such as *dass*, *weil*, or *wenn*. For details, please see ibid.: 12ff.

Dependent clauses with conjunction are all conjunctional clauses, but also interrogative and relative clauses. Dependent clauses without conjunction are mostly dependent main clauses, such as *Ich weiß, es ist spät*. For details, please see ibid.: 12ff.

(Graphematic) sentences are strings of at least one token that are ended by sentence ending punctuation marks: *!*, *.*, *?*. There is a broad discussion on alternative sentence definitions, see for example Schmidt 2016 for a more elaborate theoretical account. However, since sentences are identified by sentence segmentation tools, which are primarily based on punctuation, sentences are always defined as graphematic sentences. For details, please see Hancke 2013: 12ff.

Half modals are *haben*, *sein*, *scheinen*, *drohen*, *versprechen*, if they govern an infinitive with *zu* (§101 Duden (Gr) 2009: 101), e.g. *ist zu machen*, *droht zu schneien*. For details, please see Weiß 2015: 32f.

Lexical words are all nouns, adjectives, adverbs, foreign words, numbers, main verbs, and modal verbs. Note that there is an ongoing discussion on whether modals actually qualify as lexical words (Reis 2001), hence there is also a subset of **lexical words excluding modals** employed throughout the system. For details, please see Hancke 2013: 12ff.

Quasi passives are *bekommen*, *erhalten* or *kriegen* if they govern a past participle (§179 Duden (Gr) 2009: 147f), e.g. *bekommt gemacht*, *kriegt eröffnet*.

T-units are "one main clause plus any subordinate clause or non clausal structure that is attached to or embedded in it" (Hunt 1970: 4). For details, please see Weiß 2015: 33.

9.4. Evaluation

Up until now, no systematic, internal evaluation of the entire collection of 398 complexity features has been carried out to confirm the validity of the individual measures on standard and non-standard data. However, the performance of all measures can be approximated to varying degrees of certainty without such an evaluation: None of the measures of language use is prone to errors, as they rely purely on look ups in a linguistic data base. Furthermore, most linguistically motivated measures do not introduce additional error potential beyond failures in NLP. Thus, their robustness may be equated with the robustness of the NLP tools involved in their calculation, see the respective tool descriptions for details on this, see Alexopoulou et al. 2017: 8 for a similar argument. Most linguistic measures, that do introduce additional error potential, were evaluated on a small set of normalized L1 data in Weiß 2015 and either received high scores of precision and recall or were excluded from the system. Measures of discourse and meaning and measures of human language processing are the only features, that currently lack sufficient, systematic internal validation and do not allow for a sufficient approximation of their robustness, as their extraction introduces error potential beyond the NLP components they are based on. However, these measures were validated externally by their high performance on estimating L2 proficiency and text readability (Galasso 2014; Weiß 2015).

10. Measuring Linguistic Complexity on Learner Corpora

This chapter presents a descriptive cross-corpus analysis of changes in complexity across proficiency levels in the full data sets of *Merlin* and *Falko Georgetown*. The analysis is based on complexity features, that were extracted using the system described in chapter 9. For reasons of space, only a representative selection of about 100 features may be discussed here to provide a cursory overview. These measures were chosen and grouped based on theoretical considerations. However, plots for all 398 complexity measures may be found in the online supplementary material provided for this thesis at `http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/supplementary-material/complexity-plots/`.

For both corpora, the complexity analysis system described in Chapter 9 was used, to extract complexity measures automatically from the close transcriptions of the learner texts. Whether to carry out an automatic complexity analysis on close transcriptions of learner data or on corrected target hypothesis, is a crucial question, without a straight forward solution: In favor of a target hypothesis speaks, that NLP tools are typically trained on standardized L1 data, and thus more likely to perform adequately on standardized data, than on close transcriptions of L2 data (Bestgen & Granger 2014). This holds in particular for more advanced syntactic analyses carried out by parsers. Also, measures operating on the surface level are considerably impaired by non-standard data, such as word frequency data base or word list based measures, as they are used in the analysis system for the assessment of language use and connectives. However, it has been reasoned, that the analysis of target hypothesis instead of close transcriptions of learner data may fail to capture unique characteristics of the L2 inter-language system under investigation (Thomas 1994: 328). Furthermore, target hypotheses are likely to follow different annotation rules across corpora. This might impair comparability of results in cross-corpus studies. Preliminary experiments on *Merlin* indicated, that analyses carried out on close transcriptions yielded reasonable results, that were comparable with the

analyses conducted on the target hypotheses. Hence, it was decided, that the performance of the complexity analysis system was not sufficiently impaired by using the close transcriptions, to justify the disadvantage of analyzing an artificial standard version rather than the proper inter-language system.

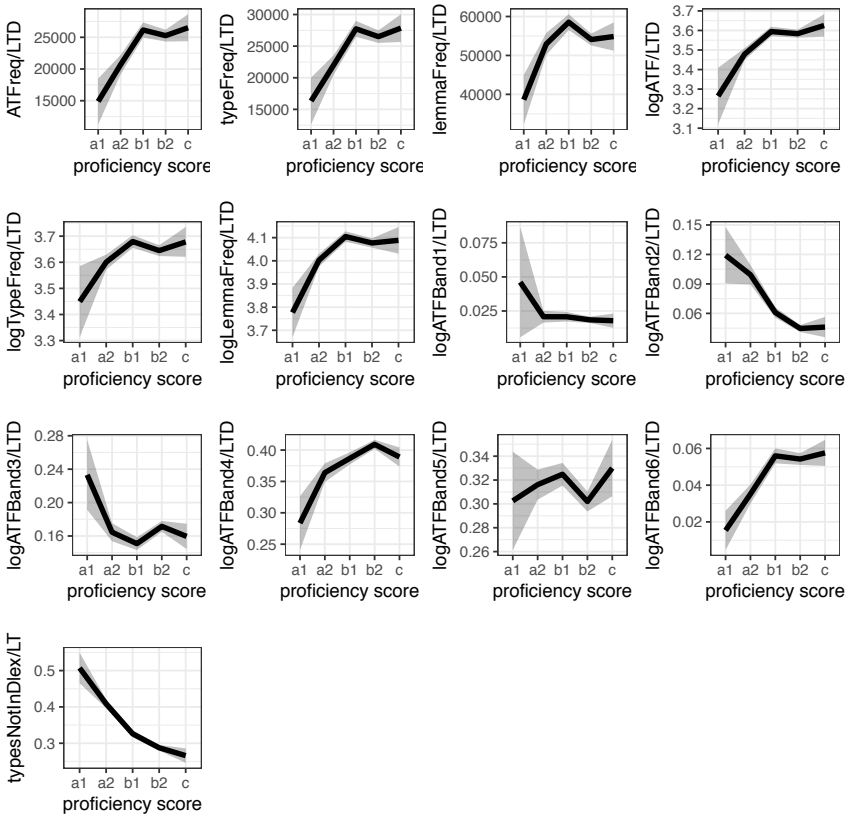
The following four sections present exemplary complexity measures of language use (Section 10.1), discourse and meaning (Section 10.2), and human language processing (Section 10.3). These are followed by three separate sections for theoretical linguistic measures of the lexical, syntactic-grammatical, and morphological domain in Section 10.4, 10.5, and 10.6. The chapter closes with a summarizing discussion of the findings in Section 10.7.

10.1. Observations for Language Use

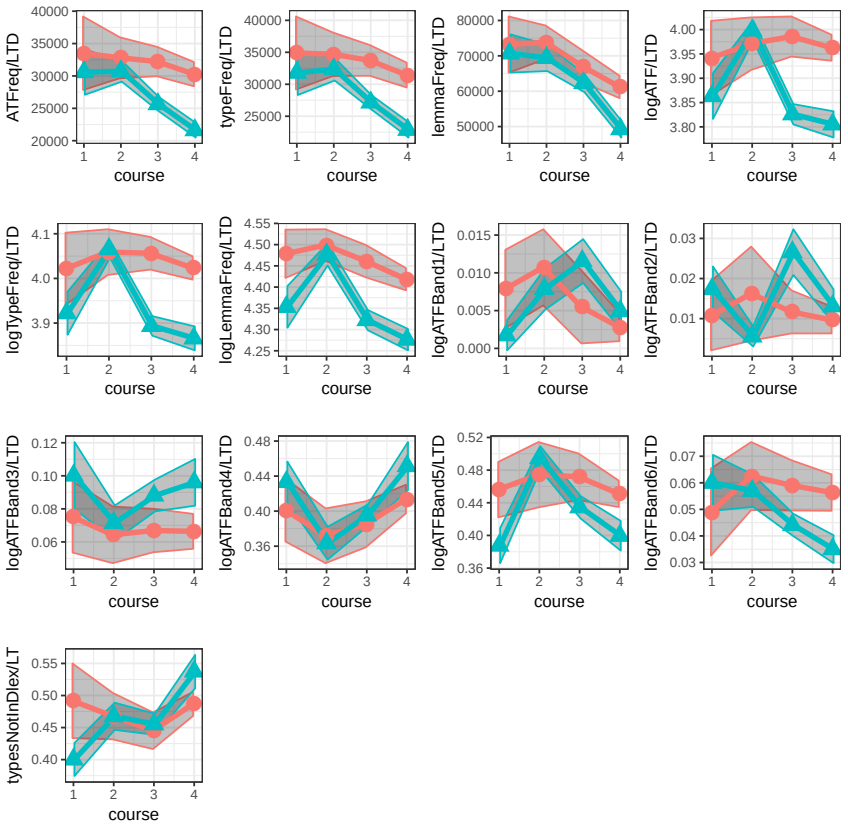
Frequencies from *dlexDB* were chosen to represent language use, because they include linguistically informed frequencies for annotated types and lemmas and thus provide a broader picture. Figure 10.1 shows the measures across proficiency levels extracted from *Merlin* and *Falko Georgetown L2*. Here and across all following plots, the grey shades indicate 95% confidence intervals and for the *Falko Georgetown L2* data, curricular tasks and book reviews are separated to allow to differentiate between general trends in the data followed by both task groups and potentially task-based developments that only show in curricular tasks. Also, individual plots within figures are referred to by their panel position, which is assigned with increasing cardinality from left to right and from top to bottom.[1]

On the *Merlin* data, frequencies of annotated lexical types (ATF), lexical types and lemmas per lexical type found in *dlexDB* (LTD) increase with increasing proficiency (panels 1 to 6). As for log annotated type frequency bands, it may be seen that words from the first three frequency bands decrease, while words from the last three increase (panels 7 to 12), i.e. learners use more frequent words with increasing proficiency. This is in line with the development shown in the first six panels. While this development is not immediately expected, panel 13 shows, that the number of unknown lexical types drastically decreases with increasing proficiency, too. Hence, the development is likely to be an artifact of the improved orthography of learners with advancing proficiency level. This also fits with the observation, that all changes level off for levels B2 and C, because learner are likely to have masters orthography

[1] I.e. *AFTFreq/LTD* in row 1, column 1 is positioned in panel 1, *typeFreq/LTD* in row 1, column 2 in panel 2, and *logTypeFreq/LTD* in row 2, column 1 in panel 5.



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.1.: DlexDB frequencies across proficiency levels.

at this level.

On the *Falko Georgetown L2* data, genuine language use effects are more likely, since the texts were written at home, where learners could use the aide of a spell-checker or proof read their texts before submitting. This also shows in the scales for all measures: Frequencies are considerably higher for *Falko Georgetown* than for *Merlin*. This makes all findings on *Falko Georgetown* more likely to show actual changes in language use, except that there seem to be clear task effects: For book review texts, there are virtually no changes across course levels, i.e. for a stable task environment, language use does not change. For curricular tasks, there is a general decrease of annotated lexical type and lemma frequencies (panels 1 to 6), which is likely to be due to the increase of words unknown to the data base for these texts with advanced course level (panel 13). Since learners may spell-check their writing before submission, a plausible hypothesis would be, that this increase indicates the use of more specific vocabulary in the curricular writing tasks. This is in line with the expected functional requirements of the tasks, since the curricular tasks for course levels 3 and 4 discuss more specific topics than the other tasks. Another indicator, that language use as measured by word frequencies is highly prone to task effects, are the idiosyncratic properties exhibited by course level 2 for some of the measures (panels 4 to 11). This also explains, that across frequency bands, there is no consistent development to be seen.

Overall these findings indicate, that language use in terms of vocabulary frequencies is a highly sensitive measure: It is considerably influenced by non-standard writing variants in beginner to intermediate learner writings, which was anticipated, though. Also, vocabulary use is highly sensitive to task differences, while it barely changes for intermediate to advanced learners, who take the same task. Thus, results on lexical sophistication measured by frequency data base entries should be interpreted with care, when they are assessed on data, where proficiency levels are correlated to different task backgrounds. Also, when applied to non-standardized beginner to intermediate learner data, measures seem to collectively assess orthography rather than language use.

10.2. Observations for Discourse and Encoding of Meaning

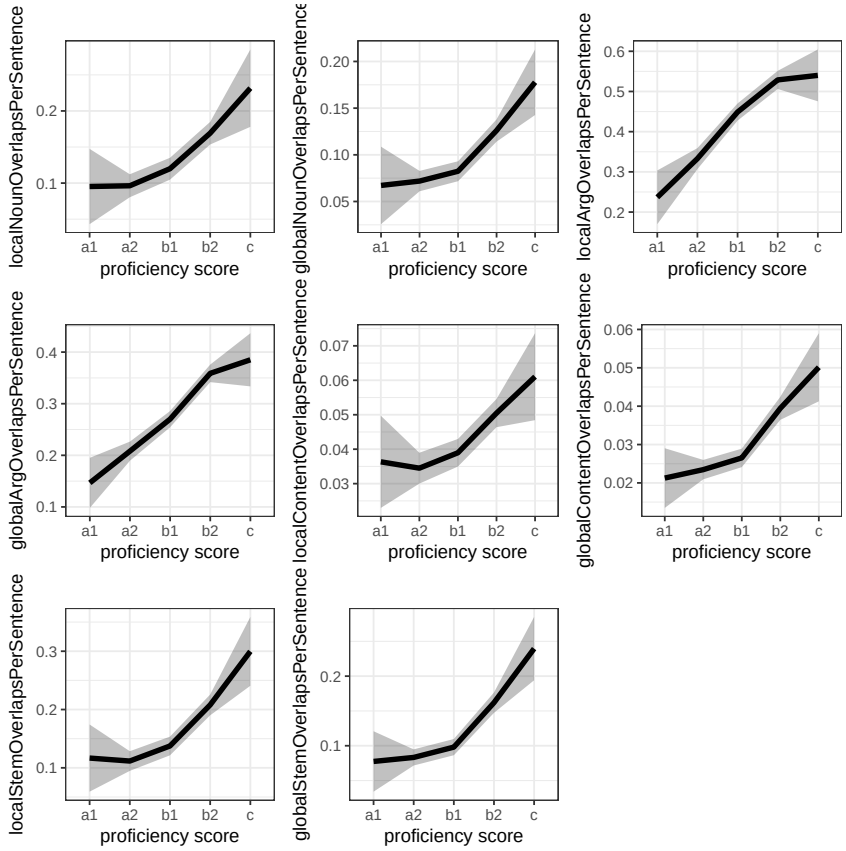
Discourse and encoding of meaning are represented with i) local and global overlaps of linguistic material; and ii) referring expressions in terms of pronouns and articles, in order to assess two distinct cohesive devices. Figure 10.2 shows the

former. On *Merlin*, the measures collectively increase throughout proficiency levels, indicating overall more cohesive writing. The increases level off for local and global argument overlaps on higher proficiency levels, though. For the *Falko Georgetown L2* curricular tasks, there are increases for local and global noun and stem overlaps over course levels, too, however, argument overlaps decrease and content overlaps remain virtually the same. Also, for book reviews, all measures gain high scores without any development across course levels. These plots indicate, that increasing proficiency is correlated with more cohesive writing in terms of overlapping of linguistic material, but that for higher proficiency levels, the effect levels off for argument overlap. Also, there seems to be a task component to shared linguistic material, since book reviews and curricular tasks differ so severely.

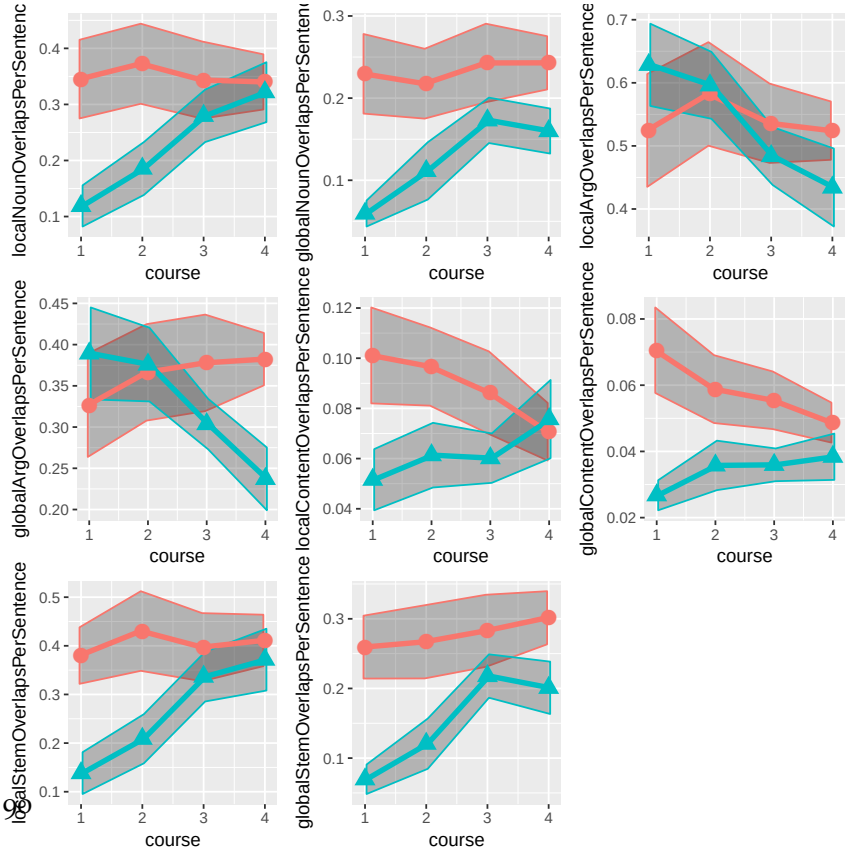
In terms of linguistic co-reference devices, the situation is a little different, though, as may be seen in Figure 10.3. Ratios for pronouns, personal pronouns and possessive pronouns per token in sentence (TIS) (panels 1 to 3) overall decrease for increasing proficiency across corpora: On the curricular task in *Falko Georgetown L2*, this trend is only interfered by some task effect at course level 2. However, there is an increase of third person possessive pronouns for book review tasks (panels 3 and 5), i.e. book reviews seem to elicit third person possessive pronouns in particular. There is also an overall decrease in the use of first and second person pronouns with increasing proficiency on both, the *Merlin* data and the *Falko Georgetown L2* curricular tasks, which is likely to be task induced, too, cf. task descriptions in Section 5.2 and 6.2. Book reviews do not elicit any kind of first or second person pronouns at any course level, which is plausible from a functional perspective.

The use of definite and indefinite articles increases throughout proficiency level for all corpora independent of task type (i.e. curricular tasks vs. book review). The use of proper names shows a general decline across proficiency levels on *Merlin*, which levels off for more advanced levels. On *Falko Georgetown L2*, curricular tasks of course level 1 and 2 are at the same low level as on *Merlin* for levels B2 and C, but book reviews seem to generally require a slightly raised amount of named entities as well as the tasks for course levels 3 and 4, i.e. articles and speeches, which also is functionally plausible.

Taken together, this shows, that general pronoun and article use seems to develop relatively stable across proficiency levels, corpora, and task backgrounds. Only the use of person is considerably influenced by task requirements, which is functionally expected, though.

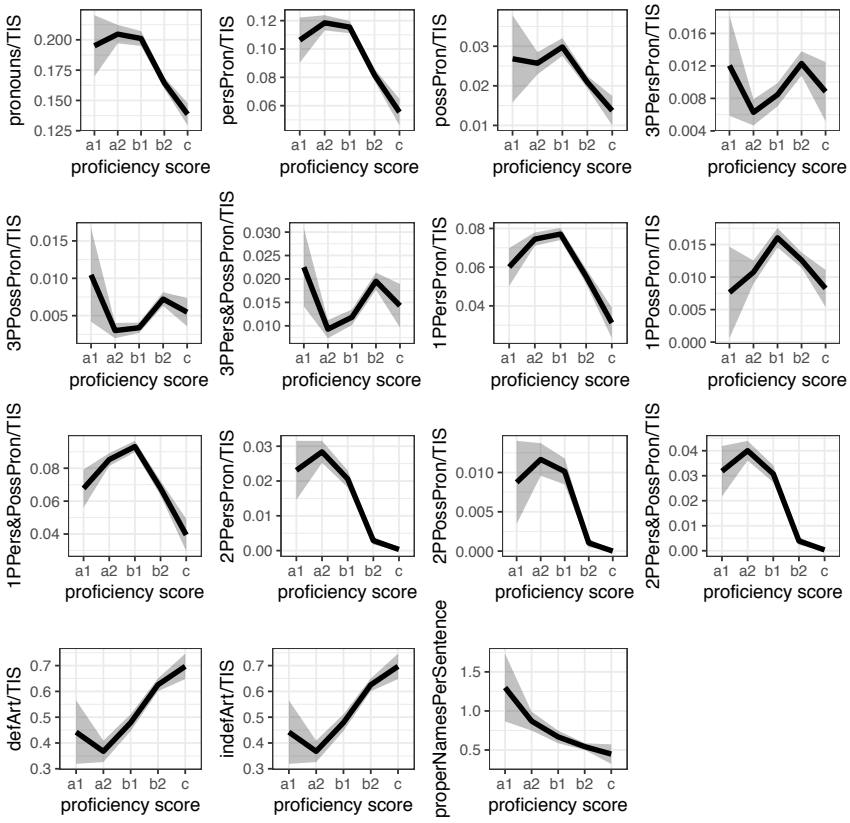


(a) Extracted from *Merlin* data.

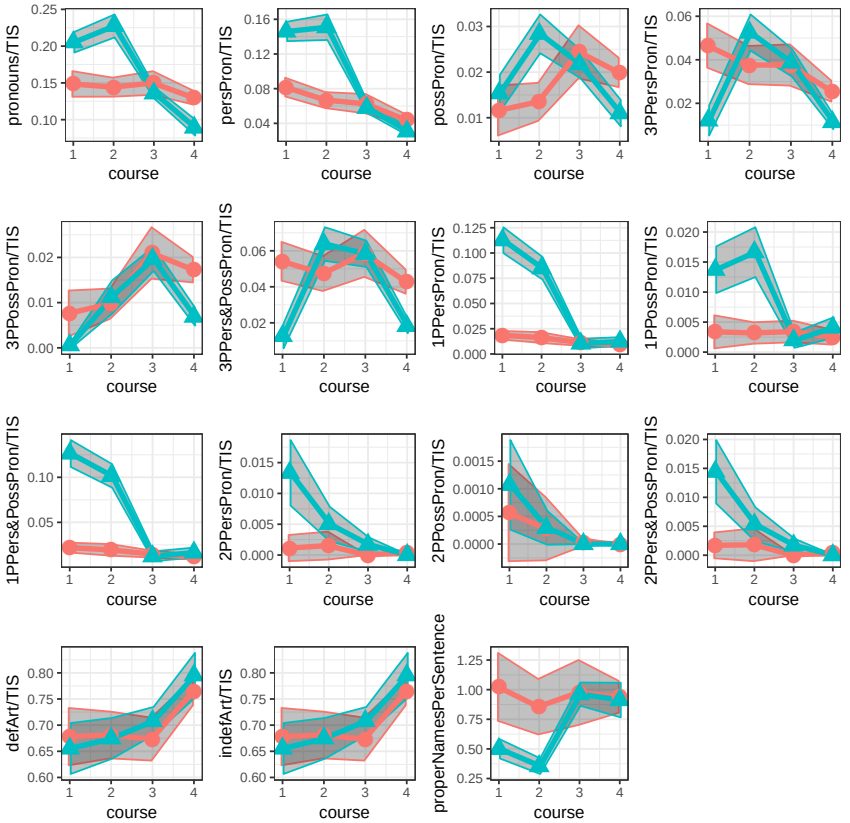


(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.2.: Overlap of linguistic material across proficiency levels.



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.3.: Pronouns, articles and names across proficiency levels.

10.3. Observations for Human Language Processing

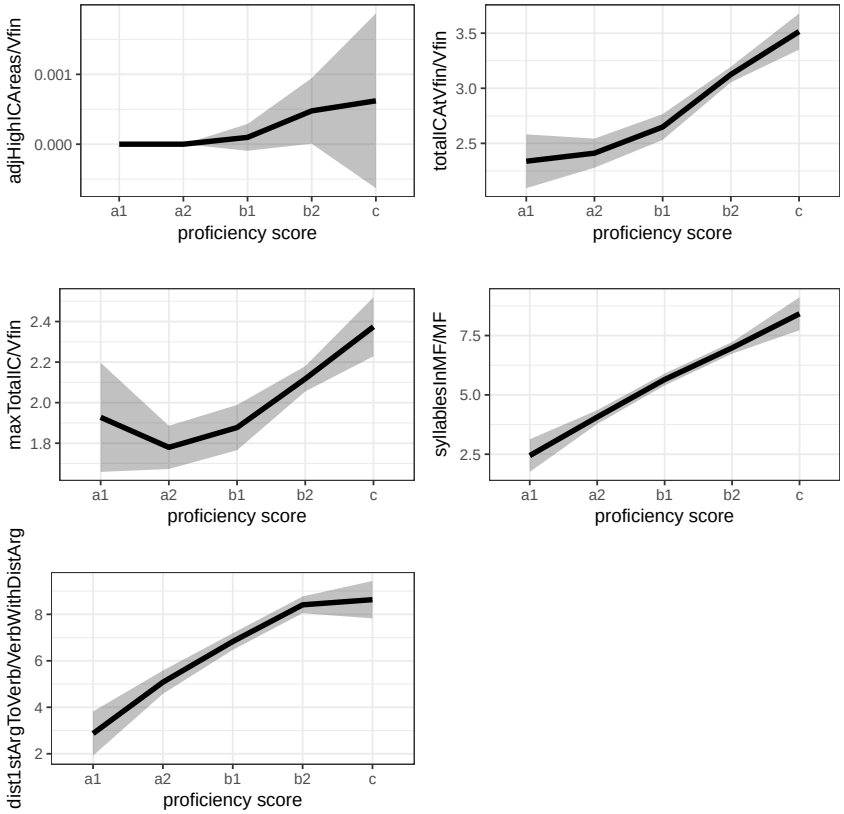
Human language processing is represented by measures of the DLT using additional verb weight and two syllable distance measures in Figure 10.4. The additional verb weight condition was chosen over the other DLT measures, because it elicits the highest possible DLT values, which makes the observation of high adjacent integration cost (IC) areas most likely. Still, as may be seen clearly from the wide confidence intervals, there are hardly any instances of such high adjacent integration cost areas in the data. It remains to test, whether this is a property of L2 language or whether this lack of adjacent high integration cost areas, which may be caused, for example, by deeply embedded clauses in the middle field, persists on L1 data, too. If the measure is more informative on L1 data, this would be an interesting discovery with regard to the differences between German L2 and L1 writing. Otherwise, this would indicate that the thresholds for high adjacent integration cost areas are set too high. Hence, the results will lead to valuable insights either for research or for improvement of the system. This issue will be further investigated in the near future.

Average total and maximal integration cost per finite verb (Vfin) increase with increasing proficiency (panels 2 and 3) on both, the *Merlin* and the *Falko Georgetown L2* data and there is no difference between curricular tasks and book reviews within course levels. Furthermore, the achieved scores for more proficient *Merlin* learners and the *Falko Georgetown L2* data are at similar levels. The same holds for the number of syllables in middle fields, which rises from on average 2.5 to more than 7.5 (panel 4) on *Merlin* and from slightly less than 6 to nearly 10 on *Falko Georgetown L2* and the number of syllables between the first argument and its governing verb per verbs that have non-adjacent arguments (panel 5).

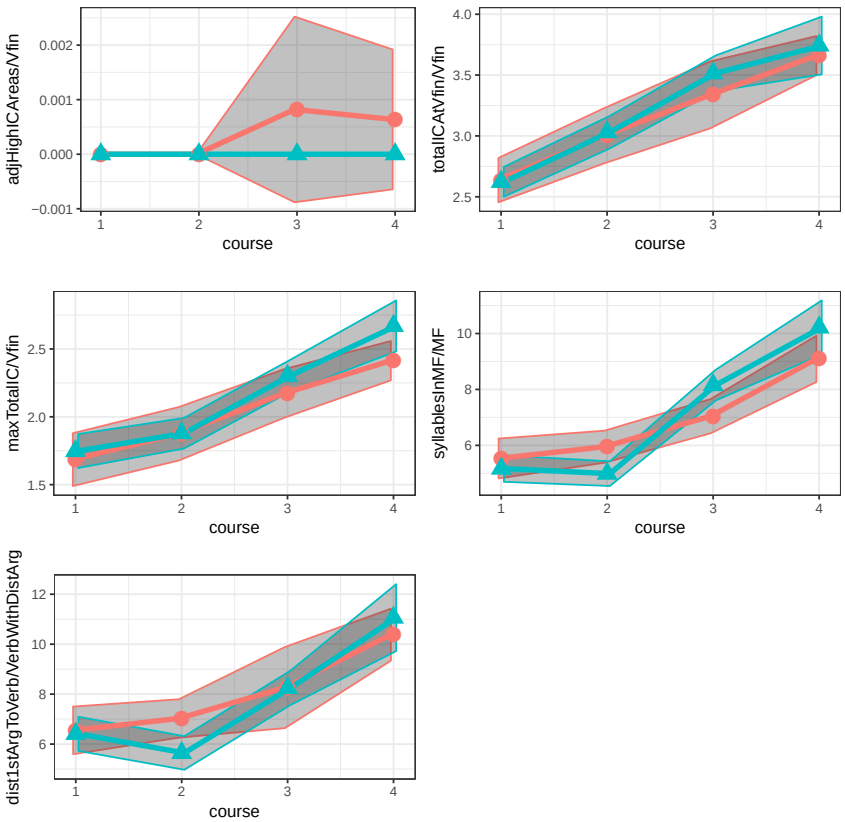
These are highly remarkable results, because they clearly indicate, that measures of human language processing are homogeneously increasing with increasing proficiency, and that this process is stable against task differences. No other group of measures throughout the entire analysis shows similarly consistent results.

10.4. Observations for Lexical Complexity

Lexical complexity is assessed in terms of lexical variation, which is one of the most commonly assessed measures of lexical complexity. Also, unlike lexcal diversity, it allows for more fine-grained insights into the development of lexical variation



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

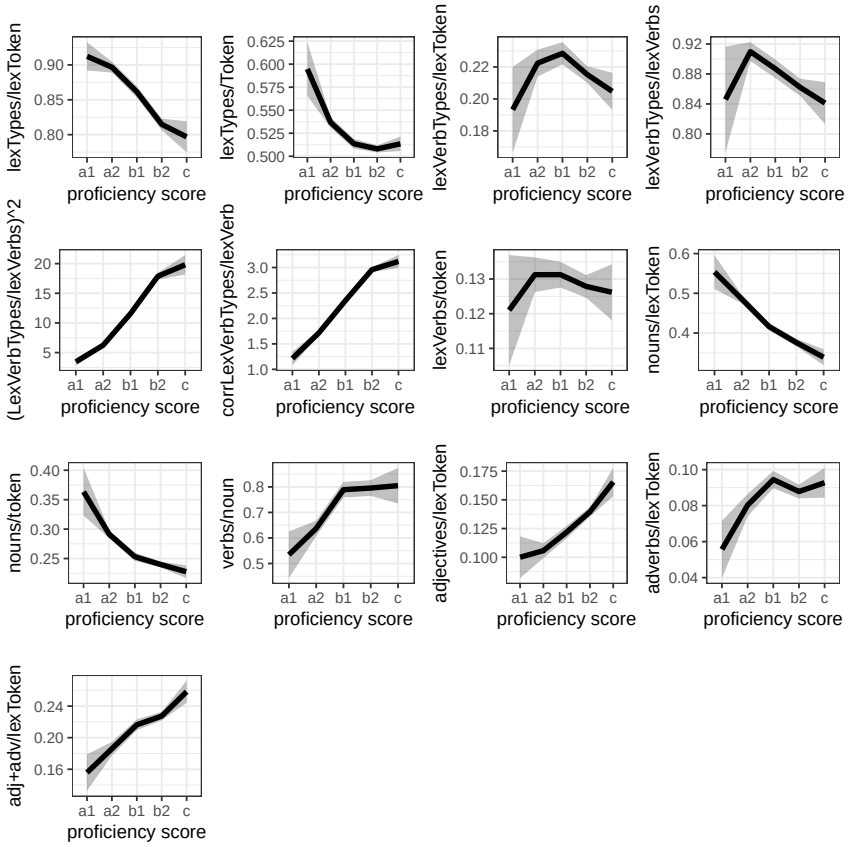
Figure 10.4.: DLT-V and syllable distance measures across proficiency levels.

for various POS. The measures are shown in Figure 10.5. The two variants of lexical type token ratio in panels 1 and 2 show a decrease of lexical variation with increasing proficiency on *Merlin*, which levels off for higher proficiency. This is likely to be due to a stabilization of orthography on the non-normalized data. Accordingly, on *Falko Georgetown L2*, there are no clear developments to be seen except for curricular tasks, which shows an increase of lexical variation, when normalized by tokens instead of lexical tokens. Since there is a tendency of an inverse development for the latter normalization, this indicates an increase in the use of functional tokens words in more proficient writing. Lexical verb variation shows a decrease with increasing proficiency across corpora (panels 3 to 7) for *Merlin* texts and curricular tasks in *Falko Georgetown L2*, while for book reviews there is again virtually no change.[2]

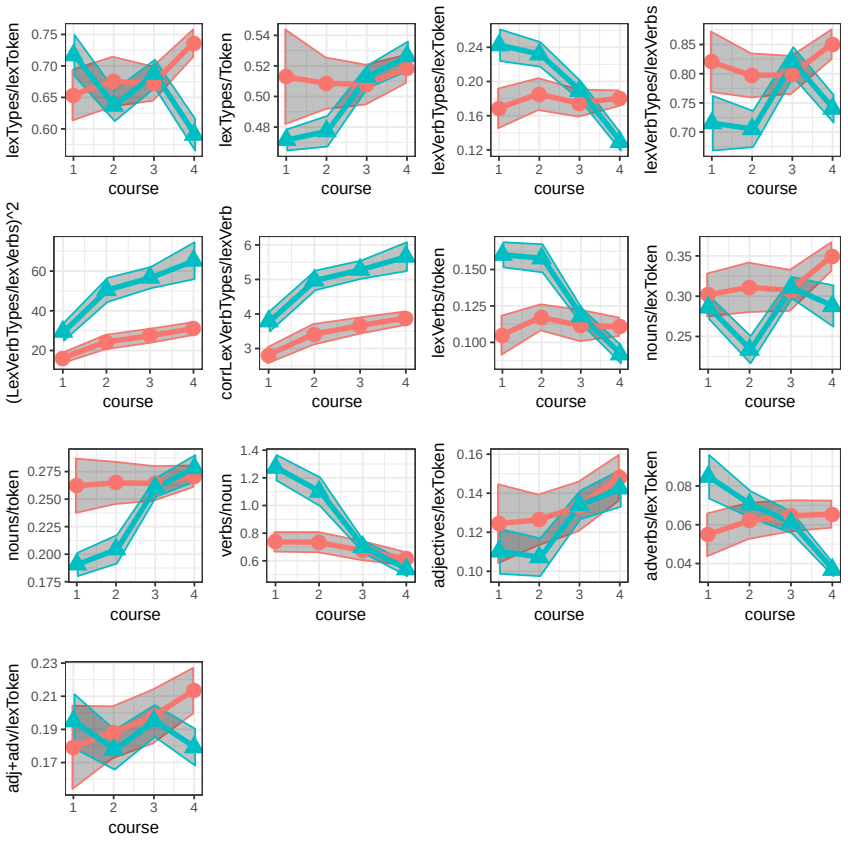
Measures of noun variation decrease more clearly with increasing proficiency (panels 8 and 9) in *Merlin* and are stable on *Falko Georgetown L2* across course levels at the level exhibited by intermediate and advanced learners in *Merlin*. There seem to be some task-specific influences, though, especially with regard to the curricular task in course level 2 and a generally lower amount of nouns per tokens for curricular tasks in course level 1 and 2. This may also clearly be seen for the verb to noun ratio in panel 10, which increases in *Merlin* with increasing proficiency up to B1 learners, but stagnates around 0.8 for intermediate to advanced learners and is around 0.8 for *Falko Georgetown L2* book reviews and curricular tasks from course level 3 and 4, but is much higher for curricular tasks from course level 1 and 2. A verb to noun ratio above 1 indicates, that texts contain more verbs than nouns. This means, that the curricular tasks from course levels 1 and 2 inhibit the usage of nouns, although intermediate to advanced learners seem generally to be prone to use slightly more nouns than verbs, leading to a ratio of 0.8. Also, the *Merlin* data shows that less proficient learners use considerably more nouns than verbs, but then progress towards a more balanced ratio with only slightly more nouns than verbs.

Finally, more proficient learners use more adjectives and adverbs in *Merlin*, but the effect levels off for the latter (panels 11 to 13). This increase may also be seen for adjectives in *Falko Georgetown L2* irrespective of whether learners write in the context of a curricular task or a book review, but with relatively wide confidence

[2]Note that verb variation seems to increase for lower proficiency levels on *Merlin* in panels 3 and 4, but due to the wide confidence intervals and the lack of similar trends in the mathematically more stable transformed variants (panels 5 to 7), this impression is unreliable.



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.5.: Lexical variation measures across proficiency levels.

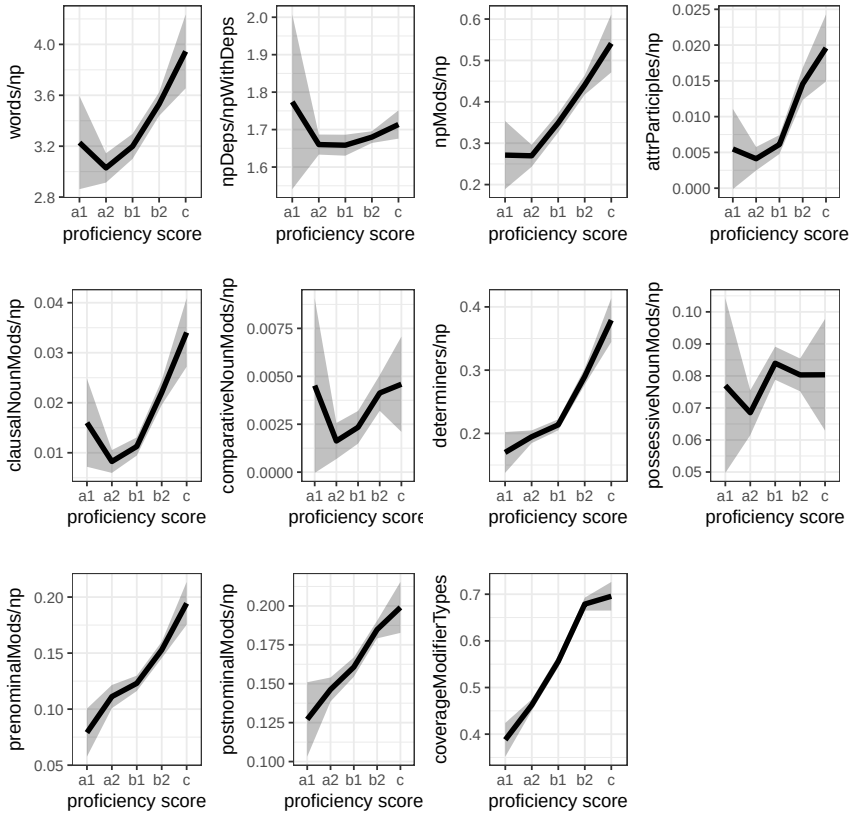
intervals. The ratio of adverbs to lexical types decreases with increasing course level for curricular tasks and is stable but considerably lower than in *Merlin*, which draws a less clear picture.

10.5. Observations for Syntactic and Grammatical Complexity

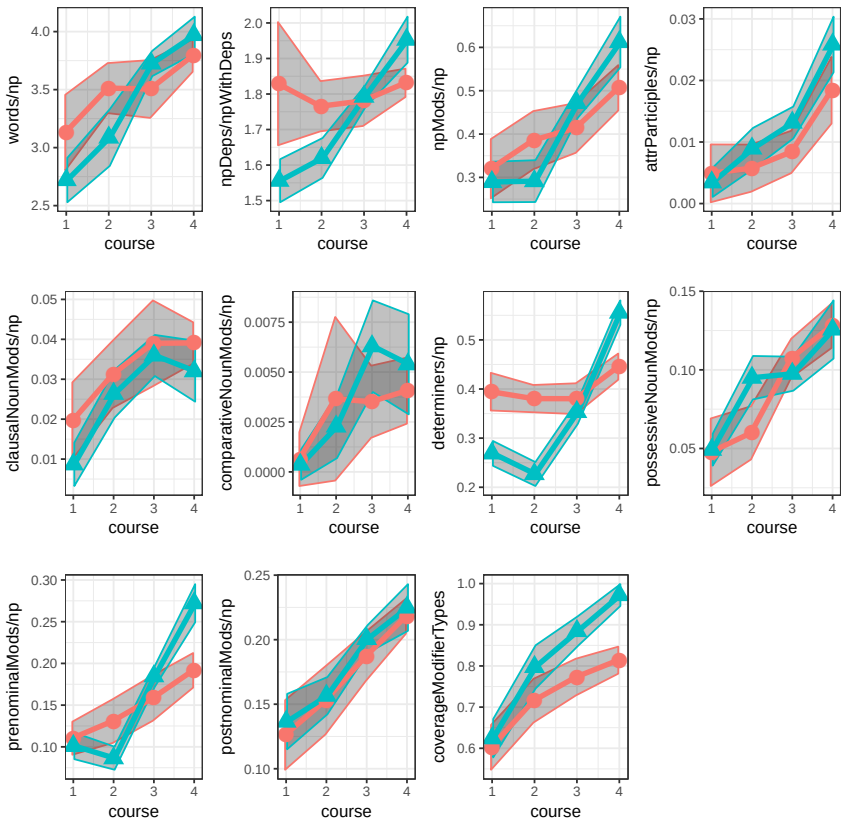
Grammatical complexity is represented in terms of i) complex noun phrases; ii) tense and passive measures; and iii) dependent clauses. The special focus on complex noun phrases is motivated by the importance of the nominal domain in German academic language (Hennig & Niemann 2013; Schlömer 2013). Furthermore, the complexity analysis system used in this thesis has a particularly elaborate set of measures to assess various aspects of the nominal domain. Periphrastic grammatical constructions in terms of tenses and passives are assessed, to target prominent measures of grammatical complexity. Dependent clauses broaden the view to clausal complexity, in particular in terms of subordination, which is not only also traditionally measured to assess syntactic complexity, but also allows to compare the development of the phrasal and clausal domain.

Phrasal Complexity Figure 10.6 shows various indices to measure complex noun phrases. Overall, the development of the complex noun phrase is quite similar across corpora and increases with increasing proficiency irrespective of tasks, with *Falko Georgetown L2* texts showing similar ratios to those obtained for intermediate to advanced learners in *Merlin*. This holds especially for words an modifiers per noun phrase (panels 1 and 3), attributive participles and clausal noun modifiers (panels 4 and 5), as well as determiners (panel 7), and prenominal and postnominal modifiers (panels 9 and 10), as well as the coverage of modifier types (panel 11). The only exception to this is the ratio of determiners per noun phrase in *Falko Georgetown L2* book reviews, which starts and remains at a rather high ratio of 0.4. This is to be expected given the previous findings for the use of articles in Figure 10.3, though. It is remarkable, that the ratio of attributive participles and clausal noun modifiers is relatively similar: Attributive participles and postnominal relative clauses, which are the most common form of clausal noun modifiers, are often equivalent in meaning and may be used interchangeably (Fabricius-Hansen 2014). Yet, the measures to not indicate a preference towards either construction in the texts.

As for the other measures: The amount of noun dependents per nouns with



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

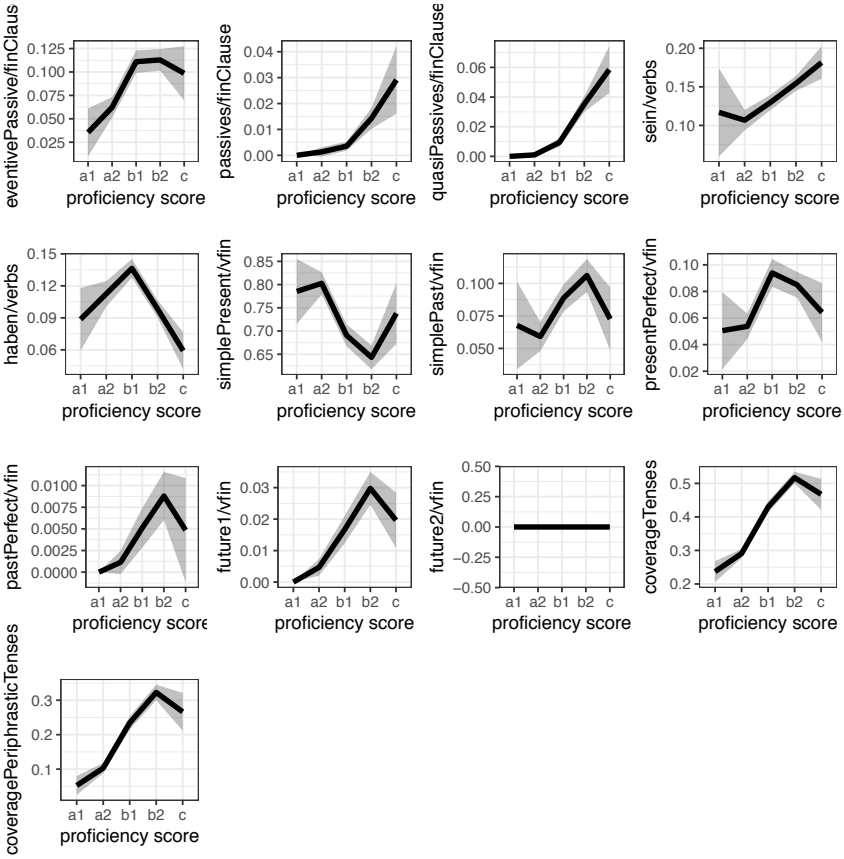
Figure 10.6.: Complex NP measures across proficiency levels.

dependents (panel 2) stays relatively stable in *Merlin* and *Falko Georgetown L2* book reviews, except for the wide confidence intervals for A1 learners, which are likely to be artifacts of issues of the automatic assignment of dependency relations on the non-normalized data. However, for the curricular tasks, there is a steady increase, which is likely to be due to the inhibited use of noun phrases in the curriular tasks for course level 1 and 2. Comparative noun modification seems to increase for more proficient learners (panel 6), but the constructions are relatively rare and the measure shows wide confidence intervals, which makes the general tendency difficult to interpret. Finally, possessive noun modifiers increase for intermediate to advanced learners in *Falko Georgetown L2* irrespective of task, but there is no clear tendency to be observed on *Merlin* due to rather wide confidence intervals.

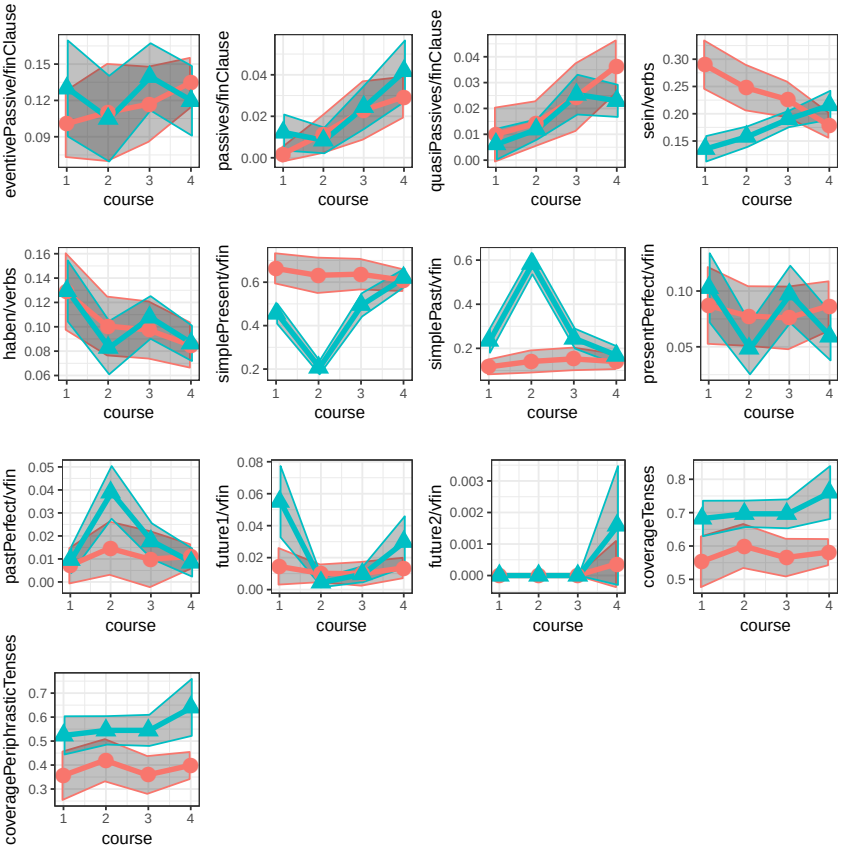
Overall, the results show a clear general trend towards a global complexification of the noun phrase and an increasing variety of used modifies, which is relatively stable across task backgrounds and corpora.

Grammatical Complexity Figure 10.7 shows ratios measuring the use of periphrastic grammatical constructions. The first three panels show various passive measures. Overall, passive and quasi passive constructions are used more often by more proficient learners (panels 2 and 3) across corpora irrespective of task. Eventive passive, though, only increases up to proficiency level B1 in *Merlin* and does not show any clear pattern in *Falko Georgetown L2*, due to wide confidence intervals, indicating that this measure assesses increased proficiency for lower proficiency levels, only.

The uses of *sein* and *haben* per verb (panels 3 and 4) are actually conceptualized as measures of verb variation in the complexity analysis system, but presented here, because they may be interpreted as coarse but robust approximation of periphrastic grammatical constructions and proved to be more interpretable, when analyzed together with periphrastic tenses for the current purposes: Learners seem to systematically increase their use of *sein* with progressing proficiency, although something in the book review task in *Falko Georgetown L2* seems to elicit an uncharacteristically frequent use of *sein* for earlier course levels, which then decreases to the level to which curricular task and *Merlin* writings progress. For *haben*, there is no clear picture on *Falko Georgetown*, due to wide confidence intervals, but on *Merlin*, the plot shows a clear peak for B1 learners, indicating an over-use of *haben*. This could be due to some form of restructuring of periphrastic grammatical knowledge at the intermediate level, which would also explain the incoherent



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.7.: Tense and passive measures across proficiency levels.

picture on *Falko Georgetown L2*, which is also written by intermediate to advanced learners and in fact oscillates on the same scale as *Merlin* B1 to C level learners. Also, B1 scores are assigned to learners taking test levels A1 to B2 on *Merlin*, which makes a task effect less likely. When comparing these findings with more concrete measures of periphrastic grammatical constructions, this seems to be related to the acquisition of present and past perfect, which show a similar pattern on the *Merlin* data and equivalently wide confidence intervals in *Falko Georgetown L2*, which again oscillate within the boundaries of the *Merlin* scale (panels 8 and 9).

As for future tense, the use of future 1 (panel 10) increases with increasing proficiency on *Merlin* and shows similar scores on *Falko Georgetown L2* as for intermediate to advanced learners in *Merlin*, but it also seems to be very prone to functional task factors. Thus, it is uncommonly frequent in the curricular task for course level 1, which is a letter about future events, and for B2 learners in *Merlin*, which could be due to functional task factors when considering the tasks posed at test level B2, but cannot be fully determined without further investigations. Future 2 (panel 11) is not or hardly used by learners in either corpus.

Similarly, the use of simple present and past (panels 6 and 7) seems to be very prone to functional task factors, allowing for no clear picture on the *Merlin* data, except for showing that the plots seem to be inverse to each other, i.e. when there is a preference for simple past, texts include fewer simple present and vice verse. The same holds for the *Falko Georgetown L2* data, except that due to the lower number of different tasks the plots are more stable. In particular one may see, that book reviews are predominantly written in simple present rather than simple past and that the curricular task for course level 2 elicits the use of past tense.

As for tense variedness, it may be seen, that the coverage of tenses and periphrastic tenses increases steadily with increasing proficiency on *Merlin* (panel 12, 13). On *Falko Georgetown*, the ratio remains stable for book reviews at a rate similar to the one observed on *Merlin* for advanced learners. For curricular writing tasks, the ratios are slightly higher, indicating a more elaborate temporal structure. This is likely to be due to the explicit instructions given for these tasks, which directly ask learners to produce certain tense constructions.

Overall, the results show a relatively stable increase of passive use with increasing proficiency independent of corpus and task backgrounds, while tenses are functionally bound to be highly influenced by task. Some indications of a genuine cross-task over-use of *haben* related to the acquisition of periphrastic past tense at proficiency level B1 in *Merlin* could be identified, though, which requires further

investigation in future work.

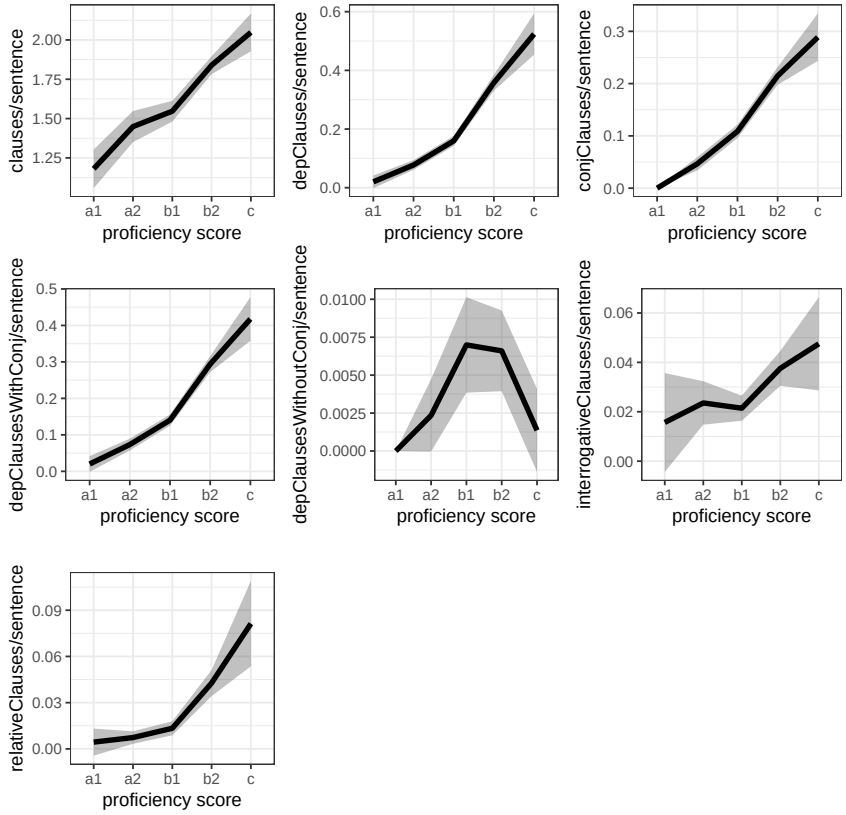
Clausal Complexity Figure 10.8 shows measures for clausal subordination. Across corpora and independent of tasks, more proficient learners use more dependent clauses with and without conjunctions as well as more conjunctional clauses and relative clauses (panels 2, 3, 4, and 7). Contrary to this, there are hardly any dependent clauses without conjunction in *Merlin* (panel 5), and due to the wide confidence intervals and the overall low frequency of instances, the observable temporary increase for proficiency levels B1 and B2 seems unreliable without further evidence. The same holds for the *Falko Georgetown L2* data, which does not show any clear changes due to its wide confidence intervals.

The overall ratio of clauses to sentences (panel 1) only increases on the *Merlin* data, whereas it remains stable in *Falko Georgetown L2* irrespective of task and with broad confidence intervals. The range of values observed in *Falko Georgetown L2*, though, is comparable to the ratios assigned to advanced learners in *Merlin*. Thus, the lack of increase seems not to indicate, that sentences have a less complex clause structure in *Falko Georgetown L2* than in *Merlin*. The same holds for the ratio of interrogative clauses per sentence (panel 6), except that the curricular task for course level 4 seems to inhibit the use of interrogative clauses.

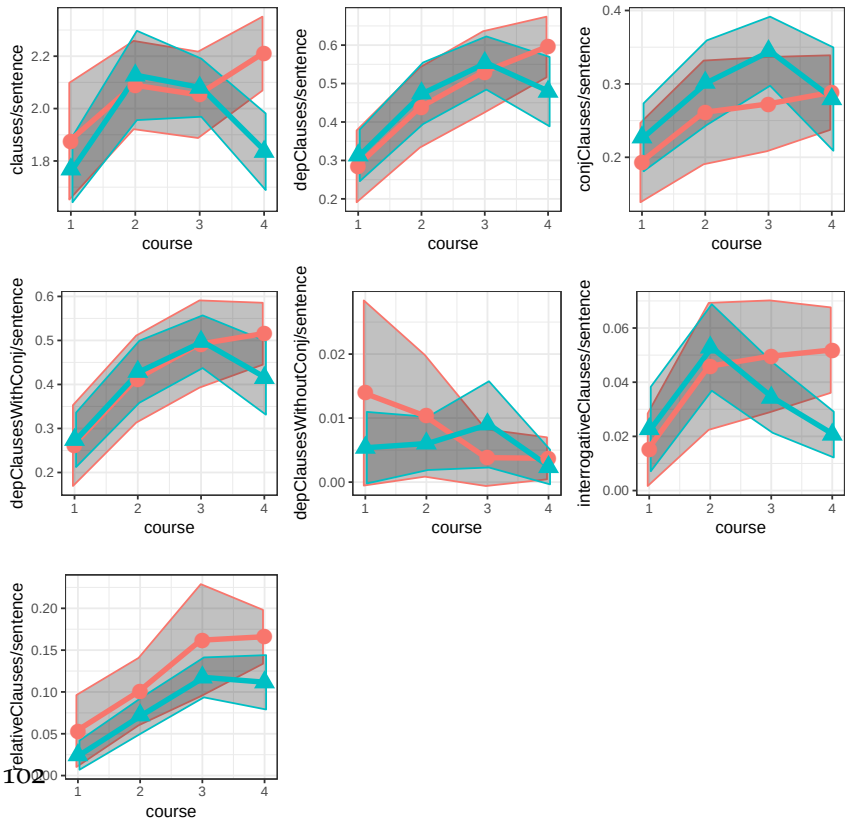
Overall, there seems to be a stable increase in the use of dependent clauses irrespective of task backgrounds. The continued increase for most measures at higher proficiency levels is unexpected, since it has been reasoned, that at later stages of acquisition, grammatical development targets especially the phrasal rather than the clausal domain (Paquot 2017). However, not all measured indices of clausal complexity show a clear pattern, for example the use of conjunctional clauses with conjunction or interrogative clauses. Hence, development in the clausal domain seems to show a heterogeneous picture depending on the indices used for assessment.

10.6. Observations for Morphological Complexity

Morphological complexity is measured in terms of noun and verb inflection. Alternatively, derivational measures could have been discussed, but the nominal domain has already been discussed in detail in the previous section. Hence, these measures were considered to convey more novel information. Figure 10.9 shows the plots for case marking across nouns and verbal inflection. For case inflection, panels 1 to 4

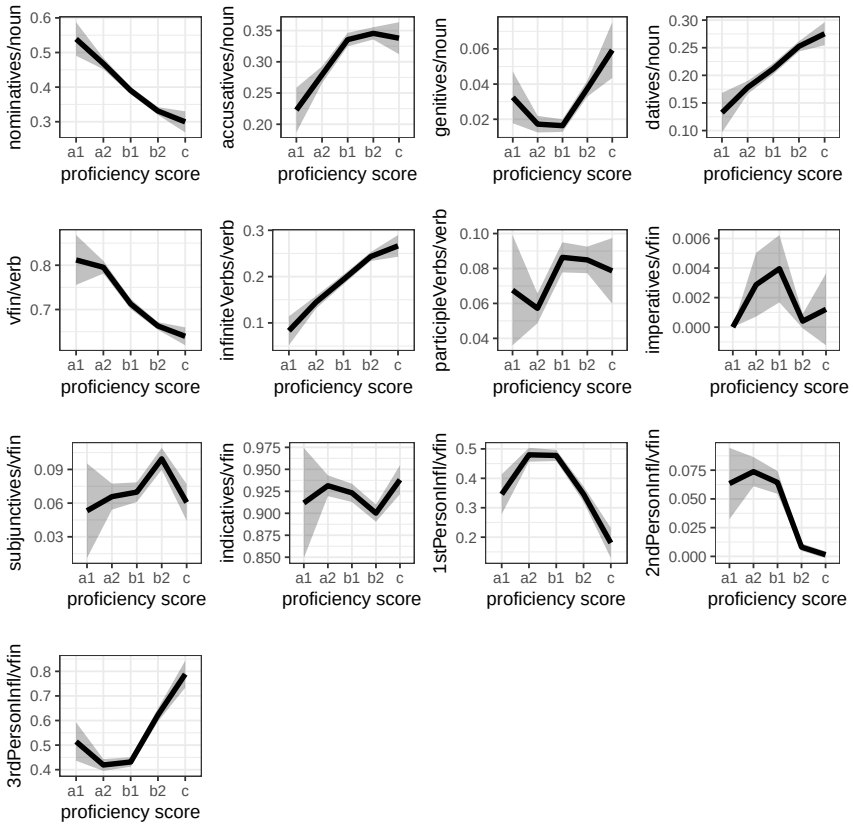


(a) Extracted from *Merlin* data.

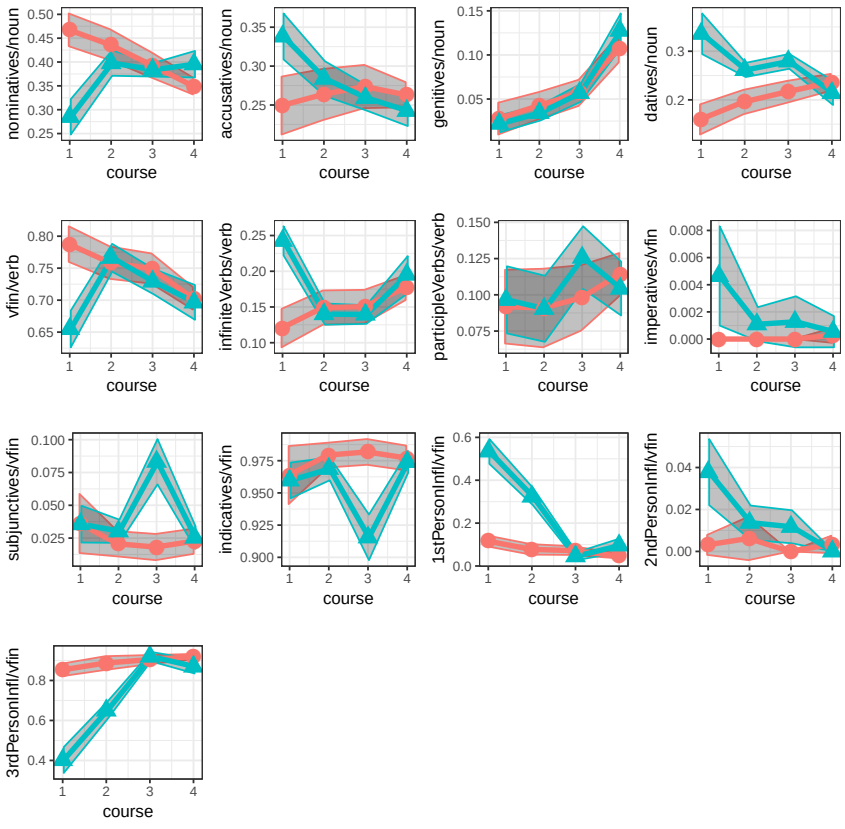


(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.8.: Dependent clause measures across proficiency levels.



(a) Extracted from *Merlin* data.



(b) Extracted from *Falko Georgetown L2* data; legend: △: curricular tasks, ○: book reviews.

Figure 10.9.: Case, status, mode, and person inflection measures across proficiency levels.

show decreases in the usage of nominative case in favor of increases for accusative, genitive, and dative case on the *Merlin* data. For accusative, though, this effect leves off for more advanced learners, which is also mirrored in the *Falko Georgetown L2* data, where nominative case decreases in favor of genitive case, but stays relatively stable for accusative case. Also, the curricular task for course level 1 seems to elicit increased use of accusative and dative case.

The use of finite verbs decreases seemingly in favor of other status of the verb, especially infinitive verbs (panels 5 and 6) on both corpora irrespective of task, except for the curricular task for course level 1, which seems to elicit considerably more infinitive verbs than finite verbs. The use of participles (panel 7), too, increases from beginners to intermediate learners, but not for intermediate to advanced learners. There, it remains stable around 0.08, which is the same ratio at which it stays throughout all *Falko Georgetown L2* texts. Note, though, that the wide confidence intervals on both corpora for the use of participles by intermediate to advanced learners indicate a relatively high inter-learner variability.

As for *genus verbi*, which is shown in panels 8 to 10, there seems to be no consistent progression: learners predominantly use indicatives across corpora irrespective of proficiency level or task, while they hardly use imperatives or subjunctives. The only exception to this seems to be the curricular task for course level 3, which shows a clear raise of subjunctives to 0.10 and a corresponding decrease of indicatives to 0.90.

Finally, the use of first and second person inflection on finite verbs decreases with increasing proficiency for *Merlin* texts as well as curricular writing tasks in *Falko Georgetown L2*, however, this is likely to be induced by functional task needs, which happen to require the respective use of first and second person in less advanced courses for both corpora, as may clearly be seen, when regarding the lack of development on the book review texts, which nearly exclusively exhibit third person markings.

10.7. Discussion

The visual inspection of complexity measures across proficiency and course levels conducted in this chapter was designed to give a first impression of the data and the features extracted from it. It provides a highly informative overview of a large quantity of measures, instead of limiting the investigation to a selected few, as is often the case in complexity analyses.

The result gave intriguing insights into to which extend certain feature groups develop homogeneously and are stable across task backgrounds. In fact, most groupings showed an overall consistent development, except for clausal complexity measures. These showed across corpora and task backgrounds increasing complexity in terms of subordination, but stagnation for some concrete measures, such as dependent clauses with conjunctions. This indicates, that it might be important to employ several measures of clausal complexity, in order to make generalizations about the clausal system. Most feature groups are stable in this regard, though. This holds even for noun complexity, although from a theoretical perspective, it would be reasonable to see a restructuring of the was noun phrases are modified, for example, from more relative clauses towards more attributive participles.

The results also confirmed several assumptions, that have been made in this thesis on grounds of theoretical or methodological considerations and preliminary analyses. In particular, the thorough comparison of measures extracted from *Merlin* and *Falko Georgetown L2* showed, that intermediate to advanced *Merlin* learners are comparable with *Falko Georgetown* learners in terms of language complexity: Most ratios, that were discussed here, showed highly similar results for B1 to C learners in *Merlin* and learners in *Falko Georgetown L2*. Since on *Falko Georgetown*, proficiency is approximated by course levels rather than expert CEFR ratings, these results are reassuring. Also, on *Falko Georgetown L2*, for several measures it could be seen, that homogeneous and heterogeneous task backgrounds elicit different language performances. This underlines the necessity of accounting for task backgrounds in learner corpora, when using them to assess language performance. At the same time, the analysis also clearly pointed out, which feature groups seem to be stable across task differences: Especially measures of human language processing seem to be insensitive to task differences.

Finally, the descriptive visual analysis yielded valuable impulses for future research, such as the need for further investigations of the measure for adjacent high integration cost areas and the acquisition of perfect tense, and allowed for a feasible, broad comparison of measures across research domains.

Part V.

Methods

11. Additive Regression Modeling

This chapter introduces ordinal Generalized Additive (Mixed) Models (GA(M)Ms), in order to provide the necessary background, to follow the studies reported in Chapters 13 and 14. First, Section 11.1 provides a very general introduction to linear models for readers, who are not familiar with regression modeling. Then, Section 11.2 introduces Generalized Linear Models (GLMs) and based on this, GAMs as generalizations of GLM. Section 11.3 elaborates on regression splines. Then, Section 11.5 gives a brief outline on mixed additive modeling. Finally, Section 11.6 gives a brief introduction on modeling responses as ordinal variables without re-analyzing them as either multinomial or numeric.

11.1. Introduction to Linear Regression

The standard Linear Model (LM) calculates an estimated response variable $\hat{y}$ to approximate a normally distributed, univariate response variable y in terms of a linear predictor η and some random noise ϵ, as shown in Equation 11.1.

$$\hat{y} = \eta + \epsilon, \; where \; \epsilon \sim N(0, \sigma^2) \; and \; \eta = \beta_0 + \sum_{i=1}^{I} x_i \beta_i \tag{11.1}$$

η is the weighted sum of I predictor variables x_i and their coefficients β_i. β_0 is the intercept of the function. Coefficients are estimated while fitting the model by minimizing the residual sum of squares (RSS), cf. Equation 11.2.

$$RSS = \sum_{i} (y_i - \hat{y})^2, \tag{11.2}$$

i.e. the sum of squared differences between y and $\hat{y}$. ϵ is a Gaussian random error with a zero mean and a uniform, i.e. *homoscedastic* variance.[1] This condition on the random error ϵ implements four main assumptions underlying LMs as well as

[1] Note that this standard definition of the model error ϵ as random noise applies throughout all formulas and is, therefore, not repeated throughout this chapter.

GLMs (cf. below):

1. The response variable y and its predictor η are in a **linear** relationship.[2] This is equivalent to the mean error ϵ being zero in a linear model.
2. All errors are **independent**.
3. All errors are **normally** distributed.
4. All errors have **uniform variance** (homoscedasticity).

If a model adequately models its data, it should adhere to these assumptions. If violations of these assumptions are observed, this indicates, that a model likely misses some structure in the underlying data, i.e. exhibits a poor model fit. Thus, proper model inspection requires to confirm, that a model's prediction errors, i.e. its residuals, adhere to these criteria. This holds, even if the *coefficient of determination* R^2, i.e. amount of observed variance in the data that is explained by the model, is high. R^2 is calculated by dividing the sum of squared differences between the predicted and the mean response by the sum of squared differences between the observed and the mean response, cf. Equation 11.3.

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{11.3}$$

Since this does not account for model bias, R^2 cannot indicate whether the model systematically misses certain structures in the data. Only the additional inspection of the residuals allows to draw informed conclusions about the model fit.

The weighted sum of predictors may not only contain linear combinations of predictors with their coefficients, but also interactions between predictors. Equation 11.4 illustrates this scenario for a LM with two predictors x_1 and x_2 with an interaction between them.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \tag{11.4}$$

For sake of simplicity, assume $\hat{y}$ to denote *writing progress*, x_1 *time spent writing*, and x_2 a binary factor for *concentration*, which is either high or low. In such a scenario, it is straight forward to imagine how writing progress depends on time spent writing, but the impact of time spent writing on writing progress is moderated by whether the author is concentrated or not (and vice verse). Put differently, this means, that

[2] That is, y may be estimated from η, with η being a linear combination of predictors and their parameters β.

the slope differs depending on which values x_2 takes. It is crucial to note, how this changes the interpretation of the coefficients for the components of the interaction: Without an interaction, β_1 denotes the unique effect of x_1 on $\hat{y}$. However, if x_1 interacts with another predictor, its unique effect is mediated by this predictor and the interaction's coefficient, too, i.e. $\beta_1 + \beta_3 x_2$. β_1 is only the partial effect for $x_2 = 0$. Because of this, the interpretability of partial effects drastically increases, if interacting predictors are centered such that $x_n = 0$ denotes the predictors mean. Naturally, the same holds for β_2, which is the partial effect of x_2 if $x_1 = 0$, since its unique effect is $\beta_2 + \beta_3 x_1$. In this case the mediating variable x_1 is continuous, though, so the interpretation of β_2 is less straight forward, as x_1 will virtually never be zero.

11.2. Generalized Linear and Additive Models

Generalized Linear Models (GLMs) relax the LM's assumption of normality for the response variable by means of a monotonic link function $g(\hat{y})$, that links the linear estimated response $\hat{y}$ to any distribution from the exponential family (Wood 2006: 59ff).[3] Equation 11.5 shows a basic form of the GLM, where the only novelty compared to Equation 11.1 is the link function g.

$$g(\hat{y}) = \eta + \epsilon, \text{ where } \eta = \beta_0 + \sum_{i=1}^{I} x_i \beta_i \tag{11.5}$$

GLMs may be used to model, for example, categorical or binary data. For each distribution, another link function $g(\hat{y})$ is required: For the normal distribution, the identity link is used. The logit function is used as link function for binary data, and in its generalized form for multinomial data. Another example is the log function, which is used as link function for loglinear or poisson regression.

Note, that parameters of GLMs are not fitted by least squares, but by means of maximum likelihood estimation techniques, such as Re-weighted Iterative Least Squares (IRLS): The core component of this approach is a *weighted* least squares estimation, which employs estimated weights w to allow for heteroscedasticity in the data, i.e. multivariate normally distributed errors. Large weights are assigned to data points with small error variance, and small weights to data points with

[3] The exponential family of distributions includes the normal as well as a series of other distributions such as gamma, Poisson, Bernoulli, categorical, chi-squared or the Dirichlet distribution. For more details, please see Wood 2006: 61ff

large error variances, in order to give more credit to more informative data points. The procedure is iterative, since it uses a temporary estimation of $\hat{\beta}_i$ in order to calculate a temporary dependent variable z_i and some temporary iterative weights w_i, which are then used for calculation of an updated estimation of $\hat{\beta}$. This process is iterated until convergence. For details on this procedure, please see Wood 2006, Chapter 2.

Generalized Additive Models (GAMs) (Hastie & Tibshirani 1986, 1990) are a generalization of GLMs, that allow for non-linear relations between a response (from any of the exponential families) and its predictors by introducing monotonic smooth functions $s(x)$ over at least one of the predictors x to the sum of predictor variables η (Baayen et al. 2017: 4,Wood 2006: 119f). Equation 11.6 illustrates this for a GAM with smooths over all I predictor variables:

$$g(\hat{y}) = \eta + \epsilon, \textit{ where } \eta = \beta_0 + \sum_{i=1}^{I} s_i(x_i) \tag{11.6}$$

In order to integrate non-linear relations into a linear model, regression smooths are represented as linear combinations of a finite set of K weighted basis functions $b(x)$ over a predictive variable x (Baayen et al. 2017; Wood 2006), cf. Equation 11.7:

$$s(x) = \sum_{k=1}^{K} b_k(x)\beta_k, \tag{11.7}$$

where $b_k(x)$ denotes the smooth's known k^{th} basis function over x, and β_k its coefficient, which is to be estimated.[4] Together, all basis functions build a function space, that contains $s(x)$ or an approximation of $s(x)$ (Wood 2006: 120). Equation 11.8 illustrates this for a C^{th}-degree polynomial smooth.

$$s(x) = \sum_{c=1}^{C+1} x^{c-1}\beta_c \tag{11.8}$$

The polynomial smooth consists of the sum of $C+1$ polynomial basis functions of degree 0 to C, which are each multiplied by their coefficient β_c. A graphic illustration of how the sum of weighted basis functions is combined to a smooth is shown in Figure 11.1 from ibid.: 121. It shows a 4^{th}-degree polynomial smooth and its five weighted basis functions sorted by increasing polynomial degree, i.e.

[4]Note, that here and throughout the thesis $s(x)$ is used as a shorthand to refer to $s_i(x_i)$ for any i, for sake of simplicity. The same holds for all other indexed variables, too.

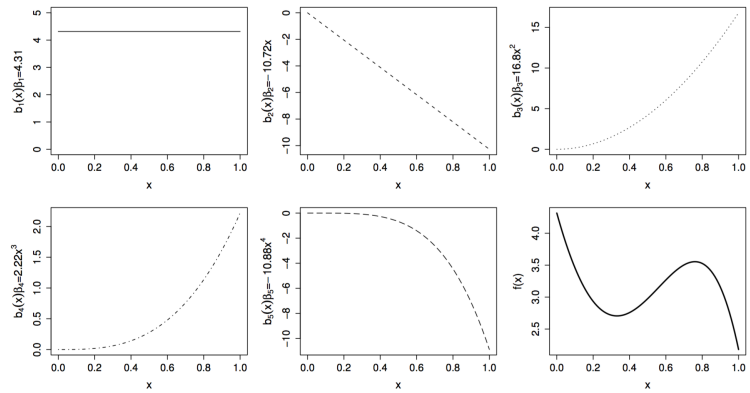


Figure 11.1.: A 4th-degree polynomial smooth preceded by its five weighted basis functions, cf. Wood 2006: 121, Figure 3.2.

the first basis function shows β_0, the second panel $x\beta_2$, the third $x^2\beta_3$ and so forth. The concrete values for the basis functions' coefficients β may be found at y-axis. The sixth panel shows the final polynomial smooth, which results from the sum of the preceding basis functions.[5] The smooth function is determined by its basis functions, insofar as it can take the form of any polynomial up to the 4^{th} degree. However, its concrete shape is determined by the coefficients β.

While polynomial smooths are conceptually straight forward, higher-degree polynomial smooths may be computationally expensive and exhibit severe involuntary oscillation at the edges of the interval. This oscillation issue is also known as the *Runge's phenomenon* and illustrated in Figure 11.2, where the *Runge* function is depicted with increasing degrees of polynomial interpolation. For these reasons, this type of smooth is less often used. In general, according to Larsen 2015: 8, there are three major types of smooths used in GAMs: i) local regression or *loess* smooths, ii) smoothing splines, and iii) regression splines. Only the latter type of smooths will be discussed in further detail in this chapter: They may be calculated independent of the response variable, are cheap in their computation, and highly interpretable (ibid.: 8). Therefore, only this type of smooths is used in both studies in Chapters 13 and 14.

[5]Note, that Wood 2006 refers to smooth functions as $f(x)$, hence the y-axis' label. Thus, for this an all following figures from ibid., $f(x)$ is equivalent to the usage of $s(x)$ in this thesis.

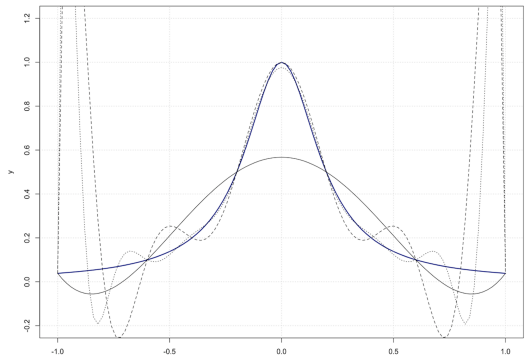


Figure 11.2.: Runge's test function for interpolation techniques (blue line), 5th-order polynomial (black line), 10th-order polynomial (dashed line), and 15th-order polynomial (dotted line) for the interval [-1; 1], cf. help page for *runge* function in *R* package *pracma* (Borchers 2017).

11.3. Regression Splines

Regression splines are globally flexible, non-linear smooth functions, which are defined interval-wise by *low-degree* polynomials. These local functions are smoothly connected at data points, that constitute the integral boundaries, so called *knots*. Knots may occur at every data point, resulting in a conventional spline, or be restricted to a subset of selected data points. Typically, knots are evenly spaced over the relevant function interval or placed at quantile ranges (cf. Larsen 2015; Wood 2006).[6] Figure 11.3 from ibid.: 122 illustrates this for a *natural cubic spline* consisting of seven basis functions. Here, two knots always constitute the boundaries for each of the basis functions and inner knots link adjacent basis functions to each other, while the two outer basis functions constitute the boundaries of the smooth's overall interval. At each inner knot, a dashed line illustrates its gradient and a curved line shows the quadratic function, that matches the first and second derivative at the knot. The second derivative at the spline's outer knots is zero. This is called a *natural* spline.

The spline in Figure 11.3 is a *cubic spline*, because it is build from a sum of weighted cubic polynomials. This is illustrated in Figure 11.3 from ibid.: 122 for a spline consisting of five basis functions. There are clear parallels to the set up of the polynomial smooth from the previous section: The smooth is build of five basis

[6]The placement of knots will not be of further concern here, because it done automatically by the type of regression splines used in this thesis.

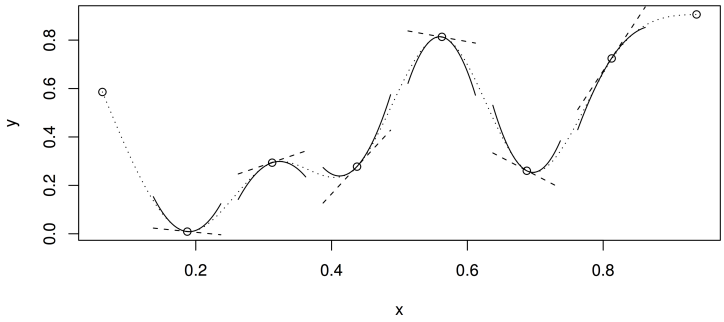


Figure 11.3.: Cubic spline build with 7 basis functions linked from Wood 2006: 122, Figure 3.3. Dashed lines show gradients at knots, curved lines show quadratic functions that match the first and second derivatives at each knot.

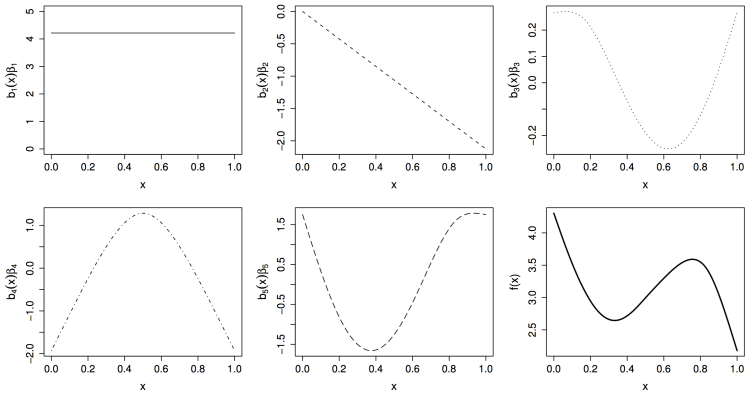


Figure 11.4.: A rank 5 cubic regression spline preceded by its weighted basis functions, cf. Wood 2006: 123, Figure 3.5.

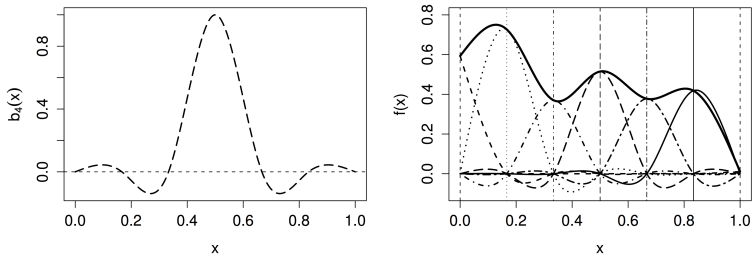


Figure 11.5.: Single cubic basis function (left) and full cubic regression spline (right), cf. Wood 2006: 147, Figure 4.1.

functions, which are continuous up to and including their second derivative and again, the first two basis functions are the intercept and a linear function. However, note that the basis functions are all cubic functions, not polynomials of increasingly higher degrees.

Figure 11.5 shows in more detail, how a single cubic basis function contributes to a full cubic, cf. Wood 2006: 147. The left panel shows a single cubic basis function $b_4(x)$, which has its maximum at 1 at a single knot and is zero at all others. This is illustrated even more clearly on the right panel, where a cubic spline (thick black curve) is plotted with all its basis functions (dashed and dotted curves) on the interval [0; 1]: at each knot, exactly one cubic function has its maximum, while all others are zero at this knot. Unfortunately, the written form of cubic splines is quite space consuming. Thus, the formula is not written out in this section, please see ibid.: 122ff, Gu 2002: 37 for details.

Thin-plate splines build their basis functions from weighted sums of linearly independent, low-degree polynomials. This is illustrated in Figure 11.6 from Wood 2006. The polynomial degree c of each basis function depends on the number of basis functions and their position in relation to the other basis functions, since basis functions, again, are sorted by increasing wiggliness (Baayen et al. 2017: 6f). Note, that the first two basis functions are again completely smooth, while all following basis functions introduce an increasing amount of wiggliness. Thin plate regression splines have certain mathematical properties, that make them generally more favorable than other types of regression splines (cf. Wood 2006: 150). These include the automatic identification of knot positions and the automatic identification of suited basis functions (see ibid.: 150ff for more details on this). One particular strength of thin-plate regression splines is, that they may be multivariate, i.e. smooth over more than one predictor. In particular, any n-variate thin-plate regression

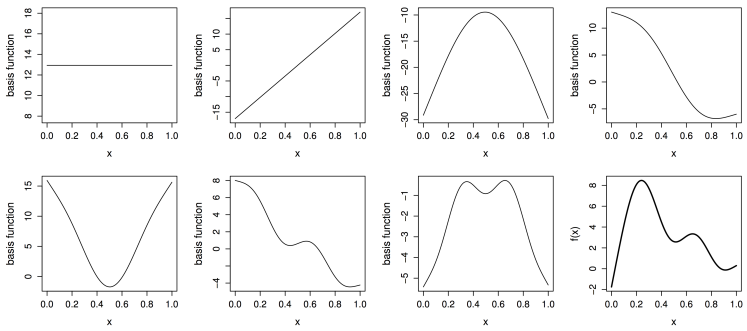


Figure 11.6.: A rank 7 thin plate regression spline preceded by its weighted basis functions, cf. Wood 2006: 153, Figure 4.5.

smooth $s(x_1, x_2, ..., x_n)$ may be used to build non-linear interaction surfaces or objects as weighted sums of simpler surfaces or objects (Baayen et al. 2017: 13). The only restriction on this is, that thin-plate splines require all interacting predictors to be isometric, i.e. measured on the same scale. For non-isometric interactions, tensor product smooths may be used, which do not require predictors to be on the same scale with the caveat of being less precise (ibid.: 14).[7] Since no multivariate regression splines were used in any of the studies in this thesis, these aspects of regression splines will not be discussed further here. For details please see Baayen et al. 2017; Wood 2003, 2006.

11.4. Penalization of Splines

It has been mentioned previously, that while the overall shape of the regression spline is determined by its basis functions, its concrete form is determined by its coefficients β. Therefore, spline penalization typically targets β, while the number of basis functions is neglegible. Put differently: If the number of basis functions is set to some initial K, *penalized* parameter estimation will suffice to control a smooth's wiggliness, because superfluous basis functions may simply be muted by zero coefficients. Thus, if the number of basis functions is set to some initial K, penalized parameter estimation will suffice to prevent the model from over-fitting. A potentional issue when setting K is under-smoothing, though, due to a lack of

[7] Tensors describe linear mappings, such as dot or cross products, between geometric vectors, scalars and other tensors. The mathematical details of tensors are beyond the scope of this thesis, please see Kolechi 2002 for an overview.

basis functions.[8] Therefore, K should ideally be slightly above the optimal number of basis functions. To ensure that K is not set too low, it is often suggested to compare a model with K basis functions to its pendant with $2K$ basis functions, to see if this increases model performance (Baayen & Divjak 2016; Wood 2006).[9]

After setting the number of basis functions to K, the smooth's wiggliness is controlled by including a penalty term to the estimation of β. For simple additive models, i.e. models whose response follows a normal distribution, this is done by minimizing the *penalized* RSS Q of the regression spline $s(x)$, as shown in Equation 11.9.

$$Q = \sum_{i=1}^{N}(y_i - s(x_i))^2 + \lambda \int (s''(x))^2 dx \tag{11.9}$$

In this set up, Q may be increased independent from the plain RSS $(y_i - s(x_i))^2$ due to the integrated, squared second derivative of the smooth, which serves as penalty term. Intuitively, this penalizes more wiggly curves, as the second derivative measures the slope of slopes, which increases with decreasing smoothness. The weight placed on this penalty is determined by the smoothing parameter λ. If $\lambda = 0$, Q is measured only in terms of model fit as measured by RSS, while disregarding smooth complexity. This results in an *un-penalized regression smooth*. Higher λ values penalize undersmoothing of complex splines and thus enforce smoother curves (Baayen et al. 2017; Larsen 2015). Hence, λ has to be estimated such, that a good trade-off between smoothness and model fit is reached. Note, that when dealing with GAMs, i.e. allowing for response variables following non-normal distributions, β is to be fitted by penalized likelihood estimation (Wood 2006: 136). This is straight forward, as the same applies to GLMs, too. One commonly suggested type of penalized likelihood estimation is Penalized Re-weighted Iterative Least Squares (P-IRLS) (Larsen 2015; Wood 2006). As the name suggests, P-IRLS extends IRLS (cf. Section 11.2) to GAMs, by introducing λ as penalty for wiggliness. For a more detailed discussion of P-IRLS, please see Larsen 2015: 12ff or Wood 2006: 126ff.

The smoothing parameter λ may be estimated by two approaches: i) by generalized cross-validation criteria (GCV), which chooses λ such that the mean prediction

[8] Note that there is a scenario in which an excessive number of basis functions may pose a problem, though, namely when the model is not provided with sufficient data support to estimate all parameters. Also, data with severe outliers might require a reduction of basis functions, too.

[9] Note that this comparison should not rely on a comparison of the explained variance, since additional basis functions are likely to reduce the model's residual error. Instead, the trade-off between model fit and consumed effective degrees of freedom (edf) should be tested by significance testing, since additional basis functions are bound to consume more edf, as they require the estimation of more parameters.

error is minimized when fitting the model in a leave one out cross-validation scenario (Larsen 2015: 16), see Wood 2006: 126ff for details on this approach; ii) by restricted maximum likelihood (REML) in a Bayesian approach, which is used in this thesis. It is based on drawing random β coefficients from a prior distribution of β and re-fitting λ such that the drawn coefficients have a high average likelihood (Baayen & Divjak 2016: 7). This β prior follows a normal distribution with mean zero an a variation linked to the smoothing penalty: For $\lambda = 0$ all β are free to vary, while variance of β decreases for increasing λ, which ultimately leads to smaller β as they deviate less from their zero mean (Larsen 2015: 15). How much β decreases due to the penalty is measured by the *shrinkage factor* or edf, which ranges from 0 to 1 to indicate how variable β is: A basis function with an edf of 1 does not receive any penalty for wiggliness. In the case of regression splines, this is the case for the first two basis functions, which are the intercept and a linear function. With an increasing number of basis functions, though, the penalty is bound to increase as the functions' wiggliness constantly increases. Accordingly, the edf of the smooth is going to decrease. This makes the non-linearity of a smooth interpretable from its cumulative edf: Smooths for a linear predictor will show edf of approximately 1, while higher edf indicate that the smooth models in fact a non-linear relationship (Baayen et al. 2017: 8f).

11.5. Additive Mixed Models

So far, only fixed-effect models have been discussed. These models work based on the assumption of a completely random error ϵ. *Mixed-effect* models may account for known structures in the underlying noise. One common example for such structures are *repeated measures*, which violate the assumption of independence in fixed-effect models. Such repeated measures are present in study 2 in this thesis (cf. Chapter 14), for example, which is based on a longitudinal learner corpus (cf. Chapter 6) and thus contains multiple data points elicited from the same learner. Therefore, it seems in order to give a brief impression on how GAMs may be extended to Generalized Additive Mixed Model (GAMM). GAMMs may account for by-subject variations in the model by means of three types of random effects: i) random intercepts, ii) random slopes, and iii) random smooths. Equation 11.10 shows a the GAM from Equation 11.6 with a random intercept b_j, that introduces

by-subject variation to the intercept β_0.

$$g(y_j) = \eta + b_j + \epsilon_j, \text{ where } \epsilon_j \sim N(0, \sigma^2),\ b_j \sim N(0, \sigma_b^2), \tag{11.10}$$

with η as defined in Equation 11.6. Note that b_j is drawn from a normal distribution wit zero mean and a homoscedastic by-subject variance σ_b^2, i.e. by-subject variance in the intercept is treated as a sort of noise, like the random error ϵ. Furthermore, note that the random error ϵ, too, is subject dependent in this model. GAMMs may also additionally include random by-subject slopes. These two types of random effects are similar to random effects in regular linear mixed model. However, note that the estimation of random effects in GAMMs is computationally more costly than for regular linear mixed models (Baayen & Divjak 2016: 9). This thesis dispenses with a more elaborate account on parameter estimation in GAMM, though, due to the marginal role mixed-effects play in Study 2 in Chapter 14. For the same reasons, the topic of by-subject smooths will not be discussed. Please see Baayen & Divjak 2016; Wood 2006 for details.

11.6. Ordinal Regression

It was noted earlier in Section 11.2, that while GLMs and by extension GAMs allow for response variables, that are not normal distributed by means of monotonic link functions, they still require responses to belong to some distribution from the exponential family. Ordinal data, such as operationalizations of proficiency levels, as they are investigated in this thesis, do not belong to the exponential family. As Baayen et al. 2017 points out, in practice this often leads to the re-analysis of such responses to some data type, that may be captured by common link functions. Ordinal data may, for example, be re-analyzed as multinomial or binomial data, in order to capture the categorical nature of the response. This looses all information on the intrinsic order of the levels, though. Another, more common approach is to re-analyze the data as numeric, which preserves the intrinsic order of values. However, this ignores, that distances between ordered levels are not meaningful in the way distances between numeric data are (Baayen & Divjak 2016: 2). A prominent example for this are *Likert* scales, which measure consent typically on a five-level scale from strong agreement to strong disagreement. They are commonly mapped to integer values ranging from 1 to 5 for statistical analyses. This introduces a quantifiability to the level precedence, that does not apply to ordinal data: While

2 is twice as high as 4, it does not make sense to quantify the difference between *agree* and *disagree* in that manner. The same holds for proficiency levels: While $B2$ learners are certainly more proficient than $A2$ learners, they cannot be said to be twice as proficient.

Clearly, neither of the two approaches adequately represents ordinal data, since they fail to capture important characteristics of the data type. Wood, Pya & Säfken 2016 address this issue, by introducing methods for estimating smoothing parameters in GAMs with responses from non-exponential distributions, such as ordered categorical responses. They approximate ordinal responses through a latent numeric variable $u = \mu + \epsilon$. This variable is then linked to the ordinal estimated response $\hat{y}$, cf. Equation 11.11.

$$u = \eta + \epsilon, \ \textit{where } \eta = \beta_0 + \sum_{i=1}^{I} s_i(x_i) \tag{11.11}$$

u is bound to fall within the limits of $\pm\infty$. This real-valued range is partitioned into C intervals given an ordinal response with C levels. In the case of CEFR scores, this would be $C = 6$. Intervals are defined by estimating $C + 1$ boundaries $\alpha_{c=0}^{C}$ on the scale $\pm\infty$, such that $\alpha_0 = -\infty$, $\alpha_C = \infty$. The inner boundaries, which are not set to $\pm\infty$, i.e. α_1^{C-1}, are estimated by penalized likelihood maximization during model fitting alongside the other model coefficients, in order to obtain C intervals on that scale (Baayen & Divjak 2016: 5).[10] After parameter estimation, u_i is mapped to level $\hat{y}_c$, if $\alpha_{c-1} < u_i \leq \alpha_c$ and the probability of obtaining category c can be estimated by means of a cumulative distribution function $F(u)$ (ibid.: 5):

$$\begin{aligned} P(y = k) = P(\alpha_{k-1} \leq u_k \leq \alpha_k) = F(\alpha_k - \mu) - F(\alpha_{k-1} - \mu) \\ \textit{where } F(u) = \frac{e^u}{1 + e^u} \end{aligned} \tag{11.12}$$

Figure 11.7 shows some example boundaries α_0^C which map u to CEFR scores. The boundary values were estimated for the *interaction* model presented in Study 1 (cf. Chapter 13). Given this set up, the model predicts a B1 score for a text, if $\alpha_2 < u_i \leq \alpha_3 = 4.92 < u_i \leq 10.48$. Note, that for this study CEFR scores C1 and C2 were combined, hence, the number of levels is $C = 5$, and the outer boundaries α_0, α_5 at $\pm\infty$ were not depicted. Therefore, the scale is partitioned by four estimated boundaries.

[10] The technical details of boundary estimation go beyond the scope of this thesis. Please see Wood, Pya & Säfken 2016 for a more elaborate discussion.

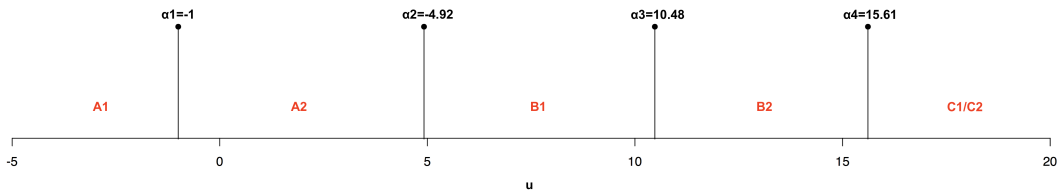


Figure 11.7.: Mapping of latent variable u to CEFR levels using estimated boundaries from the *interaction* model of study 1, cf. Section 13.2.

Part VI.

Empirical Studies

12. Exploratory Model Design

This chapter discusses the general set up of the following inferential regression studies, that were conducted on both corpora. Again, complexity measures were extracted from the close transcriptions of the learner data using the system described in Chapter 9, see the discussion at the beginning of Chapter 10 for details.

However, unlike before, the quantity of measures, that are employed in the studies, is more restricted, because the number of predictors, that may be supported by a regression model, is limited by the amount of available data. For this more rigorous selection, measures were selected based on a data-driven feature selection approach rather than theoretical considerations. The following section briefly outlines i) the details behind the restriction of predictors, that may be used in a regression model; ii) potential approaches to feature selection; and iii) which approach was chosen for this study and how the identification of candidate predictors was performed. Section 12.2 then describes the iterative model design approach used throughout all regression studies.

12.1. Feature Ranking

The number of predictors in a regression model that may be supported by a data set is limited by its size. A common practical heuristic to define an adequate maximal number of predictors is, that a data set of size n should not support more than between $\frac{n}{20}$ and $\frac{n}{15}$ predictors (Babynak 2004; Peduzzi et al. 1995, 1996). Exceeding this threshold makes the model increasingly prone to over-fitting, which occurs when a model measures idiosyncratic properties of the sample data, which fail to generalize well for the whole population. This may be caused by using too many predictors, because a model's *degrees of freedom* decrease with an increasing set of predictors: The estimation of each predictor's parameters consumes degrees of freedom, i.e. the amount to which data points are free to vary in the model. A model's degrees of freedom may be calculated by subtracting the number of estimated parameters from the number of data points used to train the model

(Babynak 2004: 414). If a model consumes too many degrees of freedom by exceeding the number of predictors, that may be supported by the data, the model is not free to vary anymore and depicts idiosyncratic properties of the sample rather than generalized characteristics of the population. This heuristic of 15 to 20 data points supporting each predictor is based on GLMs, though, which estimate one additional parameter for each new predictor. GAMs may estimate more than one parameter per predictor, if the predictor is smoothed, cf. Sections 11.2 and 11.3. Thus, depending on the amount of non-linearity, that a GAM has to account for, an even larger data support per predictor may be required.

Clearly, neither *Merlin* nor *Falko Georgetown L2* include enough data points to entertain all 398 complexity measures in a regression model. This lack of data support given the number of potential predictors is a common issue in regression modeling and there are two general strategies to solve it: i) A set of candidate predictors may be ranked by some criterion such as variability across or informativeness with respect to the response variable. This procedure is referred to as *feature screening*. Measures are then added iteratively to the model, after confirming their normality and that they are uncorrelated. This type of approach is for example employed for complexity-based regression analyses by Crossley & McNamara 2009; Crossley et al. 2010; McNamara, Crossley & McCarthy 2009. ii) Another approach to shrink the amount of measures in a large feature set, in order to get a more manageable amount of measures for a regression analysis, is *feature aggregation*. In this procedure, measures that correlate with each other are combined to a single measure, for example by clustering, factor analysis, or principal component analysis (Babynak 2004: 419), see for example Biber, Gray & Staples 2014 for a factor analysis approach in language analyses.

While feature aggregation reduces the number of predictors and eliminates unwanted correlations between independent variables for a regresison model (Boslaugh & Watters 2008: 298ff), it may lead to not necessarily theoretically interpretable feature groupings. Thus, for the following studies, measures were ranked in terms of their informativeness and then used to iteratively augment the regression models. Informativeness was assessed by information gain. This is a measure from information theory, which assesses the expected reduction in entropy for some class (the response variable) given a feature (the complexity measure for which the information gain is calculated). Information gain was calculated using the Waikato Environment for Knowledge Analysis (WEKA) machine learning suit API, version 3.8 (Frank, Hall & Witten 2016; Hall et al. 2009). Then, the most

informative predictors were used to augment an increasingly complex regression model in an iterative approach. For this, an initial GAM was set up, which predicts L2 proficiency from only the designated task factor, that was chosen for the given analysis. This and all following steps were conducted in R version 3.3.3 (R Core Team 2017), using the packages *mgcv* (Wood 2003, 2004, 2011) and *itsadug* (van Rij et al. 2016) for regression modeling and model evaluation. Furthermore, the *caret* package (Kuhn et al. 2016) was used for random stratified partitioning when conducting 10-folds cross-validation.

12.2. Model Augmentation

Complexity measures from the ranked feature collection were added to the initial model in a step-wise approach:

1. The most informative complexity measure as established by information gain ranking (cf. above), that had not been used in the model up to this point, was selected.
2. The designated measure was transformed to approximate a normal distribution more closely. This was achieved by means of square root or log transformation. Measures that could not be transformed to approximate a normal distribution, for example due to a high amount of zero values, were discretized to binary or ternary measures.
3. The Pearson rank correlation between a designated transformed measure and all measures already included in the most recent model was calculated. If a designated measure ...
    a. ... did not exhibit a correlation more extreme than ± 0.70 to any of the predictors included in the most recent model, it was added to the model.
    b. ... did exhibit a correlation more extreme than ± 0.70 to any of the predictors included in the most recent model, all correlated predictors were removed from the most recent model in exchange for the designated measure.[1]
4. The designated measure was added with a smooth, unless it was a discrete measure. Thin-plate regression splines were used for smooths. Initially,

[1] Evaluating feature correlation with Pearson rank correlation and setting the threshold to ± 0.70 follows the approach outlined in Crossley et al. 2010.

the default number of 9 knots was used, but it was always tested whether doubling the number of knots increased model fit significantly. If smoothed measures proved to be non-linear, they were checked for concurvity with any of the other smoothed measures in the model: Concurvity is a measure similar to colinearity, but applies to non-linear functions. It is bound between 0 and 1 and indicates whether smooths encode redundant information (cf. help page for *concurvity* in *mgcv* package). If measures showed concurvity higher than 0.70 they were treated equivalent to measures with too extreme correlations, cf. Step 3.

5. The model updates performed in Steps 3 and 4 were kept, if they lead to a significantly better model fit. Model fit was evaluated using the *compareML* function from the *itsadug* package, which performs a "Chi-Square test [...] on two times the difference in minimized smoothing parameter selection score [ZW: i.e. REML in the case of GAMs] [...], and the difference in degrees of freedom specified in the model." (help page of *compareML* from the *itsadug* package) assuming $\alpha = 0.05$.

The augmentation process was set to abort after 20 iterations had passed without yielding a significantly better model or after the threshold of 20 data points per predictor was reached.

After choosing the most informative predictors, in a second step, all measures were centered around their mean, before potential interactions between all complexity measures and the designated task factor were introduced and evaluated using model comparison. Due to the centering, differences between task factors become more interpretable, since the comparison is conducted between the mean values of the measures. Interactions had to satisfy two requirements to be admitted to a model: i) When added to the model, they needed to lead to a significantly improved model fit compared to the model without this interaction. For significance testing, a χ^2 test was conducted for $\alpha < 0.05$ using the *compareML* function from the *itsadug* package. ii) Given a significant model improvement, interactions were required to show at least one significant difference in slopes across task factor levels for at least $\alpha < 0.10$, which stayed stable across 10 iterations of 10-folds cross-validation for more than 80% of the trials.

13. Modeling L2 Proficiency in Merlin

This chapter studies German L2 proficiency in the German *Merlin* corpus using the data set outlined in Section 5.3. The German section of the *Merlin* corpus is one of the largest collections of German learner data. It has already been used in previous proficiency assessment studies based on complexity measures (Hancke 2013), but the effect of task factors on these analyses has not yet been addressed properly. The following studies investigate the informativeness of functional task factors for complexity analyses on the *Merlin* data based on *task theme*.[1] For this, the original research questions formulated in the introduction were specified as follows:

1a. Which complexity measures are elicited for model design in an exploratory, data-driven approach?

1b. What do these measures show about L2 proficiency?

2a. Are these measures influenced by task theme?

2b. If so, how are measures influenced by task theme?

3a. How well do the most suited complexity measures perform in a classification task?

The next section briefly outlines the set up of the studies conducted in *Merlin*, before Section 13.2 reports and discusses a series of GAMs, that predict assigned overall CEFR scores from complexity measures and task factors. Then, an ancillary study on performance effects is conducted in Section 13.3. While performance effects are not part of the main research questions guiding these studies (cf. above), the residual analyses of the models in Study 1 indicated, that it could be worthwhile to account for differences between test level and proficiency scores. This is supported by a study by ibid., who finds that complexity measures are sensitive to performance differences in *Merlin*, as well as by theoretical considerations: It seems reasonable to assume, that learners, who outperform a test for a given level, might exhibit different language characteristics than learners, who are over-strained by their test.

[1] Please see Section 13.1 for an explanation of the choice of task factor.

Hence, an ancillary study including performance effects was added as excursion. Both studies include classification experiments additional to the discussion of the models and their fit the data to assess the predictive power of the models. The chapter closes with a brief summary of the findings in Section 13.4.

13.1. Set Up

On *Merlin*, L2 proficiency was operationalized as the holisitic overall CEFR score, which was assigned to each text by expert raters, cf. Section 5.1. Furthermore, *task theme* was chosen as task factor for the analysis of task-effects, because i) it is a commonly investigated functional task factor in SLA research, cf. Chapter 4; and ii) it is one of the two factors identified as most suited for statistical analyses with respect to its variability across proficiency scores and test levels, as well as its lack of idiosyncratic structure in its distribution across test levels, cf. Section 7.2.[2] Complexity measures were extracted from the close transcriptions of the *Merlin* texts by employing the system described in Chapter 9. These were preferred over the target hypotheses provided for each text, for the reasons already discussed at the beginning of Chapter 10.

The resulting set of 398 complexity measures (cf. Chapter 8) was reduced by employing the feature screening procedure described in Chapter 12. Then, a GAM with overall CEFR score as response and task theme as an a priori given predictor was augmented with complexity measures following the procedure outlined in Section 12.2. The augmentation process was aborted after 153 iterations, because the final 20 iterations were completed without obtaining an improved model. This lead to a final model, with overall 13 complexity measures including members of all four main domains included in the full feature collection: language use, human language processing, discourse and meaning, and theoretical linguistics, in particular syntax, lexicon, and morphology, cf. Chapter 8. Four of the complexity measures were binarized before being added to the model. All other complexity measures are continuous.

[2] *Task type* would have been an equally good candidate for analysis. The analysis of more than one task factor was unfortunately beyond the scope of this thesis, though. Thus, it was decided to leave the analysis of *task type* for future work.

```
gam.merlin.interactions <- gam(OverallCefrScore ~
                                 hasTransitionsFromSubjectToNot +
                                 has3rdPersPossessivePronouns +
                                 containsToInfinitives +
                                 usesConjunctionalClauses +
                                 halfModalClusterPerVP +
                                 logSumNonTerminalNodesPerSentence +
                                 logATFBand2PerTypesFoundInDlex +
                                 avgVTotalIntegrationCostAtFiniteVerb +
                                 lexTypesFoundInDlexPerLexType +
                                 typeTokenRato +
                                 logSumNonTerminalNodesPerWord +
                                 usesConjunctionalClauses:TaskTheme +
                                 logATFBand2PerTypesFoundInDlex:TaskTheme +
                                 typeTokenRato:TaskTheme +
                                 sumNonTerminalNodesPerWord:TaskTheme +
                                 s(charactersPerWord) +
                                 s(numberOfSentencesSquared) +
                                 TaskTheme,
                               data=merlin,
                               family=ocat(R=5))
```

Figure 13.1.: Model formula of *Merlin* interaction model predicting overall CEFR scores from scaled and transformed complexity measures.

13.2. Study 1: Modeling Task-Effects

13.2.1. Model Description

The first study investigates L2 proficiency and how task-effects influence the analyses. For this, three different models were built from the set of potential measures: The *reference* model includes task theme as a predictor, but does not use any interactions. It contains overall three univariate thin-plate regression spline smooths to account for non-linear relations to the response. Also, a *complexity* model was build by removing task theme as a predictor. This model requires overall five univariate thin-plate regression spline smooths to account for non-linearity. These two models serve as grounds for comparison with the *interaction* model: For this model, interactions between the 13 complexity measures and task theme were introduced. The *interaction* model is shown in Figure 13.1. It features four interactions between complexity measures and task theme and contains two smooths.[3] For all three models, the default number of nine knots was sufficient to capture all non-linearity, i.e. doubling the number of knots did lead to more wiggly smooths. Also, concurvity was less extreme than 0.70 for all smooths in the three models.

[3]Model formulas for the reference and the complexity model may be found in Figure A.1 in Appendix A.

Model	AIC	Df	REML	Edf	Compared with	χ^2	Edf difference	$Pr(>\chi^2)$
Complexity	1315.05	30.37	658.56	19				
Reference	1287.08	28.41	642.77	20	Complexity	15.790	1	1.914e-08
Interaction	1281.00	39.27	628.84	31	Complexity	29.717	12	2.861e-08
					Reference	13.928	11	0.003

Table 13.1.: Model comparison for complexity, reference, and interaction model build on the *Merlin* data.

13.2.2. Model Fit

When comparing all three models, the interaction model captures more deviance than the other models: It has $R^2 = 0.7660$, while the others explain $R^2 = 0.7540$ for the reference model and $R^2 = 0.7410$ for the complexity model.[4] Although these differences are small, when comparing all models using χ^2 tests, the deeper integration of task theme information systematically leads to significantly improved performance, as may be seen in Table 13.1. By investing 11 (respectively 12) edf, the interaction model gains a moderate decrease in REML score, which is highly significant for $\alpha < 0.01$. Similarly, the reference model gains a significantly improved REML score compared to the complexity model by investing 1 edf. Hence, overall, the observable differences between models are in fact significantly more informative.

When analyzing the residual errors across all three models, the results show overall good model fit without any clear structures: The distribution of the standard deviation is homoscedastic for all models, see Figure 13.2 below and Figure A.4 in Appendix A for the respective plots of the full residual errors in the upper left. Furthermore, the mean residual errors are close to zero, with $\mu = -0.27; \sigma^2 = 26.18$ for the reference model and $\mu = 0.11; \sigma^2 = 22.28$ for the interaction model. It is a little off, though, for the complexity model with $\mu = -1.02; \sigma^2 = 43.25$. Despite the more or less adequate mean values, the wide standard deviations are clearly troubling. s the visual inspection shows, all three models exhibit few but severe outliers, i.e. data points whose residual errors deviate more than two standard deviations from the mean: For the reference model, there are 8 outliers, for the interaction model, there are 10 outliers, and for the complexity model, there are 5 outliers. These are the cause of the wide standard deviations (cf. below).

These outliers and other wide spread residual errors were investigated further,

[4]Note, that these results are based on the *Merlin* data without the four most sever coutliers, cf. below. Only the first description of residual errors below is based on results obtained for the full data set.

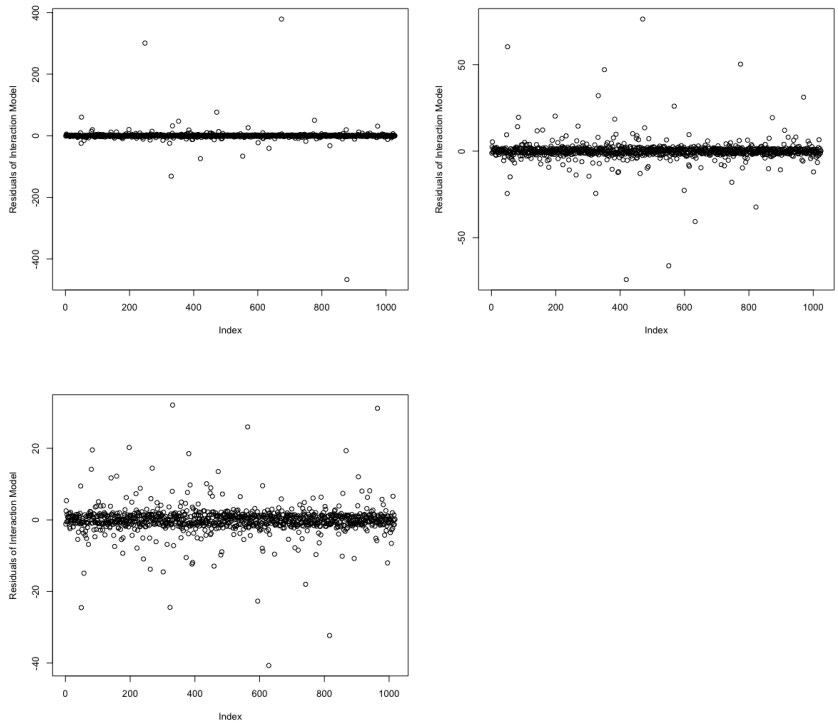


Figure 13.2.: Residuals of *Merlin* interaction model on i) full data (upper left); ii) data without 4 most severe outliers (upper right); iii) data without any outliers (lower left).

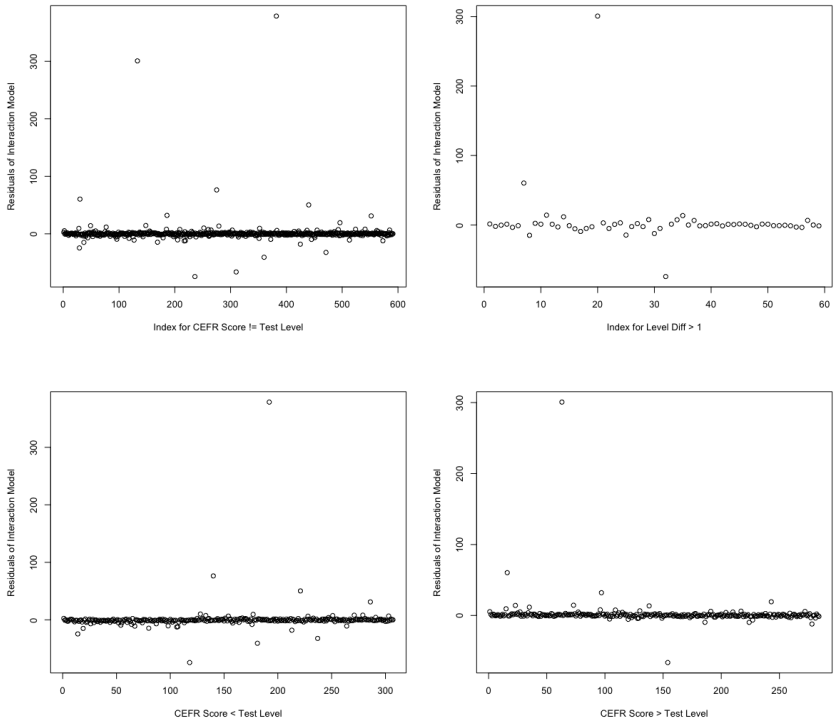


Figure 13.3.: Residuals of *Merlin* interaction model for test level and CEFR score mismatches: i) score and test level differ (panel 1); ii) the difference is > 1 (panel 2); iii) score $<$ test level (panel 3); iv) score $>$ test level (panel 4).

in order to identify a pattern, which might explain, what causes these data points to deviate from the rest of the data. It was found, that the most extreme residual errors were highly redundant across all three models and that the four most severe outliers are caused by the same texts across all models. Also, high residual errors occur nearly exclusively for texts, which were rated as much as two levels higher or lower than their test level: The – across models consistently – most extreme outlier receives an A1 score in a B1 test, i.e. severely underperforms. Most other data points with relatively large residual errors belong to texts, which outperformed their test level by up to 2 levels. Yet, most learners who over- or underperformed even as much as two test levels had residual errors close to zero across models. Figure 13.3 shows the residual plots for various configurations of CEFR score and test level mismatches for the interaction model.[5] Thus, it does not seem like the

[5]Figure A.5 in Appendix A shows the respective plots for the reference and the complexity model.

models systematically fail to account for over- and underperforming writers in general. Unfortunately, no further common pattern could be identified. However, the findings prompted the ancillary study in Section 13.3, which includes task performance as an independent variable to the models.

When investigating how much the outliers effected the models, it was found, that only removing the four consistently most severe outliers showed considerable changes in model fit: It was found that removing these four most severe outliers lead to considerabe model improvement in terms of i) strengthened significance of the predictors; ii) higher R^2 scores; and iii) lower REML scores, while for neither of the models, removing any additional outliers had similar effects.[6] A comparison of the REML scores and AIC for all three versions of the reference model may be found in Table A.7, in Table A.6 for the interaction model, and in Table A.8 for the complexity model in Appendix A. Due to these findings, the four most severe outliers were removed from the data, while all others were kept in the models: All evidence indicates, that these four outliers exhibit some idiosyncratic structure, which does not generalize across other data points, influences model fit, and cannot be pinned down to a specific property that could be included to the model. This lead to mean residual errors and standard deviations of $\mu = 0.05, sd = 5.74$ for the reference model, $\mu = 0.04, sd = 7.26$ for the interaction model, and $\mu = 0.15, sd = 6.39$ for the complexity model, see the upper right in Figures 13.2 and A.4 in Appendix A.

13.2.3. Model Discussion

Table 13.2 shows the model summary for the interaction model. Since the reference and complexity model serve as grounds for comparison, but are very similar to the interaction model, similarities and differences between the models are pointed out while discussing the interaction model. Tables A.1 and A.2 in Appendix A show the model summaries for the reference and the complexity model. The appendix also includes summary tables for the interaction model, which use the other task themes as reference levels, cf. Tables A.3, A.4, and A.5. Furthermore, Figure 13.4 shows the component smooth functions of the two non-linear measures of the interaction

[6] Note, though, that this reasoning is not based on significance testing, because a χ^2 is not admissible for model comparisons used to decide on outlier removal, because it is bound to be widely uninformative: The test is heavily based on comparing differences in edf, which are bound to be highly similar, when fitting the same model on an only marginally smaller version of the same data set. Hence, it remains at the discretion of the researcher to evaluate Akaike's 'An Information Criterion' (AIC) and REML differences according to their research objectives, cf. help page of compareML (van Rij et al. 2016).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	8.3759	0.3833	21.8509	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.5349	0.2387	-2.2408	0.0250
has3rdPersPossessivePronouns[TRUE]	-0.8906	0.2030	-4.3873	< 0.0001
containsToInfinitives[TRUE]	-0.5541	0.2282	-2.4284	0.0152
usesConjunctionalClauses[TRUE]	-0.6051	0.3173	-1.9074	0.0565
halfModalClusterPerVP	0.1831	0.1011	1.8113	0.0701
logSumNonTerminalNodesPerSentence	1.9714	0.1785	11.0435	< 0.0001
logATFBand2PerTypesFoundInDlex	-0.3003	0.1091	-2.7528	0.0059
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1059	3.4968	0.0005
lexTypesFoundInDlexPerLexType	0.8840	0.0942	9.3858	< 0.0001
typeTokenRato	1.2853	0.2038	6.3068	< 0.0001
logSumNonTerminalNodesPerWord	-0.7130	0.1598	-4.4619	< 0.0001
TaskTheme[Society]	0.4921	0.7085	0.6947	0.4873
TaskTheme[Profession]	1.0774	0.5508	1.9560	0.0505
TaskTheme[Smalltalk]	-0.8117	0.3529	-2.3004	0.0214
usesConjunctionalClauses:TaskTheme[Society]	2.1839	0.9603	2.2742	0.0230
usesConjunctionalClauses:TaskTheme[Profession]	-0.4185	0.5417	-0.7726	0.4398
usesConjunctionalClauses:TaskTheme[Smalltalk]	0.5155	0.4714	1.0937	0.2741
logATFBand2PerTypesFoundInDlex:TaskTheme[Society]	-0.1827	0.4194	-0.4357	0.6631
logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]	0.5517	0.3530	1.5628	0.1181
logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]	0.5392	0.2197	2.4539	0.0141
typeTokenRato:TaskTheme[Society]	-0.4750	0.3634	-1.3072	0.1912
typeTokenRato:TaskTheme[Profession]	-0.5975	0.3998	-1.4947	0.1350
typeTokenRato:TaskTheme[Smalltalk]	-0.8335	0.2925	-2.8494	0.0044
logSumNonTerminalNodesPerWord:TaskTheme[Society]	-0.9369	0.4216	-2.2224	0.0263
logSumNonTerminalNodesPerWord:TaskTheme[Profession]	-0.1522	0.3409	-0.4465	0.6552
logSumNonTerminalNodesPerWord:TaskTheme[Smalltalk]	0.2680	0.2344	1.1429	0.2531
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord)	2.7714	3.5484	18.5670	0.0007
s(numberOfSentencesSquared)	4.6262	5.7193	254.0399	< 0.0001

Table 13.2.: Interaction model predicting *Merlin* overall CEFR scores from scaled and transformed complexity measures fitted on *Merlin* data without the four most severe outliers. Uses 'demand' as reference level for task theme.

model scaled to their linear predictor (cf. help page of *plot.gam* from the *mgcv* package). It is necessary to plot the smoothed measures, because unlike for the parametric effects, the shape of the smooths cannot be fully determined from the summary table alone. The grey shades indicate the standard error for each plot. Each plot's data support is shown at its x-axis. Plots of the smooths for the reference and complexity model may be found in Figures A.2 and A.3 in Appendix A.

Length measures The two non linear measures in the model are highly superficial length measures: the *number of characters per word*, i.e. word length, and the *squared number of sentences*, i.e. text length. The *number of characters per word* increases significantly in a close to linear manner for $\alpha < 0.01$, but only after passing a

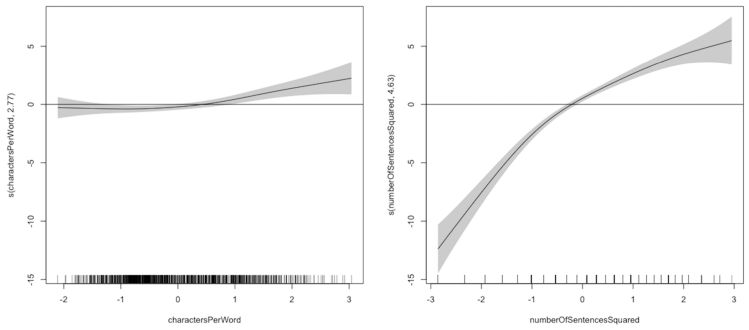


Figure 13.4.: Smooths of *Merlin* interaction model.

threshold, which is roughly around zero, cf. Figure 13.4, left plot. Note that due to the centering of all measures, zero indicates the mean value. Thus, the threshold indicates that if words contain more characters than they do on average, more characters are indicative for higher proficiency. Shorter than average words are not indicative of lower proficiency, though, which is straight-forward, since functional words fall into this category, which are not to be expected to indicate lower proficiency. The *squared number of sentences* significantly increases with increasing proficiency for $\alpha < 0.01$: The measure shows a steep increase, which levels off for higher proficiency, following approximately the shape of a logarithmic function, cf. Figure 13.4, panel 2. Put differently, this shows, that more sentences are related to more proficient writing, but that for higher proficiency levels, differences in the number of sentences become less informative.

Unfortunately, neither of these two measures may be mapped directly to a certain concept, because they operate globally and are influenced by multiple factors, which is also likely to contribute to their non linearity: Word length, for example, may be increased by derivation or composition, but it also decreases, when using many frequent words, which are often shorter than less frequent words, or when using a high amount of function words. The number of sentences is even less interpretable, because it may be influenced not only by various linguistic but also by technical factors: More sentences might indicate longer texts and thus measure writing fluency. Yet, a shift from paratactic to hypotactic writing, i.e. increased use of dependent clauses, might counteract this increase, thus causing the leveling off of the effect for more proficient writers. On uncorrected learner data, the number of sentences is also highly influenced by proper punctuation, though, since automatic sentence segmentation as applied here is predominantly based on

sentence boundary identification, i.e. it only separates sentences at punctuation marks, see Hancke 2013: 22ff for a brief survey. Thus, spurious punctuation is bound to result in fewer sentences. Since ibid. also reports, that spurious punctuation is in fact an issue for texts of less proficient writers in the *Merlin* corpus, the stronger effect for sentence length for less proficient writers is likely to be at least partially due to insufficient punctuation and the resulting repercussions on sentence boundary identification. While this does not invalidate the overall finding of increases in the number of sentences for more proficient writers, it makes the non-linearity of the effect difficult to interpret without further investigation.

Syntactic or grammatical complexity Syntactic or grammatical complexity is assessed by two global measures of the grammatical system, namely the *number of non terminal nodes per words* and *per sentence*, and by two measures of structure complexity, namely the *number of half modal verb clusters per verb phrase* and the *use of conjunctional clauses*. The two global measures are, again, not straight forward to interpret, because they are sensitive to an abundance of changes in the grammatical system and likely to be caused by multiple factors. However, they do allow for some conclusions, when considered together with other measures. For this, it is helpful to briefly look at why both non terminal node measures exhibit inverse slopes: The *number of non terminal nodes per sentence* shows significant increases for increasing proficiency levels at $\alpha < 0.01$. In principle, this could be caused by either an increase in the number of non terminals or a decrease in the number of sentences. The latter, however, is clearly not the case, since the *squared number of sentences* was found to increase significantly for increasing proficiency. Thus, the increase has to be due to an increase in the number of non terminal nodes for higher proficiency levels, which has to be more extreme than the increase in the number of sentences. Interestingly, increases of the *log sum of non terminal nodes per word* are significantly correlated with decreasing proficiency throughout all task themes for $\alpha < 0.05$ albeit with differences in the steepness of the respective slope, since this measure interacts with task theme: When using *demand* as reference level, the slope significantly levels off for *small talk*, while the strongest effect is found for *society* texts; see the explanation for the differences between *demand* and *society* texts in Study 2 in Section 13.3.3, though. If with increasing proficiency the number of non terminal nodes increases even more than the number of sentences and yet simultaneously the *ratio of non terminal nodes per word* decreases, the number of words per text has to increase for increasing proficiency levels, too, and again,

more extreme than the increase for non terminal nodes. Unfortunately, there are several linguistic and technical reasons which might influence the number of non terminal nodes that are introduced per word: coordination of phrases and clauses leads to fewer nodes in non-binary parses, for example. Increased parse accuracy, too, might decrease the ratio of non terminal nodes to words, though, since some types of parsing errors tend to inflate the number on non terminals by introducing superfluous nodes to the structure. Thus, without further evidence, it is not possible to draw any certain conclusions from these findings.

There are three structural measures of grammatical complexity in the interaction model, two of which measure phrasal complexity, or more precisely complexity of the verbal domain: The *ratio of half modal clusters per verb phrase* and the *use of to infinitives*. Interestingly, these two measures are closely linked, since half modals are defined as *haben* (to have), *sein* (to be), *scheinen* (to seem), *drohen* (to threaten), and *versprechen* (to promise), if they govern a *to* infinitive (§101 Duden (Gr) 2009: 101). Yet, both measures show inverse slopes: using *to* infinitives is associated with lower proficiency for $\alpha < 0.01$, while the ratio of half modal clusters shows marginally but constantly significant increases for increasing proficiency with $\alpha < 0.10$, which is stable at its significance level across 10 iterations of 10-folds cross-validation for more than 80% of the iterations and thus assumed to be genuine. Since half modal clusters by definition can only occur in texts that contain *to* infinitives, the most plausible interpretation of these two findings is, that *to* infinitives tend to occur in texts of less proficient writers, but if they occur in a text, more proficient writers use them in half modal clusters.

The third structural measure of grammatical complexity in the interaction model is the *use of conjunctional clauses*, which measures clausal complexity. It interacts with task theme and shows highly heterogeneous effects on proficiency across task theme: In *demand* and *profession* texts, conjunctional clauses have a negative slope, which is marginally significant at $\alpha < 0.10$, but consistent across 10 iterations of 10-folds cross validation more than 80% of the trials. For *society* texts, the slope is increasing with $\alpha < 0.05$ and for *small talk*, the slope is significantly less extreme, than for either of the other texts and there is no significant effect for conjunctional clauses in *small talk* texts to be found. These inverse slopes across task theme lead to the measure being insignificant in the complexity model. For *society* texts, the significant positive correlation with proficiency seems to satisfy a straight forward functional need of argumentative texts: The formulation of opinions and arguments about society necessarily elicit conjunctional clauses such as adverbial clauses. The

marginally significant decrease for *demand* and *profession* is less expected, but could be due to a preference for other types of subordinate clauses, such as relative clauses. Since there are interpretations of the findings, where there are increases in other types of subordinate clauses, these results do not allow to draw conclusions about clausal complexity for any texts other than *society* texts, for which at least this type of subordinate clause is indicative of higher proficiency.

Taken together, the findings for syntactic or grammatical complexity do not allow for a clear conclusion with regard to how sentential and phrasal complexity develop with increasing proficiency. However, the results do give intriguing insights in details of the development of the grammatical system: The results for *to* infinitives and *half modal clusters* show, that while measures of verb complexity may indicate some overall trend based on their presence or absence, there are still nuances in how they relate to proficiency, depending on how they are used, given that they are present in a text. Also, it is highly interesting to observe, that grammatical complexity seems indeed to be highly influenced by functional task-effects on both, the sentential and the phrasal level: While the sum of non terminal nodes per words and the use of conjunctional clauses significantly interact with task theme, the ratio of half modal clusters and the sum of non terminal nodes per sentence are the only measures without significant interactions that do differ severely across models: The ratio of half modal clusters per verb is more significant in the complexity model, than in any of the other models with $\alpha < 0.05$. This indicates that some considerable portion of the variance that is captured by this measure in the complexity model is in fact due to differences in task theme. Also, the ratio of non terminal nodes per sentences is non linear for both, the reference and the complexity model, and only becomes linear when adding the interactions.

Lexical complexity Lexical complexity is assessed in terms of lexical diversity by a standard *type token ratio*, which interacts with task theme: Across themes, the measure shows a significantly increasing slope (cf. Tables A.3, A.4, and A.5 for the respective reference levels), yet, the slopes show varying degrees of steepness. It is strongest for *demand* texts with $\alpha < 0.01$ and levels slightly off for the other levels, although *small talk* texts are the only ones that significantly differ from the others. *Small talk*, which shows the lowest effect, is consistently significant for $\alpha < 0.10$ across more than 80% of iterations of 10 times 10-folds cross-validation. I.e. small talk prompts less lexical diversity in texts than demand tasks. Also, when task theme is not included in the model, type token ratio becomes a non linear measure,

in order to account for the differences between slopes.

To ensure, that the observed effects may be genuinely attributed to lexical diversity, it was tested, whether text length, too, interacts with task theme: Since standard type token ratios are known to be highly sensitive to text length, it might be possible, that the observed effect is an artifact of an interaction between task theme and text length.[7] Another GAM was build to test, whether the number of words per text interacts with task theme in a model, that includes only number of words and standard type token ratio as complexity predictors. Since this proved to be the case, this model was augmented with another interaction for the standard type token ratio, to see, whether both interactions capture different aspects of the variance. This is in fact the case: although the effect becomes weaker, allowing both measures to interact leads to a significantly improved model fit. This shows, that while the interaction of the standard type token ratio and task theme in the interaction model may capture some variance, which is due to differences in text length, some of the variance is genuinely captured by differences in the lexical diversity of the texts.[8]

Language use Two measures in the model assess language use: The *frequency of annotated lexical types, that were grouped into the second frequency bin of the dlexDB data base* show a significant leveling off of their effect on proficiency when being allowed to interact with task theme: While *demand* and *society* tasks have highly significant negative slopes for $\alpha < 0.01$, neither the slopes for *small talk* nor *profession* texts significantly differ from the intercept. However, only for *small talk* is the difference to the slopes of *demand* and *society* significant. Without further information these results do not lend themselves to a straight forward interpretation. Since the results are specific to the second frequency band and not to other *dlexDB* frequency measures, it is likely, that the type of vocabulary, which is included in the frequency bin is inappropriate for *demand* and *society* texts. An inspection of this frequency band showed, that most words included in it are nouns, adjectives, and finite main verbs, however, within these parts of speech, no common pattern could be identified. Also, it is not quite clear, which word fields would be unsuited for these two texts types, especially since *demand* includes only informal settings across test levels, while *society* texts are all very likely to contain higher register vocabulary, because

[7] Note that this was not tested, when iteratively augmenting the model, because it was not among the selected measures.

[8] The GAM with both interactions and a model comparison may be found in Tables A.9 and A.10 in Appendix A.

they are confounded with test level C1 and formal writing occasions. Section 13.3.3 discusses these interactions further, considering performance effects, too.

The *ratio of lexical types found in dlexDB to lexical types* shows a significant increase for increasing proficiency at $\alpha < 0.01$. This might be due to the type of language represented in *dlexDB*: The data base is based predominantly on written language and the findings might indicate, that more proficient learners use more of the vocabulary that is commonly used in written German. However, keeping the observations from Section 10.1 in mind, the more likely explanation is, that more proficient learners make less writing errors, i.e. that this measure of complexity is influenced by accuracy: all complexity measures were calculated on the non-normalized student writings, so it is to be expected, that for less proficient writers more misspelled words occur per text, which is bound to decrease the number of words found in a frequency data base. An interaction with task theme, which would make a genuine effect for language use more likely, cannot be found for this measure, because the measure neither interacts with task theme, nor does it drastically change in terms of significance or its slope across models. In fact, this lack of task theme influence makes it likely, that in this context, the amount of lexical words found in *dlexDB* is rather indicative for increasing accuracy than for an approximation of language use in written registers, since this would be expected to differ across task themes, both on theoretical grounds as well as given the results for the other frequency measure.

Human language processing The *average total integration cost at the finite verb using the enhanced verb weight modification*, which measures human language processing, shows significant increases with increasing proficiency with $\alpha < 0.01$. The measure is not influenced by task theme, neither in terms of interactions nor in terms of differences across models. This is consistent with the findings from Section 10.3. From a theoretical point of view, both, the increase with increasing proficiency level as well as the lack of functional task-effects are plausible: On the one hand, linguistic structures, that facilitate high integration costs, are themselves associated with higher proficiency: Examples are complex clausal structures with subordination in the middle field, especially when nesting subordinate clauses inside subordinate clauses, or – although less so – high amounts of phrasal modification in the middle field. On the other hand, from a theoretical point of view functional task factors should not effect cognitive processing load as much, i.e. while an interaction of a cognitive task factor could be expected to interact with a measure of cognitive

complexity, this is not necessarily the case for a functional task factor like task theme.

Discourse and cohesion The models contain two measures of discourse and cohesion: the *use of transitions of grammatical roles from subject to no grammatical role* measures whether sentences introduce subjects, without assigning any grammatical role to them in subsequent sentences. The observed negative estimate for this measure with $\alpha < 0.05$ seems, therefore, very reasonable, since such a subject drop may be taken as a sign of lacking cohesion and should occur less often with increasing proficiency. This is also consistent with the findings in Section 10.2, which show, that learners write more cohesive especially in terms of overlaps of linguistic materials, which is bound to assign some semantic role to the re-occurring material. The other measure is the *use of third person possessive pronouns*, which, too, indicates lower proficiency when present. This, again, is in line with the observations already made in Section 10.2. Overall, these findings are taken to confirm the earlier hypothesis from the descriptive study, that learners seem to restructure the coherence of their writing by repetition of aforementioned material, rather than establishing pronoun co-reference.

13.2.4. Classification Experiment

As a final evaluation step, the predictive power of the interaction model was investigated in 10 iterations of 10-folds cross-validation. Classification performance was assessed with averaged weighted F1 scores, precision, and recall. The model was compared to the complexity model and the reference model. In a second step, classification errors were investigated to determine, whether any clear structures are visible in the data, which would allow to gain insights in the shortcomings of the models.

Table 13.3 shows the mean classification results across models for 10 iterations of 10-folds cross-validation and their standard deviations. It also includes a majority baseline, that always predicts A2 as CEFR level. All three models clearly outperform this baseline with F1 scores around $71 - 72\%$. Each model also exhibits precision and recall values which are close to each other, indicating that the models are balanced in that regard, although they show a tendency towards higher precision. Within measures, there are no significant differences across models to be found, when performing two-sided t-tests, which was conducted after confirming, that all measures had equal variance by administering an F test. The difference between the

Model	μ F1	±SD	μ Recall	±SD	μ Precision	±SD
Majority Baseline	7.37	11.59	7.44	11.33	7.37	11.37
Complexity	70.97	4.25	71.63	4.74	72.30	4.09
Reference	71.32	4.33	71.78	4.87	72.74	4.10
Interaction	72.17	4.43	72.69	4.94	73.39	4.15

Table 13.3.: Weighted average precision, recall, and f1 score for complexity, reference, and interaction model for 10 iterations of 10-folds cross-validation.

Predicted↓ / Observed→	A1	A2	B1	B2	C
A1	22.9	13.3	0.0	0.0	0.0
A2	**32.1**	**238.8**	52.6	0.0	0.0
B1	0.0	51.9	**223.5**	36.3	0.0
B2	0.0	0.0	52.0	**246.7**	**37.5**
C	0.0	0.0	0.0	8.0	8.5

Table 13.4.: Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the complexity model.

F1 score of the complexity and the interaction model is marginally significant for $p = 0.05151 (t = 1.959; df = 198)$, though. The lack of more convincing differences between the models is facilitated by the models' considerable standard deviations, which are all around 4-5%. This indicates that the models are prone to overfit to a certain extend. Yet, even with the broad standard deviations, all three models show highly successful classification performance given the unfavorably unbalanced class distribution. Also, when comparing the results to other classification experiments, that were reported for this data set, the model performs at the same level: Hancke 2013: 55 reports F1 scores up to 72.4% for the same classification task when using over 100 complexity measures in a support vector machine based classification approach. The GAMs require considerably fewer predictors, though, and are more interpretable in terms of how individual measures contribute to the prediction.[9]

In order to further investigate the nature of the errors made in the classification experiment, the models' averaged confusion matrices were investigated, cf. Tables 13.4, 13.5, and 13.6. As the tables show, none of the miss-classifications deviates more than one level from the observed CEFR scores. Furthermore, the

[9]Note that a proper replication of Hancke 2013's study was beyond the scope of this thesis, but for future work, it seems desirable to set up a replication study in order to compare model performance more thoroughly.

Predicted↓ / Observed→	A1	A2	B1	B2	C
A1	23.8	12.7	0.0	0.0	0.0
A2	**31.2**	**237.3**	38.8	0.0	0.0
B1	0.0	54.0	**230.1**	36.2	0.0
B2	0.0	0.0	49.1	**247.0**	**39.4**
C	0.0	0.0	0.0	7.8	6.6

Table 13.5.: Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the reference model.

Predicted↓ / Observed→	A1	A2	B1	B2	C
A1	25.5	10.1	0.0	0.0	0.0
A2	**29.5**	**241.3**	45.4	0.0	0.0
B1	0.0	52.6	**233.5**	37.8	0.0
B2	0.0	0.0	49.1	**243.1**	**37.9**
C	0.0	0.0	0.0	10.1	8.1

Table 13.6.: Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the interaction model.

three most common CEFR levels (A2, B1, and B2) are predicted correctly most of the time by all three models: The most frequently predicted CEFR score for each given CEFR score is marked with bold font in each column, to make this trend clearer. All three models also share a tendency to make prediction errors towards the less extreme values, i.e. lower levels are more often over-rated by one level, higher levels are more often under-rated by one level. This corresponds to the results for test levels, that were observed in Section 5.2, where learners taking lower level tests tended to outperform their tests, while learners on higher levels tended to underperform. For the two levels at the margins, A1 and C, this tendency is an issue, though: Each model systematically favors the adjacent proficiency level instead the correct one for both models, i.e. A1 levels are predicted to be A2 levels 53.6% of the time and C levels are predicted to be B2 level in 82.4% of the cases. There are two potential explanations for this behavior: This might be an artifact of the idiosyncratic properties of the distribution of the response variable, since A1 and C are the two levels with the least data instances in the model, cf. Section 5.3. However, it might also be the case, that A1 is missclassified, because for the lowest proficiency level the automatic complexity analysis might be too impaired by erroneous NLP annotations to capture a beginning level learner profile

in sufficient detail. For C level learners, another plausible explanation would be, that high-intermediate and advanced learners do not differ on the levels that are assessed by the model, but mostly in terms of phraseological, phrasal, and stylistic differences, which are not captured in either model (Biber, Gray & Poonpon 2011; Ortega 2012; Paquot 2017). Since this was found mostly for academic writings, though, and C levels are also very under-represented on this data, further research is required to come to a conclusion. In particular, it seems advisable to i) repeat the study on *Merlin*'s target hypothesis; ii) to build a balanced data set; and iii) to investigate phrasal, phraseological, and stylistic development of high-intermediate to advanced learners in more detail.

13.3. Ancilliary Study 2: Modeling Performance-Effects

13.3.1. Model Description

The second study investigates, how performance in terms of succeeding or failing in a test effects the analyses: As mentioned at the beginning of this chapter, this ancillary study was prompted by the analysis of residual errors in Study 1, which indicated that the model missed some information related to divergences between test levels and proficiency scores. It was also motivated by results from previous studies and theoretical considerations. For this, a binary performance predictor was added to the reference model, cf. Section 5.1. Interactions between performance and complexity measures were introduced following the procedure for the introduction of interactions outlined before. This lead to two significant interactions for success with the two smoothed measures *characters per word* and *squared number of sentences*.

However, when re-introducing interactions for task theme, only the interaction between task theme and *log of occurrences of annotated types in frequency band 2 per annotated types found in dlexDB* significantly improved the model. Note that except for *use of conjunctional clauses*, all other interactions kept significantly diverging slopes, but did not improve model fit significantly. Also, the *ratio of half modal clusters per verb phrase* developed an interaction with task theme, that did not contribute significantly to model fit, but lead the measure to become insignificant without the interaction. Please see Section 13.3.3 for a more detailed reasoning on how to interpret these changes. For the purposes of this ancillary study on performance effects, it was decided to be conservative and to proceed without measures and interactions, that did not significantly improve model fit, when discussing model

```
506 gam.merlin.success <- gam(OverallCefrScore ~
507                            hasTransitionsFromSubjectToNot +
508                            has3rdPersPossessivePronouns +
509                            containsToInfinitives +
510                            usesConjunctionalClauses +
511                            logSumNonTerminalNodesPerSentence +
512                            logATFBand2PerTypesFoundInDlex +
513                            avgVTotalIntegrationCostAtFiniteVerb +
514                            lexTypesFoundInDlexPerLexType +
515                            typeTokenRato +
516                            logSumNonTerminalNodesPerWord +
517                            logATFBand2PerTypesFoundInDlex:TaskTheme +
518                            s(charactersPerWord, by = Passed, k = 6) +
519                            s(numberOfSentencesSquared, by = Passed) +
520                            Passed +
521                            TaskTheme,
522                          data=merlin,
523                          family=ocat(R=n_cat))
```

Figure 13.5.: Model formula of *Merlin* success model predicting overall CEFR scores from scaled and transformed complexity measures

fit and performing the classification experiment. This lead to the *conventional success* model, whose model formula is shown in Figure 13.5. However, for model discussion, it seemed more interesting to focus on how the interactions between task theme and complexity measures changed when adding performance as a predictor and to reason about what the changes mean for the interpretation of the task theme interactions in Study 1. Thus, the model discussion primarily focuses on the extended success model, whose model formula may be found in Figure A.6 in Appendix A.

13.3.2. Model Fit

The *conservative success* model explains $R^2 = 90.0\%$ of the deviance in the *Merlin* data after removing the four most severe outliers (see below), while the *extended success* model explains $R^2 = 90.6\%$. This is considerably more, than the results obtained for the previous models. Also, the conservative success model consumes fewer edf than the interaction model, because it includes fewer interactions and predictors. This may bee seen in Table 13.7, which compares the success model to the reference, the complexity, and the interaction model. The success model obtains considerably lower REML scores than the complexity and reference model by investigating 7 respectively 6 edf. The interaction model investigates five more edf and still has a considerably lower REML score.

The residual errors are comparable to the residuals of the previous models, see

Model	AIC	Df	REML	Edf	Comparison with	χ^2	Edf diff.	$Pr(>\chi^2)$
Complexity	1315.05	30.37	658.56	19				
Reference model	1287.08	28.41	642.77	20				
Interaction model	1281.00	39.27	628.84	31				
Success model	821.11	35.76	401.19	26	Complexity	257.36	7	$< 2e-16$
Success model					Reference	241.573	6	$< 2e-16$
Success model					Interaction	227.65	-5	

Table 13.7.: Model comparison for reference, complexity, interaction, and success GAMs modeling L2 proficiency from complexity measures and task theme on the *Merlin* data.

Figure 13.6. However, there are overall five severe outliers, that skew the mean residual error to $\mu = 2.72; sd = 80.00$. When removing the four most severe outliers, the mean residual error get close to zero, though, and the standard deviation considerably decreases to $\mu = -0.14; \sigma^2 = 12.91$. As for the other models, removing more outliers did not effect model fit sufficiently, hence, only these four outliers were removed.

Interestingly, these outliers are again identical to those from the previous models, except for the originally most severe outlier, which is now not among the ten most severe outliers anymore: The performance predictor allows the model to account for learners who failed their tests. However, with regard to the outliers were learners outperformed their test level, no changes may be found, which is plausible, since the measure does not differentiate whether learners scored at or above their test level.

13.3.3. Model Discussion

Table A.11 shows the model summary for the extended success model. Although the more conservative model is considered to be more adequate and thus reported in terms of model fit and classification results, most interesting in the context of this thesis are the changes in the model with respect to the interactions with task-effect after adding performance as a predictor: In fact, the major question posed by these results is to establish, what these changes mean for the validity of the findings in Study 1, since the lack of significantly informative interactions with task theme after introducing performance as a predictor could indicate, that presumed task-effects might in fact be performance effects. Hence, it seems most adequate to discuss the extended success model in more detail and to refer to the respective estimates in the conservative success model while discussing the extended success model. The

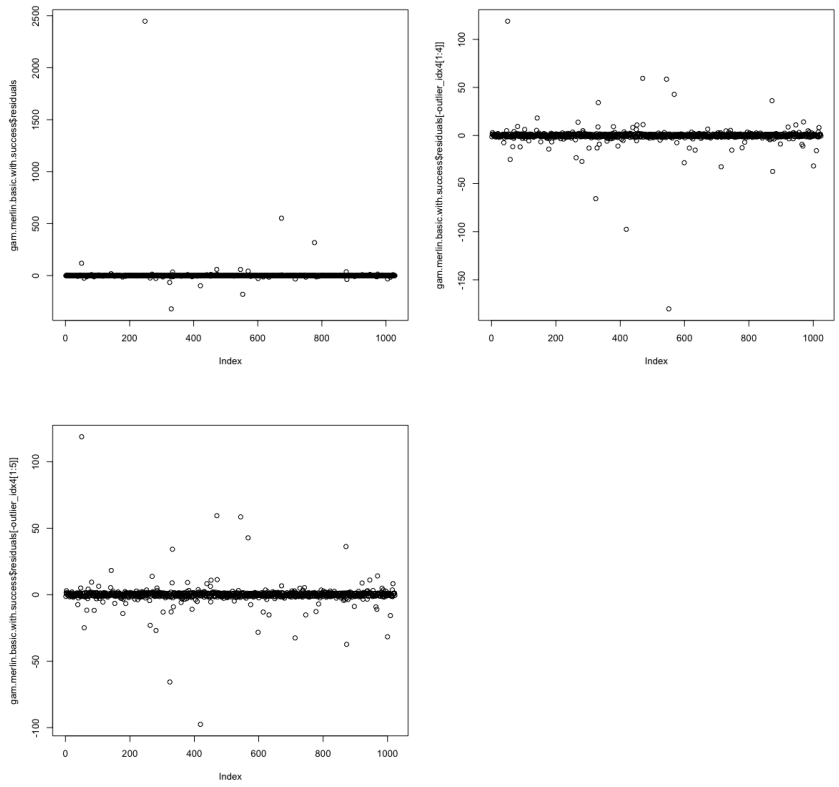


Figure 13.6.: Residuals of *Merlin* success model on i) full data (panel 1); ii) data without 4 most severe outliers (panel 2); iii) data without any outliers (panel 3).

table for the conservative success model may be found in Table A.11 in Appendix A.

Non Interacting Measures For both, the extended and the conservative success model, predictors that showed no interaction with task theme in the interaction model do not change except for minor differences in their estimates and mildly strengthened significance: To be more precise, the *average total integration cost at the finite verb using the enhanced verb weight modification*, *ratio of lexical types found in dlexDB to lexical types*, the *use of transitions of grammatical roles from subject to no grammatical role*, *use of third person possessive pronouns*, the *number of non terminal nodes per sentence*, and the *use of to infinitives* are overall stable and if they exhibit differences to the interaction model, these are due to a strengthening of the respective effects.

Task Theme Interactions As mentioned above when describing the two success models, introducing performance as a predictor does effect measures that interact with task theme: The interaction between task theme and *log of occurrences of annotated types in frequency band 2 per annotated types found in dlexDB* is the only measure, that remains significant in terms of both, differences between by task theme slopes and improved model fit. The slopes for *demand* texts become marginally steeper, but are overall stable and neither *society* nor *small talk* texts significantly differ from *demand*. This is remarkable for *small talk* texts, because this was the only theme for which the slope significantly differed from *demand*. However, *profession* texts become considerably strengthened in terms of their positive slope and significantly differ from *demand* texts with $\alpha < 0.05$. This *profession* slope is significantly different from the intercept with $\alpha < 0.10$ for more than 80% of 10 iterations of 10-folds cross-validation. Although this effect is weak, it is particularly interesting. It shows, that information on task performance provides the model with sufficient information to identify differences between task themes with respect to this measure, that go beyond the distribution of test performance across task themes. The latter has to be the case, because adding the interaction would otherwise not lead to a significantly better model fit, since task theme and performance are already included as separate predictors in the model.

As for the interactions between task theme and *non terminal nodes per word* and task theme and standard *type token ratio* persist insofar as there are significant differences between the by task theme slopes, but including them to the model does not lead to a significantly better model fit, i.e. they do not explain enough

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	4.4304	0.6788	6.5267	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.8561	0.2884	-2.9687	0.0030
has3rdPersPossessivePronouns[TRUE]	-1.2599	0.2666	-4.7251	< 0.0001
containsToInfinitives[TRUE]	-0.7498	0.3130	-2.3957	0.0166
usesConjunctionalClauses[TRUE]	-0.4352	0.3592	-1.2116	0.2257
logATFBand2PerTypesFoundInDlex	-0.4607	0.1269	-3.6291	0.0003
avgVTotalIntegrationCostAtFiniteVerb	0.5146	0.1437	3.5814	0.0003
lexTypesFoundInDlexPerLexType	1.0218	0.1165	8.7722	< 0.0001
typeTokenRato	1.6093	0.2365	6.8037	< 0.0001
sumNonTerminalNodesPerWord	-0.8154	0.1831	-4.4530	< 0.0001
logSumNonTerminalNodesPerSentence	2.5348	0.2241	11.3088	< 0.0001
halfModalClusterPerVP	1.5797	0.7079	2.2316	0.0256
Passed[TRUE]	7.0886	0.3383	20.9528	< 0.0001
TaskTheme[Society]	11.0696	1.1406	9.7055	< 0.0001
TaskTheme[Profession]	7.3308	0.8728	8.3989	< 0.0001
TaskTheme[Smalltalk]	0.9417	0.5599	1.6820	0.0926
logATFBand2PerTypesFoundInDlex:TaskTheme[Society]	0.4645	0.6222	0.7465	0.4553
logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]	1.4155	0.5311	2.6651	0.0077
logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]	0.3894	0.2583	1.5075	0.1317
usesConjunctionalClauses:TaskTheme[Society]	1.6233	1.2559	1.2925	0.1962
usesConjunctionalClauses:TaskTheme[Profession]	-0.1916	0.9070	-0.2113	0.8327
usesConjunctionalClauses:TaskTheme[Smalltalk]	-0.4305	0.6182	-0.6964	0.4862
sumNonTerminalNodesPerWord:TaskTheme[Society]	-0.1837	0.6428	-0.2858	0.7750
sumNonTerminalNodesPerWord:TaskTheme[Profession]	0.7729	0.5283	1.4630	0.1435
sumNonTerminalNodesPerWord:TaskTheme[Smalltalk]	-0.2753	0.2709	-1.0164	0.3094
typeTokenRato:TaskTheme[Society]	-0.7026	0.5412	-1.2982	0.1942
typeTokenRato:TaskTheme[Profession]	-1.1525	0.6222	-1.8525	0.0640
typeTokenRato:TaskTheme[Smalltalk]	-0.6862	0.3602	-1.9050	0.0568
halfModalClusterPerVP:TaskTheme[Society]	-1.5613	0.7292	-2.1412	0.0323
halfModalClusterPerVP:TaskTheme[Profession]	-1.8006	0.7761	-2.3199	0.0203
halfModalClusterPerVP:TaskTheme[Smalltalk]	-1.7694	0.7563	-2.3395	0.0193
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord):Passed[FALSE]	2.6387	3.2708	6.6442	0.1152
s(charactersPerWord):Passed[TRUE]	1.9877	2.5346	8.3244	0.0267
s(numberOfSentencesSquared):Passed[FALSE]	3.9049	4.8709	66.3053	< 0.0001
s(numberOfSentencesSquared):Passed[TRUE]	4.3656	5.3919	258.7651	< 0.0001

Table 13.8.: Summary of extended success model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses 'demand' as reference level.

additional variance to justify the investment of edf that is required to estimate the different slopes. In fact, all slopes for the standard *type token ratio* interaction remain virtually the same, except for the slope for *profession* texts, which starts to differ marginally significant from *demand* with $\alpha < 0.10$ for more than 80% for 10 iterations of 10-folds cross-validation when introducing performance as a predictor. Note, though, that the resulting slope is not significantly different from the intercept anymore. The *ratio of non terminal nodes per word* remains unchanged in terms of the slopes for *demand* texts, but the previously significant difference for *society* texts disappears. Given the fact, the *demand* texts are predominantly passed by learners (97.36%) and society texts are predominantly failed by learners (80.30%), it seems that the differences in slopes in the interaction model are actually artifacts of the skewed performance distributions. For *small talk*, which shows also significant differences to the slope for *demand* in the interaction model, the interactions stays stable at $\alpha < 0.10$ for more than 80% of 10 iterations of 10-folds cross-validation, but the slope's polarity changes from negative to positive. This could be due to performance capturing the skewed distribution of whether learners failed or passed, which could allow the ratio of non terminal nodes to capture differences beyond this distinction. However, more investigations would be needed to find more evidence for this hypothesis. Interestingly, again, the originally not different slope for *profession* texts becomes significantly different at $\alpha < 0.10$ for more than 80% of 10 iterations of 10-folds cross-validation, because the coefficient changes polarity and increases considerably. When ignoring these differences in the conservative success models, the slopes follow the general trends of the respective coeficients, i.e. *non terminal nodes per word* shows a highly significant negative slope, and *type token ratio* a highly significant positive slope.

The *use of conjunctional clauses* is the only interaction from the interaction model, that does not show any significant differences between slopes after including task theme as a predictor, although the individual estimates hardly change except for *smalltalk*, which changes polarity and *society*, which shows a slightly decreased estimate. However, this change sufficed to make the difference between *society* and the other slopes insignificant. Again, this seems to indicate, that significant differences between *society* texts and other texts may be explained by performance differences instead. In the conservative success model, presence of conjunctions is correlated with lower proficiency for $\alpha < 0.05$.

The *ratio of half modal clusters per verb phrase* is the only measure that did not interact with task theme in the interaction model, but is effected by introducing

performance as a predictor: The measure only shows a significant slope, if it is allowed to interact with task theme in the extended success model. This shows significant differences between *demand* texts and all other task themes at $\alpha < 0.05$: *demand* texts show a strong positive correlation with task theme, but not for *society*, *profession*, or *small talk* texts, which are all insignificant from zero. Since *profession* and *small talk* texts are also skewed in favor of passed tests, it seems unlikely that this interaction is not genuine, but may only surface when accounting for performance differences. However, it should be noted that despite these significant differences, the interaction does not contribute enough information beyond what is already included by the performance predictor, to lead to a significantly improved model fit. It seems desirable, though, to investigate potential functional task-effects on half modal verb clusters or verb clusters in general on a data set that shows a more favorable performance distribution.

Overall, these changes show, that while task theme clearly improves model fit in Study 1 and shows significant interactions with complexity measures, idiosyncratic distributional properties of the data seem to interfere with the analyses. Hence, it is not quite clear to which extend the findings are theoretically interpretable, before conducting further research on more controlled data. Unfortunately, this went beyond the scope of this thesis, but has to be addressed in future work. Still, the persistence of significant differences in slopes are taken to indicate, that tasks do influence complexity in *Merlin*, only the more concrete nature of these influences cannot be narrowed down with certainty until further investigations.

Performance Interactions The two non-linear superficial length measures, *number of characters per word* and *squared number of sentences* show significant interactions with performance. This is shown in Figure 13.7, which shows each smooth once for learners who failed their tests (left panels) and once for learners who passed their tests (right panels). The smooth for *characters per word* given a failed test is insignificant and thus not interpretable. For learners, who pass their tests, though, characters per words increases with proficiency level. Although this effect is nearly linear, it still shows a significant increase only after passing the threshold of zero. The *squared number of sentences* is similar for success and failure, insofar as the number of sentences increases for increasing proficiency levels, but levels off for higher proficiency levels. However, for learners who failed the onset of leveling of is earlier and the leveling off effect in itself is more extreme. This interaction with the differences between global performance scores and test level and these

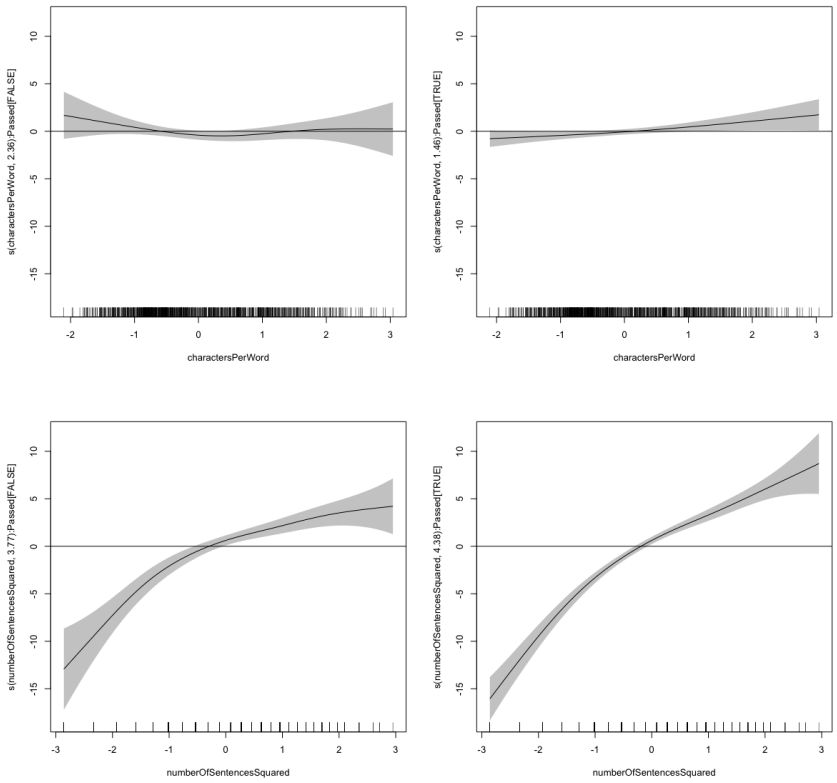


Figure 13.7.: Smooths of *Merlin* success model.

global length measures is also seems theoretically straight forward, because these two measures are sensitive to a variety of language changes, which would be expected to be caused by over-strained or under-challenged writing. Hence, they may capture overall differences in the language system that are not induced by a single characteristic but by multifarious changes. Furthermore, it stands to reason, that high-intermediate learners, who are over-strained by a given task, might exhibit superficially the same results as advanced learners, who perform at or above test level, because they attempt but fail to show adequate behavior. Put differently, learners who fail a C test might show comparably high numbers of sentences as learners who pass a C test, but either the conveyed information or the produced linguistic structures are insufficient to pass the test. Since the global measure is insensitive to such structural, stylistic, or content differences, the performance predictor allows the model to give less credit to the measure, for failed tests. To investigate this hypothesis further, it was tested, whether learners, who fail a C1 test, show wider standard deviations for the number of sentences, than learners, who pass a C1 test. The intuition behind this was, that broader standard deviations would indicate a less systematic realization of the measure, which fits into the profile of a more erratic strategy of an overstrained learner. In fact, *squared number of sentences* shows a higher standard deviation for learners who failed a C1 test ($\mu = 3.92; sd = 0.57$), than for learners, who passed it ($\mu = 3.81; sd = 0.48$).

13.3.4. Classification Experiment

As the previous models, the *success* model was also evaluated in terms of its predictive power. For this, it was trained and tested in 10 iterations of 10-folds cross-validation and evaluated by means of weighted average precision, recall, and F1 score. The results are shown in Table 13.9, which is a repetition of Table 13.3 from the previous section, but augmented by the results for the *success* model in the last row of the table. The conservative success model significantly outperforms all previous models on all measures: It not only shows scores that improved on average by about 10%, but also decreased the result's standard deviation considerably. When inspecting the results further in the averaged confusion matrix across trials, cf. Table 13.10, it becomes apparent, that this improvement is due to two major differences:

First, when looking at levels A2 to B2, the tendency to overestimate proficiency levels is clearly reduced. Second and more prominently, performance at the edges of the classification spectrum improved considerably, especially for texts with C

Model	μ F1	±SD	μ Recall	±SD	μ Precision	±SD
Majority Baseline	7.37	11.59	7.44	11.33	7.37	11.37
Complexity	71.20	4.25	71.89	4.71	72.53	4.03
Reference	71.32	4.33	71.78	4.87	72.74	4.10
Interaction	72.17	4.43	72.69	4.94	73.39	4.15
Success	84.98	2.75	85.60	2.80	85.28	2.74

Table 13.9.: Weighted average precision, recall, and F1 score for complexity, reference, interaction and success model for 10 iterations of 10-folds cross-validation.

Pred.↓ / Obs.→	A1	A2	B1	B2	C
A1	**28.1**	12.7	0.0	0.0	0.0
A2	26.9	**260.9**	34.5	0.0	0.0
B1	0.0	30.4	**271.9**	18.2	0.0
B2	0.0	0.0	21.6	**271.9**	6.0
C	0.0	0.0	0.0	0.4	**40.0**

Table 13.10.: Averaged confusion matrix for classification of L2 proficiency in *Merlin* using the success model.

scores. This increased discriminatory power between B2 and C ratings due to the performance predictor is unexpected at first. When inspecting the distribution of passed and failed texts, that received B2 and C scores, it should be noted, that ambiguous candidates, i.e. learners who might receive either a B2 or C rating by the model, can be uniquely identified as B2 rated learners if they failed their tests:[10] it is not possible to receive a C score when failing a test, because there is no higher test level than C1. However, most B2-rated learners passed their tests, too, so while this criterion may facilitate the classification, it does not seem reasonable to assume that this is the only reason for the model to distinguish the two adjacent levels so well. Another explanation might be the observed interaction between performance and the global measures, which showed different slopes in particular with respect to more proficient learners. This comes back to the question, whether the initial struggle of the interaction, reference, and complexity models is mostly due to the under-representation of instances for C-rated texts or to the insensitivity of the model to the linguistic differences between high-intermediate and advanced

[10]Since no C learners were mistaken for any other proficiency level than B2, scores B2 and C were the only ones considered as ambiguous candidates in this scenario.

learners, which have been argued to be mostly stylistic and at the phrasal level. Hence, no final conclusion can be drawn until further investigation in this issue as suggested in Study 1.

For A1 scores, the improvement is less extreme, but still considerable. The weaker impact might be due to the fact that for ambiguous candidates – i.e. learners, who might be rated as A1 or A2 learners by the model – the information whether they failed or passed their test alone is not very informative, because for both levels, the percentage of failed and passed is virtually the same: 78% of learners with an A2 score passed their tests, for learners with A1 scores the success rate is 77%.[11] Thus, any improvement seems to be due to the influence of performance on the complexity measures. Since the two interactions with performance show differences in the slopes mainly at higher proficiency levels, the only changes in the model, that might cause these effects, may be the changed task theme interactions. At the levels in question, differences between *demand* and *small talk* texts are particularly relevant, since the other task themes are not or under-represented at these proficiency scores. Interestingly, with respect to small talk, the coefficients changed their polarity when it came to the *use of conjunctional clauses* and the *sum of non terminal nodes per word* in the extended success model, which brought it each time close to the overall coefficient for the entire measure that was estimated in the conservative success model. Since in terms of test levels, *small talk* is predominantly represented at test level B1, but also at test level A1, one possible explanation for the slightly improved results with respect to A1 classification and the changes in the task theme interactions could be, that due to the idiosyncratic distributional properties in *Merlin*, small talk texts with low complexity could not properly be differentiate between average A1 and under-performing B1 test takers, leading to a tendency to rate them all as A2 learners. However, when adding the information, that learners pass or fail, this ambiguity is resolved. While this is one potential theory on how to interpret these changes, again, more investigations on more controlled data sets is required for more informed conclusions.

13.4. Summary

The four models built in the course of Studies 1 and 2 gave interesting insights in L2 performance and task-effects. First, with regard to the first two research questions,

[11] Since no A1 learners were mistaken for any other proficiency level than A2, scores A2 and A1 were the only ones considered as ambiguous candidates in this scenario.

1a. Which complexity measures are elicited for model design in an exploratory, data-driven approach?

1b. Which insights do these measures give into L2 proficiency?

it could be seen, that the exploratory model design yielded in a set of 13 global system and local structure complexity predictors from all four main domains covered by the original collection of 398 measures. These showed a combination of increase and decline with advancing proficiency:

More proficient learners were modeled by more coherent writing in terms of fewer drops of subjects in adjacent sentences. They also use fewer third person possessive pronouns. Lexical complexity is somewhat underrepresented in the model, but it shows higher lexical diversity for more proficient learners as measured by higher type token ratios. As for language use, measures are likely to be heavily influenced by spelling improvements, which may be seen from the increase of words found in *dlexDB*. At the same time, the use of words for log frequency band 2 decreases, which requires further investigations before allowing for interpretations. More proficient learners also produced longer words, more sentences, more non terminal nodes per sentence, but less non terminal nodes per word, and higher total integration costs at the finite verb using additional verb weight. This may be interpreted as an overall global increase of complexity in the grammatical system, which goes along with an increase in processing cost for discourse referents. On a more local scale, interesting results were obtained for verb complexity in terms of verb clusters: While the models assume, that more proficient learners tend to avoid *to* infinitives, their use in half modal clusters indicates higher proficiency. These findings are highly interesting and the acquisition of German complex verb clusters should be investigated in more detail. Also, conjunctional clauses were found to indicate higher proficiency on texts reasoning about society.

This leads to the answers found for other two research questions,

2a. Are these measures influenced by task theme?

2b. If so, how are measures influenced by task theme?

Overall four interactions with task theme were found in Study 1, namely with the use of conjunctional clauses, non terminal nodes per word, type token ratio, and words found from log frequency band 2 in *dlexDB*. Mostly, these measures simply showed differences in the slopes' steepness, while following the same general trend. However, conjunctional clauses were found to indicate higher proficiency in

society texts, but lower proficiency if present in demand and profession texts. The differences in slopes became insignificant, though, when including performance as a predictor as well as performance interactions with word length and number of sentences in Study 2. Hence, this effect seems to be mostly due to performance aspects on the C1 level tests, which are confounded with task theme *society*.

As for the other interactions, except for words found in log frequency band 2 in *dlexDB*, none of the interactions contributed significantly to model fit after introducing performance as a predictor, although in the other models from Study 1, they lead to a significantly better fitting model. The differences between slopes remains significant, but not for the same task themes as before. Also, the ratio of half modal clusters starts to exhibit significant differences between slopes, that were not to be found in the previous Study. Overall, this is taken to show, that there are some task-effects for task theme, but that the distributional properties of the data with respect to test level, proficiency scores, task theme, and performance are too imbalanced, to allow for clear conclusions. Thus, further investigations are needed on more balanced data sets, especially with regard to how performance and task theme combined influence complexity.

Finally, to answer the third research question,

3a. How well do the most suited complexity measures perform in a classification task?

the classification experiments in both Studies yielded remarkable results. In Study 1, all models reached F1 scores between 71 - 72%, albeit with relatively wide standard deviations of roughly $\pm 4\%$. Still, these results are comparable with previous models reported for this data set, for example by Hancke 2013. Yet, the GAMs require considerably fewer predictors to achieve these results – 13 vs. more than 100 in ibid. – and are more interpretable in terms of the contribution of individual predictors. However, the results also show, that there is no clear improvement in terms of including task theme interactions in terms of classification performance, despite the significantly improved model fit. In particular, all three models suffer from erroneous classifications at the lowest and highest proficiency levels, which were reasoned to be due to the under-representation of these instances in the data set or caused by systematic properties, that the model misses. These issues are overcome by the model for Study 2, which reaches average F1 scores of nearly 85% and decreases the standard deviation of these findings to $\pm 2.75\%$. These results are highly remarkable. However, whether these improvements are due to the genuine

informativeness of the measure or some idiosyncratic structure in *Merlin* remains to be investigated further on a more balanced data set.

14. Modeling L2 Proficiency in Falko Georgetown

This chapter studies L2 proficiency in the *Falko Georgetown L2* corpus. The main study is conducted on the longitudinal data set, because investigating the progressive development of learners of German with respect to the language complexity of their writings is an aspect, that cannot be assessed on the German *Merlin* corpus, which is fully cross-sectional. However, due to the limited size of the longitudinal data set, the generalizability of the results was analyzed on the reference data sets (cf. Section 6.3). The following studies investigate the informativeness of cognitive task factors in terms of Skehan's code complexity. For this, the original research questions formulated in the introduction were again specified as follows:

1a. Which complexity measures are elicited for model design in an exploratory, data-driven approach?

1b. What do these measures show about L2 proficiency?

2a. Are these measures influenced by code complexity?

2b. If so, how are measures influenced by code complexity?

Although the analyses on the *Falko Georgetown L2* data do investigate the same research questions as the analyses on the German *Merlin* data, the studies have to be set up differently, because of the differences between the corpora, some of which were already discussed in Chapter 6. In terms of study design, the three most crucial differences are i) the limited size; ii) the partially repeated measures; and iii) the split of data into a reference task, which remains stable across course levels, and curricular writing tasks, which differ across courses, i.e. are perfectly confounded with course levels. To address the first two aspects, the first study uses complexity measures as predictors of course levels in a GAMM, that is estimated on the longitudinal data, which consists nearly exclusively of curricular writing tasks. Then, the generalizability of the obtained effects is tested across reference data sets, that also include the book review tasks and other data that was excluded

from the longitudinal data, see Section 6.3 for details. Only after confirming the representativeness of the findings for the single complexity measures, interactions with the designated task-effect were introduced in the second study. Here, again, the effects were identified on the longitudinal data and then confirmed by testing them on the full data. Tests on the other comparison data sets were not possible, due to a lack of variability with respect to the task factor predictor in this data. Since the limited data support only allows to build models, that assess specific aspects of complexity in terms of selected measures, but not systematic aspects of L2 proficiency, no classification experiments were conducted. After elaborating on the general set up in Section 14.1, Study 3 in Section 14.2 investigates subject effects on the *Falko Georgetown L2* data without task-effects, before Study 4 in Section 14.3 elaborates on task-effects. The chapter closes with a discussion of the findings in Section 14.4

14.1. Set Up

Since the *Falko Georgetown L2* data does not include any ratings for the learner texts, course level was used as an approximation of L2 proficiency, cf. discussion in Section 6.1. Code complexity was used as cognitive task factor: Together with cognitive complexity, it was the only measure, that showed a sufficient distribution across course levels and that is variable within curricular writings. When comparing these two measures, code complexity shows a slightly more balanced ratio of high and low condition instances and its operationalization lends itself to a more straight forward implementation, cf. Section 7.3. Thus, for the purposes of this theses, it was considered to be the preferable choice.

Complexity measures were extracted using the system described in Chapter 9. From this collection of 398 measures, a pool of potential predictors was established and features were selected from this following the iterative augmentation approach outlined in Section 12.2. For this, the longitudinal data set was used, first, to build a GAMM predicting course level from code complexity and including subject id as a random intercept. This lead to an augmentation of the model with four measures before reaching the limits of the model given the limited number of texts in the longitudinal data. A slightly larger model could have been supported by the full data set, but keeping the focus on the longitudinal development was preferred over covering more measures. Also, the resulting selection of features is small but highly diverse, since it includes predictors of morphological complexity, language use,

```
144 gamm.falko <- gam(courseLevel ~
145                     ungDerivationPerTokenSquared +
146                     logLexTypesNotFoundInKCTPerLexType +
147                     s(thirdPersPossessivePronounsPerToken, k = 4) +
148                     s(wordsPerClause, k = 4) +
149                     s(subjectID, bs="re"),
150                   data=falko.long,
151                   family=ocat(R=n_cat))
```

Figure 14.1.: Model formula of *Falko Georgetown L2* subject model predicting course level from scaled and transformed complexity measures, when trained on the longitudinal data.

discourse, and a superficial clause length measure that may be influenced by global clause complexity or fluency. All four complexity measures are continuous and were centered around zero before conducting any of the studies.

14.2. Study 3.1: Modeling Complexity across Data Sets

14.2.1. Model Description

The first study on the *Falko Georgetown L2* data was designed to investigate subject id as a random effect on the longitudinal data. Hence, code complexity was excluded from the model and varying types of random effects were introduced for subject id. However, since there was no evidence for random slopes or random smooths contributing to model fit, the model was only supported with a random intercept for subject id. This resulted in the *subject* model whose formula for fitting on the longitudinal data is shown in Figure 14.1. The model includes two linear and two non linear measures: The linear measures are *ung* derivations per noun, i.e. a measure of derivational complexity, and the ratio of log frequency of lexical types, that were not found in the KCT corpus to lexical types, which is a measure of language use. The two smoothed measures are the ratio of third person possessive pronouns per token, i.e. a measure of discourse and cohesion, and the ratio of words per clause, i.e. a surface measure that may assess clausal complexity globally, but is influenced by fluency as well. The number of knots per smooth was restricted to four, which did not alter the overall slopes, but restricted the model to a more reasonable amount of consumed edf.

This model was fitted to the longitudinal data and then also applied to the full data set, the inverse data set, and the book review data set, in order to investigate

how stable the results were for unseen data instances. However, for the inverse and the book review data set, subject id was not included as a random effect, because only one subject contributed more than one text to the data in these data sets, which made it impossible for the models to converge when estimating subject id. Also, on the other data sets, measures were investigated for non-linearity anew, which lead to some modifications discussed below.

14.2.2. Model Fit

On the longitudinal data, the model explains $R^2 = 0.8980$ of the deviance, which might be due to overfitting, though, since the model consumes 9 edf due to the wiggliness of the smooths, i.e. each predictor is only supported by 10 data points. However, overfitting is an anticipated risk due to the limited data set size. Thus, the model was refitted on the three reference data sets, where in fact the explained deviance was considerably lower, namely $R^2 = 0.5340$ for the full data set, $R^2 = 0.3460$ for the inverse data set, and $R^2 = 0.3740$ for the book review data set. These results show, that the model is considerably less informative on the other data sets, but at the same time i) these results seem more plausible given the limited number of predictors and ii) although the measures fail to explain a large quantity of the deviance, they are informative across all data sets, which is a good indicator that the selected measures are genuinely informative.

The analyses of the residual errors were overall uniformly distributed with mean values around zero and overall homoscedastic standard deviations across all data sets, see Figure 14.2. Two severe outliers could be identified on the longitudinal data set, but despite skeqing mean and standard deviation from $\mu = -0.13; sd = 1.47$ to $\mu = 4.81; sd = 36.04$, their removal did not effect the model in terms of explained deviance, parameter coefficients, or REML scores. Also, the data set is already so limited that it did not seem justifiable to remove more data. Hence, the outliers were kept in the model. The same holds for the inverse and book review data set, which show mean residual errors of $\mu = 0.01; sd = 5.56$ and $\mu = 0.28; sd = 7.15$. On the full data set, mean residual errors are $\mu = 0.72; sd = 29.95$. Removal of the visible outliers did not effect model fit, hence, they were kept.

14.2.3. Model Discussion

Table 14.1 shows the model summary for the subject model trained on the longitudinal *Falko Georgetown* data set. Figure 14.3 shows the corresponding plots for the

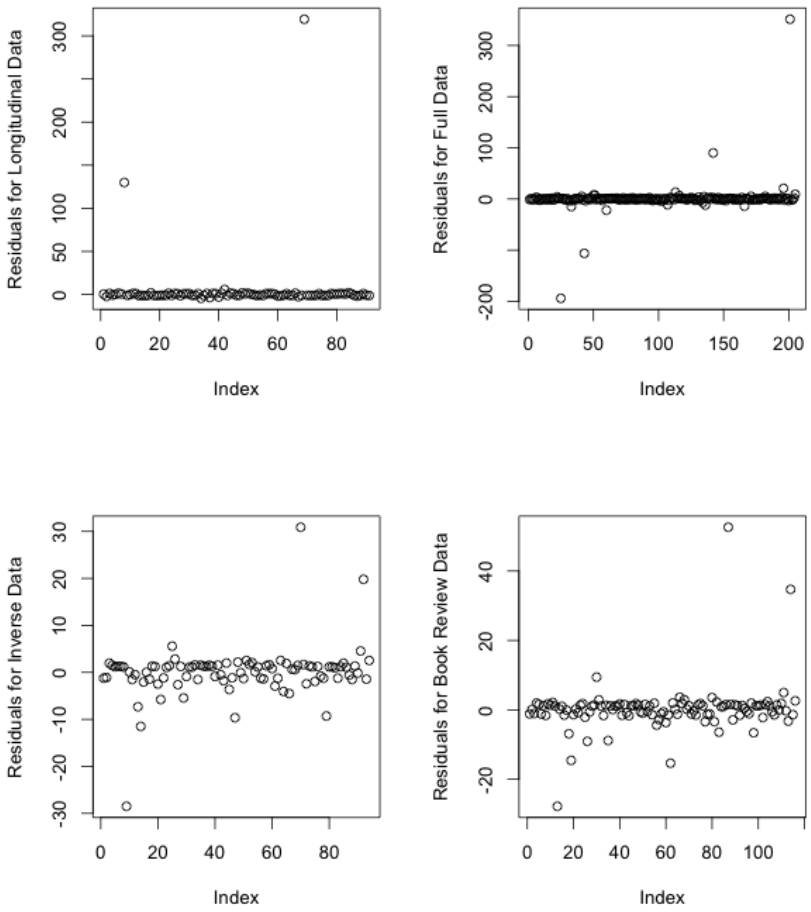


Figure 14.2.: Residuals of the subject model trained on the i) longitudinal (panel 1); ii) full (panel 2); iii) inverse (panel 3); and iv) book review (panel 4) data set.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	5.7696	0.3488	16.5414	< 0.0001
ungDerivationPerTokenSquared	2.2865	0.4815	4.7487	< 0.0001
logLexTypesNotFoundInKCTPerLexType	2.3453	0.5493	4.2698	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(3rdPersPossessivePronounsPerToken)	2.4904	2.8154	28.5833	< 0.0001
s(wordsPerClause)	2.7365	2.9479	32.9625	< 0.0001
s(subjectID)	0.0004	27.0000	0.0003	0.6245

Table 14.1.: Summary of *Falko* GAMM with by-subject random intercepts. Predicts course level from scaled and transformed complexity measures estimated on longitudinal *Falko Georgetown L2* data.

two smoothed predictors. Summaries for the three refitted versions of the model, which were estimated on the full, the inverse, and the book review data sets, may be found in Tables 14.2, 14.3, and 14.4. The corresponding plots of the smoothed measures may be found in Figure 14.8.

Two of the four measures are highly significant at $\alpha < 0.01$ across all data sets: the *squared ratio of ung derivation per token* and the *ratio of words per clause*. The former is a measure of morphological complexity, but since *ung* suffixes are among the most productive nominalization suffixes in German, they are also particularly sensitive to nominal writing style, which is a known feature of academic language (Hennig & Niemann 2013). On the longitudinal data set, the measure shows a relatively steep, positive, linear slope. On the other data sets, the slope levels off for higher proficiency, especially on the inverse and the book review data set (Figure 14.4b). Since the full data set is a combination of the inverse and the longitudinal data, the slope for *ung* derivations on this data set is a straight forward combination of these two slopes: It is close to linear, but showing some leveling off for higher course levels (Figure 14.4a). This difference in slopes indicates, that the curricular tasks in the higher course levels, namely writing articles and speeches, facilitate the use of *ung* derivations compared to the tasks in the lower course levels, namely writing private letters and novels. This is plausible, since speeches and articles are more likely to elicit academic language than novels and private letters. More importantly, though, the effect remains highly significant on the inverse and book review data despite its leveling off. This indicates, that independent from task factors the use of more *ung* derivations indicates more advanced writing.

The *ratio of words per clause*, which assesses clause length on a global level, is overall significantly increasing with increasingly advanced courses throughout data

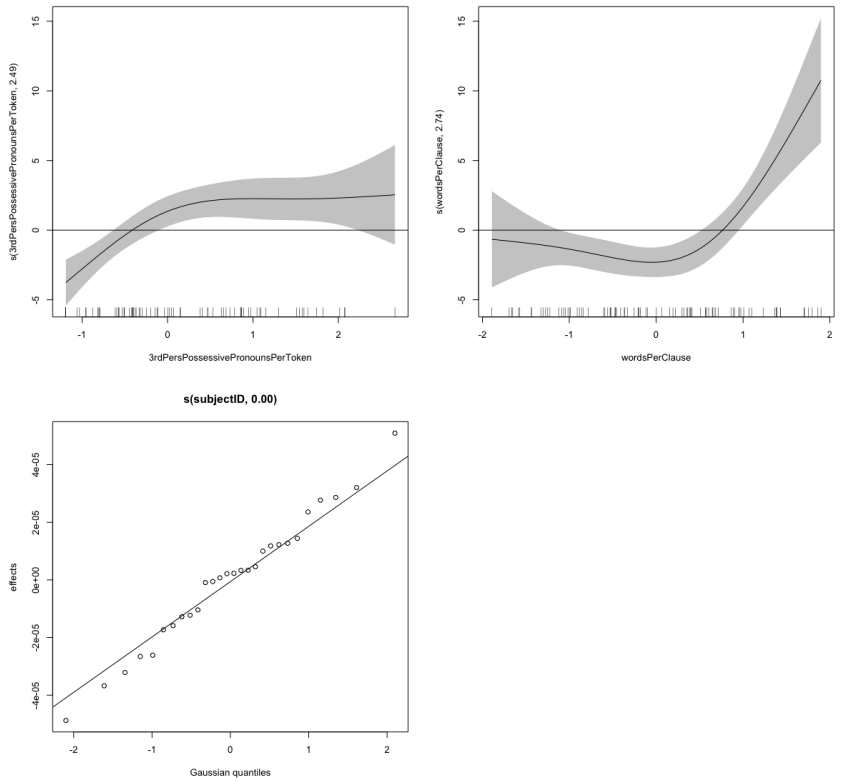


Figure 14.3.: Smooths of GAMM with by-subject random intercepts fitted on longitudinal *Falko Georgetown L2* data.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	2.3484	0.1659	14.1571	< 0.0001
3rdPersPossessivePronounsPerToken	0.7450	0.1712	4.3506	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(ungDerivationPerTokenSquared)	2.1186	2.6133	41.7028	< 0.0001
s(logLexTypesNotFoundInKCTPerLexType)	2.1566	2.7468	15.4155	0.0013
s(wordsPerClause)	2.8997	3.6589	50.6344	< 0.0001
s(subjectID)	0.0017	119.0000	0.0017	0.4643

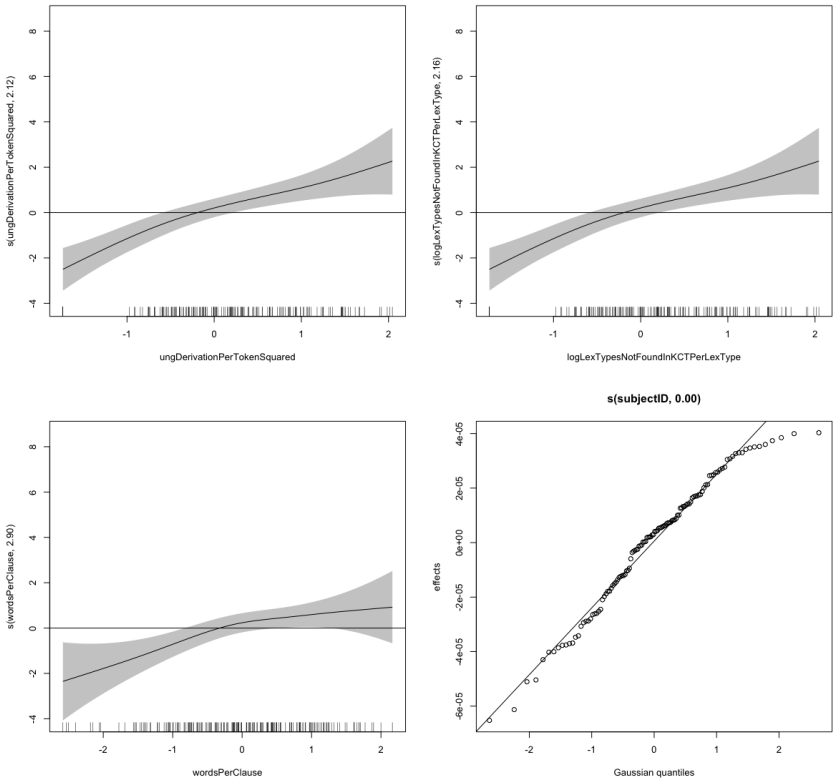
Table 14.2.: Summary of *Falko* GAMM with by-subject random intercepts. Predicts course level from scaled and transformed complexity measures estimated on full *Falko Georgetown L2* data.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	1.5108	0.2188	6.9064	< 0.0001
logLexTypesNotFoundInKCTPerLexType	0.3911	0.2328	1.6800	0.0930
3rdPersPossessivePronounsPerToken	0.2882	0.2236	1.2886	0.1975
wordsPerClause	0.7240	0.2738	2.6439	0.0082
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(ungDerivationPerTokenSquared)	2.0017	2.4543	28.1487	< 0.0001

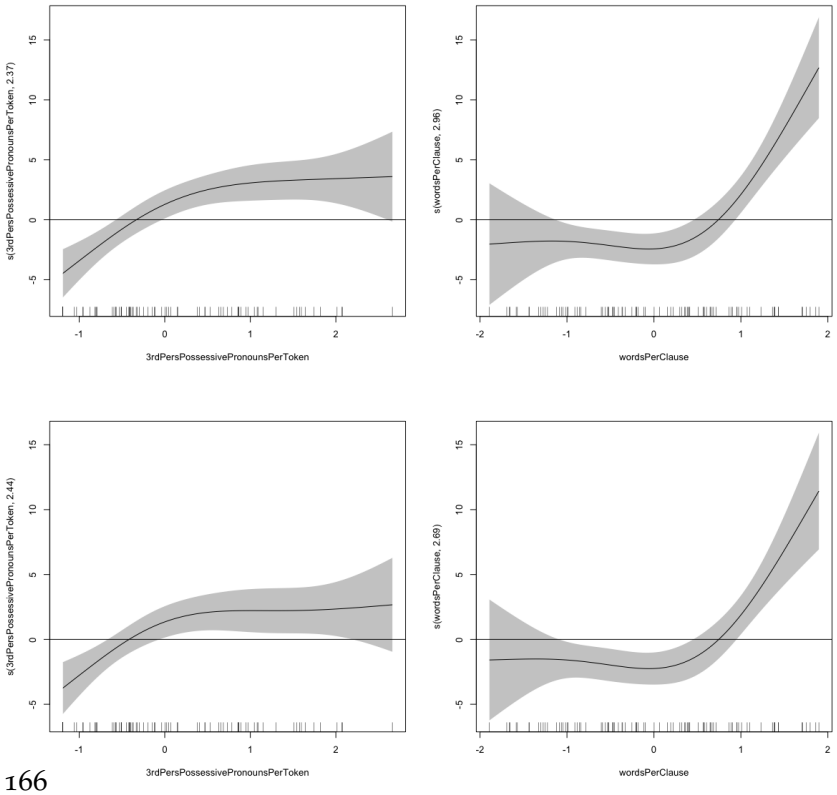
Table 14.3.: Summary of *Falko* GAM. Predicts course level from scaled and transformed complexity measures estimated on inverse *Falko Georgetown L2* data.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	1.9495	0.1988	9.8082	< 0.0001
logLexTypesNotFoundInKCTPerLexType	0.5458	0.2209	2.4706	0.0135
3rdPersPossessivePronounsPerToken	0.3900	0.1893	2.0604	0.0394
wordsPerClause	0.8822	0.2443	3.6114	0.0003
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(ungDerivationPerTokenSquared)	2.7078	3.3318	37.6659	< 0.0001

Table 14.4.: Summary of *Falko* GAM without task-effects. Predicts course level from scaled and transformed complexity measures estimated on book review *Falko Georgetown L2* data.



(a) Fitted on *Falko Georgetown L2* full data data set.



(b) Fitted on *Falko Georgetown L2* inverse (left) and book review (right) data sets.

Figure 14.4.: Smooths of GAMM with by-subject random intercepts fitted on the control data sets.

sets, too. However, the shape of the slope varies considerably: For the curricular tasks, the slope only starts to increase for later courses, while for the inverse and book review data set, the increase is linear and less steep. On the full data set, the increase is nearly linear, but levels off for higher proficiency. A possible explanation for these results might be found in the instructions given with the curricular task prompts: For the first two tasks, these require learners to focus on i) the difference between past and present tense; ii) spelling and punctuation; iii) adequate vocabulary; iv) the use of subjunctive and v) case government of prepositions. The other tasks focus more on subordination, relative clauses, and the use of adjectives as modifiers (*Das Falko-Handbuch Korpusaufbau und Annotationen*: 10ff). It could be possible, that this explicit focus on lexicon and morphology in the less advanced courses causes learners to show less development in terms of modification and subordination, thus inhibiting increases in clause length. Similarly, the focus on exactly these structures in the task prompts for the more advanced writing courses would promote their occurrence in the respective texts, which leads to a steeper slope for higher course levels. The steady, linear increase, that may be observed on the inverse and book review data would then show increases in clause length when a comparable set of learners is not directly prompted to focus in particular on a certain aspect of writing.

The *ratio of third person possessive pronouns per token* is a local measure of co-reference, which increases with more advanced writing courses across all data sets. However, the increase is linear on the full and book review data set, while on the longitudinal data it initially increases linearly, but considerably levels off for more advanced courses. Again, this is likely to be due to the task differences across levels, since the increase remains significant and stable for the full and the book review data set. It is unexpected, though, that the increase is not significant on the inverse data set, despite being significant at $\alpha < 0.05$ on the book review data set, because both data sets share most data. However, the estimate's standard deviation is quite high. This might cause the lack of significance: the book review data set is larger than the inverse data set, which might explain its larger estimate and lower standard deviation. In any case, since the book review data set is not only larger but also more theoretically principled, it was decided to disregard the insignificance of the predictor on the inverse data set for the interpretation of the measure. The results on the longitudinal and book review data set show, that more advanced writers produce more third person possessive pronouns and while the measure seems to be sensitive to task factors, as seen by the leveling off on the curricular

tasks, the linear increase on the book review data shows it to assess aspects of increased proficiency beyond that. Also, the significance of the measure on the full data set shows, that the two slopes from these two data sets may be combined to a reasonable single slope and remain significant.

The *log ratio of lexical types not found in the KCT corpus* is a measure of language use. Interestingly, it is based on a corpus of German L1 childen's writings from elementary and the first years of secondary school. Hence, it is less expected to be informative for adult L2 writings, yet, the slope shows significant linear increases on the full data set at $\alpha < 0.01$. One possible explanation, that comes to mind for the longitudinal data, is that the first two curricular tasks are similar to the writing prompts used to elicit children's writings in the KCT corpus, were pupils are asked to either elaborate on a fictional personal experience or to continue a story they heard before. This might contribute to the observation, that less shared vocabulary between a text from the longitudinal data set and the KCT corpus indicates more advanced writing. However, this would fail to account for the significance of the measure on the book review data ($\alpha < 0.05$).[1] Hence, the observed effect has to be due to differences between less and more advanced writing independent of task differences, which is highly interesting.

On a final note, it should be pointed out that on both, the longitudinal and the full data set, the random intercept for subject id seems to be completely uninformative, because the models assign zero edf to it when fitted on either data set. This is quite unexpected on the longitudinal data set, because learners should exhibit differences in their writing. Two potential explanations come to mind, though: First, it could be the case, that the longitudinal data set does not provide enough data instances for each learner. Unlike other longitudinal data sets, which typically have an even distribution of texts across learners, here, learners may be represented with 2 to 4 data points. Second, the model consists of only four measures, of which two were shown to exhibit non-linear slopes that do not quite match the theoretical expectation. This was argued to be potentially caused by the rather elaborate task prompts, which include detailed descriptions of which grammatical constructions and vocabulary to use, i.e. coincide with the aspects of language assessed by the model. Hence, it could be possible, that the strict task requirements i) make writings across learners within a course level more similar to each other in terms of their

[1] Again, less significant results on the inverse data set are disregarded following the reasoning outlined above. However, note also, that for this measure the slope is actually marginally significant at $\alpha < 0.10$, which strengthens the assumption that the lack of more significant results is due to idiosyncratic properties of the inverse data set for this measures.

writing style as it is expressed by the four measures; and ii) make writings within learners across course levels less similar to each other in terms of the subject's individual writing style as it is expressed by the four measures. This would clearly interfere with the informativeness of subject id in the model. To test this hypothesis, more investigations are required. In particular, it would be necessary to investigate whether random subject effects become significant, when building a model with measures, that are not specifically targeted by the task prompts. However, this was beyond the scope of this study, because the focus is on task and not on subject effects.

14.3. Study 3.2: Modeling Task-Effects

14.3.1. Model Description

Due to the findings in the previous study, subject id was discarded as a predictor: This is an unconventional step, because the data does contain repeated measures. Methodologically, this is bound to be accounted for with a mixed model design. However, since the GAMMs in Study 3 clearly showed, that there was no usable information to be gained from any form of random effect, it was concluded that the information was superfluous. Furthermore, the estimation of the subject id parameter did limit the amount of interactions, that could be included in a model, while remaining able to converge when estimating parameters. Given that the by-subject random effects did not contribute to model fit, while limiting the possibilities of the next study, it seemed more reasonable to discard the effect.

The model was augmented with interactions between code complexity and the complexity measures following the approach outlined in Section 12.2. When trained on the longitudinal data, this resulted in the *interaction* model shown in Figure 14.5. It includes all complexity measures from the subject model and introduces three interactions of code complexity with the ratio of *ung* derivations, the frequency of lexical types not found in KCT, and the ratio of third person possessive pronouns. Only the ratio of words to clauses does not show an interaction with code complexity on the longitudinal data. Note that for the two non-linear predictors, the number of knots per smooth was again limited to four knots per smooth, which did not alter the slopes overall shape but restricted the model to a more reasonable amount of consumed edf.

```
421 gam.falko.interaction <- gam(courseLevel ~
422                               ungDerivationPerTokenSquared:codeComplexity +
423                               logLexTypesNotFoundInKCTPerLexType:codeComplexity +
424                               s(thirdPersPossessivePronounsPerToken, by=codeComplexity, k=4) +
425                               s(wordsPerClause, k = 4) +
426                               codeComplexity,
427                             data=falko.long,
428                             family=ocat(R=4))
```

Figure 14.5.: Model formula of *Falko Georgetown L2* interaction model predicting course level from scaled and transformed complexity measures, trained on the longitudinal data.

14.3.2. Model Fit

The interaction model explains $R^2 = 0.9940$ of the deviance on the longitudinal data set and is clearly overfitting, using with 11 edf nearly twice the recommended number. On the full model, it explains $R^2 = 0.5720$ of the deviance, when including the significant interaction for the ratio of third person possessive pronouns, indicating that while the model might explain more deviance when adding more complexity measures, it is already quite informative. Unfortunately, the other two reference data sets could not be used for comparison, because they mostly or exclusively contain texts from the book review reference task and thus do not show any variance in code complexity.

The analysis of residual errors showed again a generally uniformly distributed errors with mean values around zero and overall homoscedastic standard deviations across all data sets, see Figure 14.6. Again, 2 outliers were identified on the longitudinal data set, but their removal did not effect the model in terms of explained deviance, parameter coefficients, or REML scores. However, unlike in Study 3, even with these outliers mean residual errors were adequate with $\mu = 1.21; sd = 10.94$. Also, the data set is already so limited that it did not seem justifiable to remove more data. Hence, the outliers were kept in the model. The same holds for the re-fitted model on the full data set, which has mean residual errors of $\mu = 0.02; sd = 37.72$.

Also, by subject residuals were analyzed and showed overall improvable, not quite evenly distributed errors, see Figure B.1 in Appendix B. While these results are far from ideal, it seems reasonable to assume, that the limited amount of data per subject cause this unfavorable error distribution.

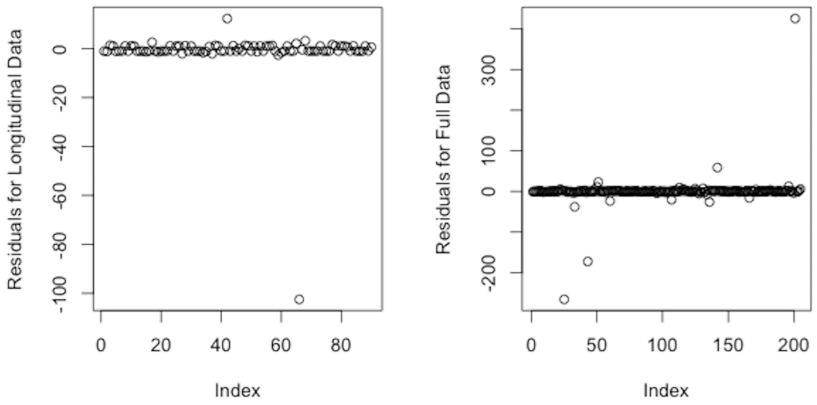


Figure 14.6.: Residuals for the subject model trained on the i) longitudinal (panel 1); and ii) full (panel 2); data set.

14.3.3. Model Discussion

As already mentioned, including all three interactions at once leads to a severely overfitting model. This even causes the ordinal model to estimate the inner boundaries for the response at $[-1, 39.59, 53.06]$, i.e. to make them considerably larger than for the previous model or any model with only a single interaction. This makes it more difficult to compare the model with the one from Study 3 or its re-fitted version on the full data set. Furthermore, it leads to small estimates becoming significantly different from the intercept, even if they weren't on a less severely overfitting version of the model. Hence, it was decided to base the model discussion on the three model summaries for each model containing only a single interaction. Before, it was confirmed, that the combination of interactions did not effect the estimates in terms of their direction or strength. These three model summaries may be found in Tables 14.6, 14.7, and 14.8 at the end of this chapter. The plotted smooths for all three models are in Figures 14.8a and 14.8b, again at the end of the chapter. Model summary and plotted smooths for the model including all three interactions may be found in Table B.1 and Figure B.2 in Appendix B.

When comparing all three model summaries and the plots of the non-linear predictors, it may be seen clearly that i) the interactions have no impact on each other, and ii) the estimates for high code complexity texts are slightly raised compared to the estimates for the same predictor without an interaction but otherwise virtually

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	2.4546	0.3129	7.8440	< 0.0001
logLexTypesNotFoundInKCTPerLexType	0.9746	0.2369	4.1133	< 0.0001
codeComplexity[HIGH]	0.3234	0.4152	0.7789	0.4360
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(ungDerivationPerTokenSquared)	2.2858	2.8080	33.7170	< 0.0001
s(3rdPersPossessivePronounsPerToken):codeComplexity[LOW]	1.0001	1.0002	0.4450	0.5048
s(3rdPersPossessivePronounsPerToken):codeComplexity[HIGH]	2.0939	2.6020	36.7289	< 0.0001
s(wordsPerClause)	2.4245	3.0906	31.6839	< 0.0001

Table 14.5.: Summary of *Falko* GAM with single code complexity interactions that is significant on full Falko Georgetown L2 data set.

the same, while none of the predictors has a slope that significantly differs from zero for low code complexity texts. Put differently, except for the ratio of words per clause, which shows no interaction with code complexity, all other measures only show a significant effect for high code complexity texts.

However, when trying to confirm these observations by on the full data set, only the effect for the *ratio of third person possessive pronouns per token* remains significant. The model summary and plots of the smoothed predictors may be found in Table 14.5 and Figure 14.7. All slopes remain virtually the same, when compared to the model without interactions, that is trained on the full data. The only exceptions are, that the *log ratio of lexical types found in the KCT corpus* becomes fully linear and *words per clauses* does not level off for more advanced classes anymore. Instead, it becomes nearly linear with wide standard deviations due to data sparsity at the respective part of the slope.

As for the interaction of *ratio of third person possessive pronouns per token* with code complexity, high code complexity texts again show an increase with advancing course level and no changes for the low code complexity condition. Thus, the interaction seems reliably to indicate, that the use of third person possessive pronouns increases with more advanced levels in high code complexity conditions. Low code complexity task conditions fail to elicit these productions, though. Since other cohesive devices, such as transitions or argument overlap, are not included in the model, it is not quite clear, whether this difference means, that learners write less cohesive in low code complexity conditions, or simply use different devices. Another explanation in line with previous observations in the cross-corpus analysis and the observations on *Merlin* would be, that learners confronted with high code complexity rather establish cohesion via pronoun use than via repetition of aforementioned discourse referents.

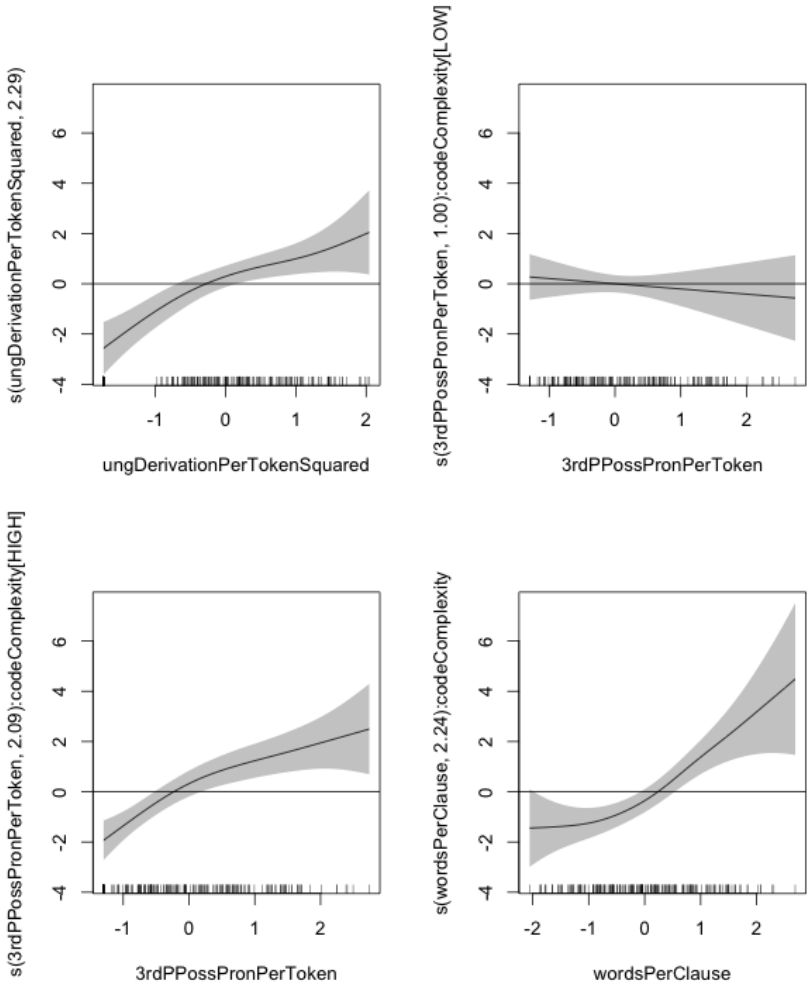


Figure 14.7.: Smooths of GAM trained on full *Falko Georgetown L2* data set.

A non-linguistic expmalantion would be, that the interaction is not due to code complexity, but an artifact of the distinction between book reviews and curricular writing tasks (since high code complexity texts are book reviews or texts from the first course level). To ensure, that this is not the case, Study 4 was repeated on the full data set using book review vs. curricular task as interaction instead of code complexity. However, this did not lead to a significant interaction with the *ratio of third person possessive pronouns per word* or any other complexity measure. This was taken to confirm the task-effect for code complexity.

Since the other two interactions from the longitudinal data set disappear on the full data set, it seemed reasonable to test, whether the interactions may be explained by idiosyncratic distributional properties of the data. For this, the longitudinal and the inverse data set were compared, since they are partitions of the full data set. This comparison showed that course level 4 is under-represented on the longitudinal data set. Hence, to confirm that the significance of the interactions on the longitudinal data is not due to the under-representation of level 4 texts, the data was augmented with 10 more texts from this course level by random re-sampling without replacement. Based on this augmented longitudinal data, interactions were again tested for their significance. However, the interactions of code complexity with the *squared ratio of ung derivations* and the word frequency measure remained significant on the augmented data, too. This shows, that the interactions are genuinely present in the longitudinal data, but for an unknown reason do not replicate to the full data set. It was thus decided not to rely on these interactions.

14.4. Summary

The studies conducted on *Falko Georgetown L2* yielded in fewer, but nevertheless interesting results with regard to the first two research questions

1a. Which complexity measures are elicited for model design in an exploratory, data-driven approach?

1b. Which insights do these measures give into L2 proficiency?

Due to the limited data support, only four measures could be investigated. As with *Merlin*, though, these assessed diverse domains of complexity. Also, the measures showed an overall coherent progression with increasing course level across data sets: More advanced writing is modeled in terms of increased nominalizations using *ung*

derivations and by increased syntactic complexity, as globally assessed by clause length. Whether this is due to increased subordination or phrasal modification could not be established, though. This matches with previous observations from the descriptive cross-corpus analysis. Furthermore, the models show, that language use in advanced writing becomes less similar to language use in a children's writing corpus. This is in principle reasonable, but it surprises, that high-intermediate learners in course levels 1 and 2 seem to exhibit enough overlap with language use in KCT to make the reduced similarity for higher course levels informative. Interestingly, as in the previous two studies on *Merlin*, the ratio of third person possessive pronouns is again included in the model. However, unlike on *Merlin*, where the binarized measure was associated with decrease proficiency when present, 3rd person possessive pronouns indicate advanced proficiency, here.

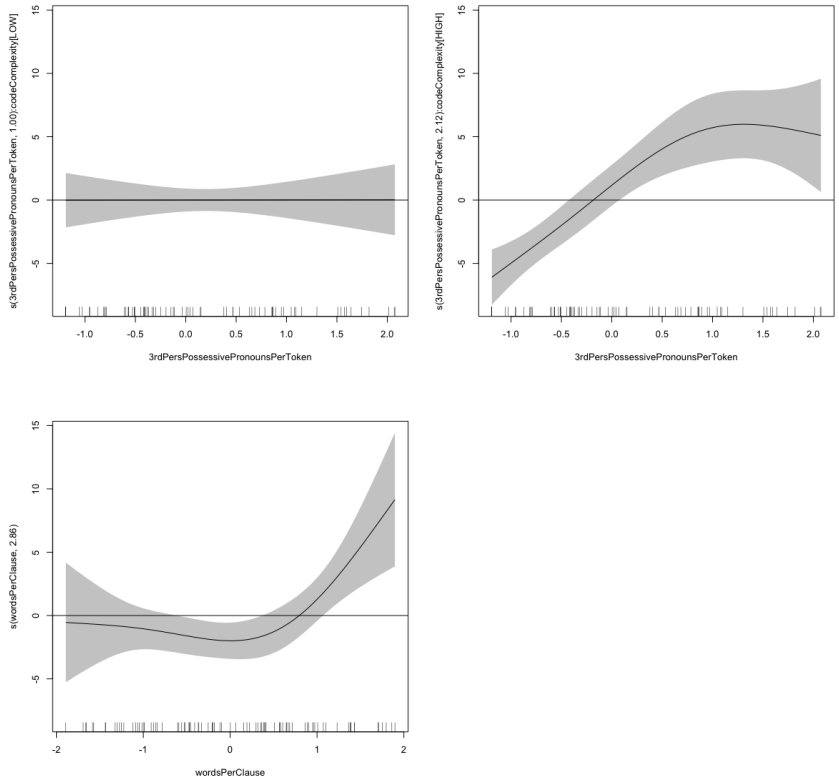
These diverging tendencies have already been observed in the descriptive cross-corpus analysis. There, they have been attributed to the task sensitivity of the person feature in pronouns use. This leads to the answers found for other two research questions,

2a. Are these measures influenced by code complexity?

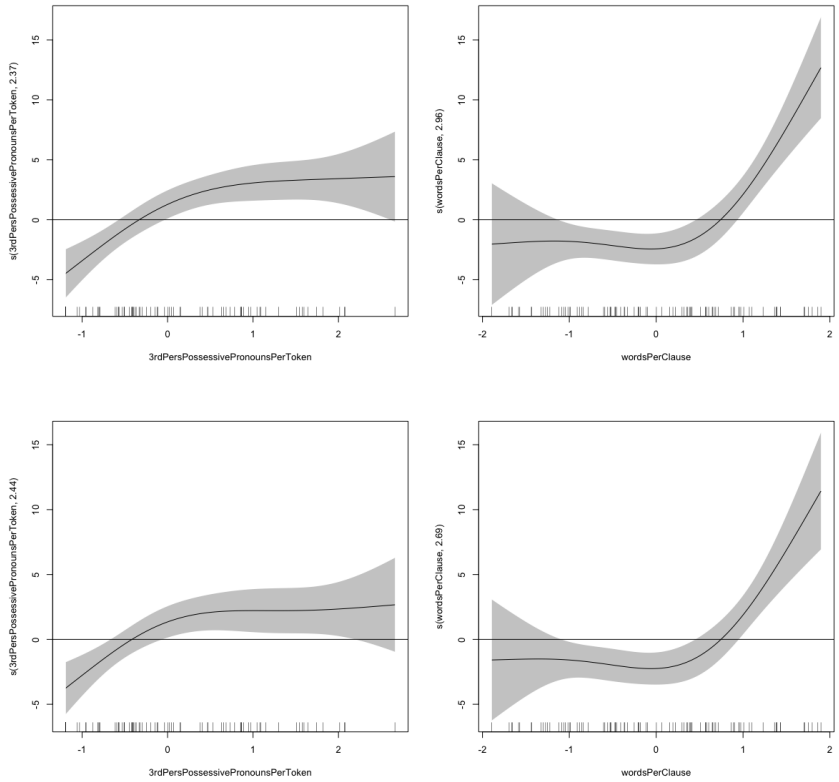
2b. If so, how are measures influenced by code complexity?

In Study 4, only a single interaction with code complexity was found in both, the longitudinal and the full data set, namely third person possessive pronouns, which is shown to increase only for high code complexity conditions, but is insignificant for low code complexity conditions. It was confirmed, that this interaction is not due to the difference between homogeneous and heterogeneous task backgrounds, which was considered possible, since high task complexity only occurs in book reviews and in the course level 1 curricular task. Interestingly, similar results were found for *ung* derivations and words not found in KCT, but these results did not remain stable on the full data set. However, when testing, whether these were only present in the longitudinal data due to idiosyncratic properties of the data, it was found that on this data set, the interactions persisted. These effects are either due to some idiosyncratic properties in *Falko Georgetown*, that could not be identified, or they genuinely show, that high code complexity elicits more complex language in terms of co-reference, *ung* derivations, and vocabulary use at least in certain task contexts. This would confirm Robinson 2001's assumption of resource-directing effects of certain cognitive task factors, rather than Skehan 1996's original assumption, that high code complexity depletes cognitive resources from language

performance. Further research on this question on a more balanced data set seems in order.



(a) GAM with interaction for *ratio of third person possessive pronouns per token*.



(b) GAM with interaction for *log ratio of lexical types found in the KCT corpus* (upper) and for *squared ratio of ung derivation per token* (lower).

Figure 14.8.: Smooths of GAMs with individual code complexity interactions on the logitudinal *Falko Georgetown L2* data.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	8.2324	0.7339	11.2179	< 0.0001
ungDerivationPerTokenSquared	2.6094	0.5345	4.8821	< 0.0001
logLexTypesNotFoundInKCTPerLexType	4.0419	0.9439	4.2823	< 0.0001
codeComplexity[HIGH]	0.2123	1.2358	0.1718	0.8636
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(3rdPersPossessivePronounsPerToken):codeComplexity[LOW]	1.0000	1.0001	0.0001	0.9903
s(3rdPersPossessivePronounsPerToken):codeComplexity[HIGH]	2.1222	2.4341	36.5027	< 0.0001
s(wordsPerClause)	2.5675	2.8641	15.9337	0.0008

Table 14.6.: Summary of Falko GAM with code complexity interaction for *ratio of third person possessive pronouns per token*. Predicts course level from scaled and transformed complexity measures estimated on longitudinal Falko Georgetown L2 data.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	8.1194	0.7828	10.3718	< 0.0001
ungDerivationPerTokenSquared	3.2446	0.4946	6.5606	< 0.0001
codeComplexity[HIGH]	1.7647	1.3434	1.3137	0.1890
logLexTypesNotFoundInKCTPerLexType[LOW]	1.5412	0.9419	1.6362	0.1018
logLexTypesNotFoundInKCTPerLexType[HIGH]	7.0638	1.4521	4.8646	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(3rdPersPossessivePronounsPerToken)	2.3676	2.7221	32.3620	< 0.0001
s(wordsPerClause)	2.7523	2.9560	47.9031	< 0.0001

Table 14.7.: Summary of Falko GAM with code complexity interaction for *log ratio of lexical types found in the KCT corpus*. Predicts course level from scaled and transformed complexity measures estimated on longitudinal Falko Georgetown L2 data.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	6.8087	0.6750	10.0870	< 0.0001
logLexTypesNotFoundInKCTPerLexType	2.8971	0.7686	3.7691	0.0002
codeComplexity[HIGH]	-0.0822	1.2204	-0.0673	0.9463
ungDerivationPerTokenSquared[LOW]	1.2325	0.7737	1.5929	0.1112
ungDerivationPerTokenSquared[HIGH]	3.3459	0.6749	4.9580	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(3rdPersPossessivePronounsPerToken)	2.4405	2.7913	17.2582	0.0009
s(wordsPerClause)	2.6941	2.9327	34.3038	< 0.0001

Table 14.8.: Summary of Falko GAM with code complexity interaction for *squared ratio of ung derivation per token*. Predicts course level from scaled and transformed complexity measures estimated on longitudinal Falko Georgetown L2 data.

Part VII.

Conclusion

15. Conclusion

This thesis investigated German L2 proficiency on *Merlin* and *Falko Georgetown*b y means of a broad variety of complexity measures under consideration of task-effects. For this, available task information and retrospective analyses of cognitive and functional task factors were used to augment i) a descriptive cross-corpus analysis of more than 100 complexity measures representing theory-based selected concepts from the domains of language use, human language processing, discourse and encoding of meaning, and three sub domains of theoretical linguistics, namely syntax, lexicon and semantics, and morphology; and ii) two exploratory regression studies using a small, data-driven selection of complexity measures to model proficiency in ordinal GAMs including interactions for either task theme (on *Merlin*) or code complexity (on *Falko Georgetown*).

These investigations addressed overall three research questions:

1. How do measures of complexity model German L2 proficiency?
2. To which extend is this influenced by cognitive or functional task-effects?
3. Does a retrospective analysis of German learner corpora with diverse task backgrounds improve complexity-based L2 proficiency modeling?

To answer the first question, the descriptive analysis of more than 100 complexity measures provided a valuable overview and general impression especially with regard to the generalizability of observable trends across corpora.[1] Of particular interest are two general insights: First, most measures addressing a certain construct, such as noun complexity, show a homogeneous development. This confirms, that for most measures, it is valid to make assumptions about the entire concept after measuring only a small selection of measures from that construct. Second, many general tendencies, that could be observed within a construct, generalized across corpora and showed similar ratios at comparable proficiency levels. This

[1] Also, in the course of the thesis, comparable analyses were conducted for all 398 complexity measures. While it was not possible to discuss all of these plots, they are made available in the online supplementary material to this thesis at `http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/supplementary-material/complexity-plots/`.

speaks for the generalizability of findings across corpora. The GAMs complement these findings with a deeper analysis of a selected few measures. Interestingly, the exploratory model design approach employed on both corpora lead to very diverse models in terms of measures. This indicates, that a varied set of complexity measures is not only theoretically desirable, but also asserts itself in data-driven approaches. Another important finding with respect to the first research question is, that ordinal GAMs allow for clear insights into the contributions of individual complexity measures, while achieving remarkable results in classification experiments, despite the comparatively small number of measures included in the model.

As for the second question, results on the descriptive study showed, that some measures were more sensitive to heterogeneous task backgrounds (e.g. tenses, person markings in pronouns) than others (e.g. integration costs (DLT), noun complexity), which is highly interesting. Given that not all data sets or applications of complexity analyses may support a thorough analysis of task factors, while still expecting heterogeneous task backgrounds, these results give a first indication of which measures to use, when generating a task-stable complexity index or model. This information might, for example, be highly valuable for other areas, such as readability assessment, too. The regression analyses were designed to allow for a focused analysis of a specific task factor. On both corpora, task-effects were found for the respective factors, and interactions lead to interesting insights and improved model fit. When comparing the results, it also seems like task factors predominantly effect local complexity measures instead of global measures of certain language structures, since neither of the length measures in any model showed significant task interactions. This impression seems to be worth further investigations, in particular with regard to the above mentioned potential uses for task-stable complexity measures.

However, despite the presence of interactions and the significantly improved model fit, task interactions did not significantly improve classification results on *Merlin*. Also, some of the significant interactions were unstable across model settings: On *Merlin*, most interactions changed considerably after introducing performance effects, and on *Falko Georgetown*, most interactions from the longitudinal data were not stable on the full data. This is problematic for the more detailed interpretation of the particular task-effects, because it cannot be ruled out, that the observed issues are due to some idiosyncratic properties of the data. This makes it difficult to either draw conclusions from the respective interactions. Thus, with regard to the third research question, i.e. the extend to which the retrospective

analysis of task factors on both corpora improves modeling results, a mixed answer has to be given: On the one hand, task factors improved model fit significantly on *Merlin*, and lead to interesting insights on both corpora. On the other hand, a lot of uncertainty remains due to the persistently unbalanced distribution of tasks and task factors across various other variables, such as proficiency scores, performance, etc.

Generally, while the analyses conducted on *Merlin* and *Falko Georgetown* already lead to some valuable insights, the studies mostly prompted new research questions due to the limited scope of the thesis and the above mentioned mixed findings. In particular, two lines of future investigations derive from the presented work: First, more detailed and controlled analyses of task-effects as well as of performance effects in *Merlin* are in order, to establish to which extend the observations made in this thesis are caused by idiosyncratic distributional properties of the data. Also, the investigation of task-effects for task type is still open. In this context, a cross-corpus validation of the classification performance of the success model from Study 2 would be highly interesting, too, provided a suited test corpus may be found. Second, the collection of *Falko* corpora invites for a series of comparative studies of L1 and L2 language performance, since in these corpora, elicitation contexts including task factors were controlled for when compiling the corpora. This allows in particular to address the issue of adequacy in L2 writing, which was briefly touched in the theoretical part of this thesis. Depending on the set up of the other *Falko* corpora, it might also be possible to compare task-effects on L1 and L2 writing. Additional to these two main lines of investigation, there are several concrete additional analyses, that were beyond the scope of this thesis and are planned for future work, such as the investigation of an L1 threshold for high integration cost areas for German L1 writings or the investigation of the acquisition of German simple and past perfect by intermediate learners.

To conclude, the conducted analyses were highly informative, but more research on task-effects is required. All results obtained in this thesis show, that complexity analyses should account for task-effects, either by analyzing data from homogeneous task backgrounds or by retrospectively analyzing tasks in learner corpora for task factors: While complexity measures are not necessarily affected by task differences, they may introduce variation to the data, that leads to erroneous conclusions about complexity. This could explain some of the variation in findings across complexity studies. The presented work contributes methodologically to this, by providing detailed operationalizations of various cognitive and functional task factors, as well

as by providing such an analysis for *Merlin*, which is one of the largest German learner corpora with a particularly diverse task background. However, the findings also show, that it would be invaluable for research on complexity and all CAF measures, to elicit learner data collection in controlled task settings, in order to promote research on task-effects.

Part VIII.

Bibliography

Bibliography

Abel, Andrea et al. 2013. *merlin: A Trilingual Learner Corpus illustrating European Reference Levels*. LRC 2013. Bergen, Norway.

Alexopoulou, Theodora et al. 2017. Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. In: *Language Learning*, pp. 1–29.

Baayen, Harald R. & Divjak, Dagmar. 2016. "Ordinal GAMMs: a new window on human ratings". Submitted.

Baayen, Harald R. et al. 2017. The Cave of Shadows. Addessing the human factor with generalized additive mixed models. In: *Journal of Memory and Language*, pp. 206–234.

Babynak, Michael A. 2004. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. In: *Psychosomatic Medicine* 66, pp. 411–421.

Barzilay, Regina & Lapata, Mirella. 2008. Modeling local coherence: An entity-based approach. In: *Computational Linguistics* 34, pp. 1–34.

Berman, Ruth A. & Slobin, Dan Isaac. 1994. *Relating events in narrative: A crosslinguistic developmental stud*. Hillsdale: Lawrence Erlbaum.

Bestgen, Yves & Granger, Sylviane. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. In: *Journal of Second Language Writing* 26, pp. 28–41.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. In: *Language* 62, pp. 384–411.

Biber, Douglas. 1988. *Variation across speech and writing*. Camebridge, United Kingdom: Camebridge University Press.

Biber, Douglas, Conrad, S. & Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*. Camebridge, United Kingdom: Camebridge University Press.

Biber, Douglas, Gray, Bethany & Poonpon, Kornwepa. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? In: *Tesol Quarterly* 45, pp. 5–35.

Biber, Douglas, Gray, Bethany & Staples, Shelley. 2014. Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. In: *Applied Linguistics*, pp. 1–31.

Birchenough, Julia M. H., Davies, Robert & Connelly, Vincent. 2016. Rated age-of-acquisition norms for over 3,200 German words. In: *Behavior research methods*, pp. 1–18.

Bley-Vroman, R. 1983. The comparative fallacy in interlanguage studies: The case of systematicity. In: *Language Learning* 33, pp. 1–17.

Bohnet, Bernd & Nivre, Joakim. 2012. "A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. Jeju Island, Korea: Association for Computational Linguistics, pp. 1455–1465.

Borchers, Hans Werner. 2017. *pracma: Practical Numerical Math Functions*. R package version 2.0.4.

Boslaugh, Sarah & Watters, Paul Andrew. 2008. *Statistics in a Nutshell*. Sebastopol, USA: O'Reilly.

Brown, Cati et al. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. In: *Behavior research methods* 40 (2), pp. 540–545.

Bruner, Jerome S. 1986. *Actual minds, possible words*. Camebridge: Harvard University Press.

Brysbaert, Marc & New, Boris. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. In: *Behavior research methods, instruments & computers* 41 (4), pp. 977–990.

Brysbaert, Marc et al. 2011. The Word Frequency Effect A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. In: *Experimental Psychology* 58, pp. 412–424.

Bulté, Bram & Housen, Alex. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. In: *Journal of Second Language Writing* 26, pp. 42–65.

Chen, Xiaobin & Meurers, Detmar. 2016. "CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pp. 113–119.

Crossley, Scott A, Kyle, Kristopher & McNamara, Danielle S. 2015. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. In: *Behavior research methods*. Springer, pp. 1–11.

Crossley, Scott A, Kyle, Kristopher & McNamara, Danielle S. 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. In: *Journal of Second Language Writing* 32, pp. 1–16.

Crossley, Scott A. & McNamara, Danielle S. 2009. Computational assessment of lexical differences in L1 and L2 writing. In: *Journal of Second Language Writing* 18, pp. 119–135.

Crossley, Scott A. & McNamara, Danielle S. 2011. Shared features of L2 writing: Intergroup homogeneity and text classification. In: *Journal of Second Language Writing* 20, pp. 271–285.

Crossley, Scott A. & McNamara, Danielle S. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. In: *Journal of Research in Reading* 35 (2), pp. 115–135.

Crossley, Scott A. & McNamara, Danielle S. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. In: *Journal of Second Language Writing* 26, pp. 66–79.

Crossley, Scott A. et al. 2010. Predicting lexical proficiency in language learner texts using computational indices. In: *Language Testing* 20 (10), pp. 1–20.

Crossley, Scott A. et al. 2011. The Development of Writing Proficiency as a Function of Grade Level: A Linguistic Analysis. In: *Written Communication* 28 (3), pp. 282–311.

Crossley, Scott A. et al. 2014. Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. In: *Journal of Writing Assessment* 7 (1), pp. 1–15.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins Publishing.

Daller, Helmut, Hout, Roeland van & Treffers-Daller, Jeanine. 2003. Lexical richness in the spontaneous speech of bilinguals. In: *Applied Linguistics* 24 (2), pp. 197–222.

Duden (Gr). 2009. *Deutsche Grammatik*. Ed. by Ursula Hoberg & Rudolf Hoberg. 4th ed. Vol. 4. Der kleine Duden. Berlin, Germany: Dudenverlag.

Ellis, R. & Barkhuizen, G. 2005. *Analysing learner language*. Oxford: Oxford University Press.

Fabricius-Hansen, Cathrine. 2014. *Vorangestellte Attribute und Relativsätze im Deutschen: Wettbewerb und Zusammenspiel*.

Falko Georgetown Dokumentation. 2007. Humnoldt-Universität zu Berlin.

Fischer, Klaus. 2017. Komplexität – dennoch ein nützlicher Begriff. In: *Linguistische Komplexität – ein Phantom?* Vol. 94. Stauffenburg Verlag.

Foster, Pauline & Skehan, Peter. 1996. The influence of planning and task type on second language performance. In: *Studies in Second Language Acquisition*.

Foster, Pauline & Tavakoli, Parvaneh. 2009. Native speakers and task performance: Comparing effects on complexity, fluency, and lexical density. In: *Language Learning* 59 (4), pp. 866–896.

Foster, Pauline, Tonkyn, Alan & Wigglesworth, Gillian. 2000. Measuring Spoken Language: A Unit for All Reasons. In: *Applied Linguistics* 21 (3), pp. 354–375.

François, Thomas & Fairon, Cédrik. 2012. "An "AI readability" formula for French as a foreign language". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics. Jeju Island, Korea, pp. 466–477.

Frank, Eibe, Hall, Mark A. & Witten, Ian H. 2016. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann.

Frogner, Ellen. 1933. Problems of sentence structure in pupils' themes. In: *English Journal* 22, pp. 742–749.

Galasso, Sabrina. 2014. *Exploring Textual Cohesion Characteristics for German Readability Classification*. B.A. Thesis.

Geyken, Alexander. 2007. The DWDS Corpus: A reference corpus for the German language of the 20th century. In: *Collocations and idioms: Linguistic, lexicographic, and computational aspects*. Ed. by C. Fellbaum. London: Continuum Press.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In: *Image, language, brain*, pp. 95–126.

Gilquin, Gaëtanelle & Gries, Stefan Th. 2009. Corpora and experimental methods: A state-of-the-art review. In: *Corpus Linguistics and Linguistic Theory* 5 (1), pp. 1–26.

Graesser, Arthur C. et al. 2004. Coh-Metrix: Analysis of text on cohesion and language. In: *Behaviour Research Methods, Instruments, and Computers* 36 (2), pp. 193–202.

Gries, Stefan Th. 2008. Phraseology and linguistic theory: A brief survey. In: *Phraseology: An interdisciplinary perspective*. Ed. by S. Granger & F. Meunier. Amsterdam, Philadelphia: John Benjamins Publishing, pp. 3–25.

Gu, Ching. 2002. *Smoothing Spline ANOVA Models*. New York: Springer.

Guiraud, Pierre. 1960. *Problèmes et méthodes de la statistique linguistique*. Presses universitaires de France.

Hall, Mark et al. 2009. The WEKA Data Mining Software: An Update. In: *SIGKDD Explorations* 11 (1).

Hancke, Julia. 2013. "Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language". MA thesis. Eberhard Karls Universität Tübingen.

Hancke, Julia, Vajjala, Sowmya & Meurers, Detmar. 2012. "Readability Classification for German using lexical, syntactic and morphological features". In: *Proceedings of COLING*. Mumbai, pp. 1063–1080.

Hastie, Trevor & Tibshirani, Robert. 1986. Generalized Additive Models. In: *Statistical Science* 1 (3), pp. 297–318.

Hastie, Trevor & Tibshirani, Robert. 1990. *Generalized additive models*. Vol. 43. Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Heister, Julian et al. 2011. dlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung. In: *Psychologische Rundschau* 62, pp. 10–20.

Hennig, Mathilde. 2017. Linguistische Komplexität – ein Phantom? In: ed. by Mathilde Hennig. Stauffenburg Verlag. Chap. 1, pp. 7–18.

Hennig, Mathilde & Niemann, Robert. 2013. Unpersönliches Schreiben in der Wissenschaft: Eine Bestandsaufnahme. In: *Informationen Deutsch als Fremdsprache* 4 (439–455).

Henrich, Verena & Hinrichs, Erhard. 2010. "GernEdiT - The GermaNet Editing Tool". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, pp. 2228–2235.

Housen, Alex & Kuiken, Folkert. 2009. Complexity, Accuracy, and Fluency in Second Language Acquisition. In: *Applied Linguistics* 30 (4), pp. 461–437.

Housen, Alex, Vedder, Ineke & Kuiken, Folkert. 2012. Document Viewing Options: Title: Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA. In: vol. 32. Language Learning & Language Teaching. Amsterdam, Philadelphia: John Benjamins Publishing. Chap. 1–2.

Hunt, Kellogg. 1965. Grammatical structures written at three grade levels. In: *National Council of Teachers of English*.

Hunt, Kellogg. 1970. Syntactic maturity in school children and adults. In: *Monographs of the Society for Research in Child Development* 35.

Jarvis, Scott et al. 2003. Exploring multiple profiles of highly rated learner compositions. In: *Journal of Second Language Writing* 12, pp. 377–403.

Kennedy, Christopher & McNally, Louise. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. In: *Language* 81 (2), pp. 345–381.

Kintsch, Walter. 1974. *The representation of meaning in memory*. Hillsdale: Lawrence Erlbaum.

Knoch, Ute, Roushad, Amir & Storch, Neomy. 2014. Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? In: *Assessing Writing* 21, pp. 1–17.

Knoch, Ute et al. 2015. What happens to ESL students' writing after three years of study at an English medium university? In: *Journal of Second Language Writing* 28, pp. 39–52.

Kolechi, Joseph C. 2002. An Introduction to Tensors for Students of Physics and Engineering. In: *NASA Technical Memorandum*.

Kuhn, Max et al. 2016. *caret: Classification and Regression Training*. R package version 6.0-73.

Kyle, Kristopher. 2016. "Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication". PhD thesis. Georgia State University.

LaBrant, Lou L. 1933. A study of certain language developments of children in grades four to twelve inclusive. In: *netic Psychology Monographs* 14, pp. 387–491.

Larsen, Kim. 2015. GAM: The Predictive Modeling Silver Bullet. In: *http://multithreaded.stitchfix.com/blog/2015/07/30/gam/*.

Larsen-Freeman, D. 1978. An ESL index of development. In: *Tesol Quarterly* 12 (4), pp. 439–448.

Lavalley, Rémi, Berkling, Kay & Stüker, Sebastian. 2015. "Preparing Children's Writing Database for Automated Processing". In: *Workshop on L1 Teaching, Learning and Technology (L1TLT)*. Leipzig, Germany, pp. 9–15.

Levy, Roger & Andrew, Galen. 2006. "Tregex and Tsurgeon: tools for querying and manipulating tree data structures". In: *5th International Conference on Language Resources and Evaluation*.

Louwerse, Max M. et al. 2004. "Variation in language and cohesion across writ- ten and spoken registers". In: *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pp. 843–848.

Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. In: *International Journal of Corpus Linguistics* 15 (4), pp. 474–496.

Lu, Xiaofei. 2011. The relationship of lexical richness to the quality of esl learners' oral narratives. In: *The Modern Languages Journal* 96 (2), pp. 190–208.

Lu, Xiaofei & Ai, Haiyang. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. In: *Journal of Second Language Writing* 29, pp. 16–27.

Malvern, David et al. 2004. Lexical diversity and language development: Quantification and assessment. In: *Quantification and Assessment*. Palgrave Macmillan.

McCarthy, Philip M. & Jarvis, Scott. 2007. A theoretical and empirical evaluation of vocd. In: *Language Testing* 24, pp. 459–488.

McCarthy, Philip M. & Jarvis, Scott. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. In: *Behavior research methods* 42 (2), pp. 381–392.

McNamara, Danielle S., Crossley, Scott A. & McCarthy, Philip M. 2009. Linguistic Features of Writing Quality. In: *Written Communication*, pp. 1–30.

McNamara, Danielle S. et al. 2014. *Automated evaluation of text and discourse with Coh-Metrix.* Camebridge University Press.

Merlin project. 2014a. *task desciption: Essay: why it's of value to learn German*. `http://merlin-platform.eu/`.

Merlin project. 2014b. *task desciption: Formal letter: apply for internship in sales department*. `http://merlin-platform.eu/`.

Merlin project. 2014c. *task desciption: Formal letter: ask for information at Au pair Agency*. `http://merlin-platform.eu/`.

Merlin project. 2014d. *task desciption: Formal letter: Au pair writes letter of complaint to Agency*. `http://merlin-platform.eu/`.

Merlin project. 2014e. *task desciption: Formal letter to housing office*. `http://merlin-platform.eu/`.

Merlin project. 2014f. *task desciption: Informal e-mail: arrange an appointment with a friend to go swimming together*. `http://merlin-platform.eu/`.

Merlin project. 2014g. *task desciption: Informal e-mail: ask a friend for help with finding an apartment*. `http://merlin-platform.eu/`.

Merlin project. 2014h. *task desciption: Informal letter: ask friend to take care of pet*. http://merlin-platform.eu/.

Merlin project. 2014i. *task desciption: Informal letter: birthday congratulations*. http://merlin-platform.eu/.

Merlin project. 2014j. *task desciption: Informal letter: congratulate to birth of a child*. http://merlin-platform.eu/.

Merlin project. 2014k. *task desciption: Informal letter for New Year to a friend*. http://merlin-platform.eu/.

Merlin project. 2014l. *task desciption: Informal letter: offer a ticket not used to a friend*. http://merlin-platform.eu/.

Merlin project. 2014m. *task desciption: Informal letter to a friend announcing a visit*. http://merlin-platform.eu/.

Merlin project. 2014n. *task desciption: Online article: about sticking to one's traditions and "assimilation" in a new environment*. http://merlin-platform.eu/.

Merlin project. 2014o. *task desciption: Report: about the housing situation*. http://merlin-platform.eu/.

Meurers, Detmar. 2017. "Fördern Schulbuchtexte die sprachliche Entwicklung? Eine Analyse sprachlicher Komplexität". Workshop "Spracherwerb und Sprachförderung über die Lebensspanne III beim DIE, Bonn.

Miestamo, M. 2008. Grammatical complexity in a cross-linguistic perspective. In: *Language complexity: Typology, contact, change*. Ed. by M. Miestamo, K. Sinnemäki & F. Karlsson. Benjamins, pp. 23–41.

Norris, J. & Ortega, Lourdes. 2003. Defining and measuring SLA. In: *The Handbook of Second Language Acquisition*. Ed. by C. Doughty & M. Long. Blackwell.

Norris, John M. & Ortega, Lourdes. 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. In: *Applied Linguistics* 30 (4), pp. 55–578.

Ortega, L. 2012. Interlanguage complexity: A construct in search of theoretical renewal. In: *Linguistic complexity: Second language acquisition, indigenization, contact*. Ed. by B. Kortmann & B. Szmrecsanyi. Berlin: de Gryuter, pp. 127–155.

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

Pallotti, G. & Ferrari, S. 2008. Lavariabilità situazionale dell'interlingua: Implicazioni per la ricerca acquisizionale e il testing linguistico. In: *Competenze Lessicali e Discorsive nell'Acquisizione di Lingue Seconde*.

Pallotti, Gabrielle. 2009. CAF: Defining, Refining and Differentiating Constructs. In: *Applied Linguistics* 30 (4), pp. 590–601.

Pallotti, Gabrielle. 2015. A simple view of linguistic complexity. In: *Second Language Research* 31 (1), pp. 117–134.

Paquot, Magali. 2017. The phraseological dimension in interlanguage complexity research. In: *Second Language Research* Special Issue on Linguistic Complexity, pp. 1–25.

Patten, Bill van & Benati, Alessandro G. 2010. *Key Terms in Second Language Acquisition*. London, New York: Continuum.

Pearson, Hazel. 2013. A Judge-Free Semantics for Predicates of Personal Tase. In: *Journal of Semantics* 30, pp. 103–154.

Peduzzi, Peter et al. 1995. The importance of events per independent variable (EPV) in proportional hazards analysis I. Background, goals and general Strategy. In: *Journal of clinical epidemiology* 48 (12), pp. 1495–1501.

Peduzzi, Peter et al. 1996. A simulation study of the number of events per variable in logistic regression analysis. In: *Journal of clinical epidemiology* 49 (12), pp. 1373–1379.

Petrov, Slav & Klein, Dan. 2007. "Improved Inference for Unlexicalized Parsing." In: *HLT-NAACL*. Vol. 7, pp. 404–411.

Polio, Charlene. 2012. How to Research Second Language Writing. In: *Research Methods in Second Language Acquisition. A Practical Guide*. Ed. by Alison Mackey & Susan M. Gass. Guides to Research Methods in Language and Linguistics. Wiley-Blackwell. Chap. 9, pp. 158–179.

Polio, Charlene & Park, J.-H. 2016. Language development in second language writing. In: *Handbook of second and foreign language writing*. Ed. by R. Manchón & P. K. Matsuda. Mouton de Gruyter.

Projekt Tiger. 2003. *Tiger Annotationsschema*. Universität des Saarlandes, Universität Stuttgart, Universität Potsdam.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ravid, Dorit & Tolchinsky, Liliana. 2002. Developing linguistic literacy: A comprehensive model. In: *Journal of Child Language* 29 (2), pp. 417–447.

Reis, Marga. 2001. Bilden Modalverben im Deutschen eine syntaktische Klasse. In: *Modalität und Modalverben im Deutschen* 9, pp. 287–318.

Rescher, Nicholas. 1998. *Complexity: A philosophical overview*. Transaction Publishers.

Reznicek, Marc et al. *Das Falko-Handbuch Korpusaufbau und Annotationen*. Humnoldt-Universität zu Berlin.

Robinson, Peter. 1995. Task Complexity and Second Language Narrative Discourse. In: *Language Learning* 45 (1), pp. 99–140.

Robinson, Peter. 2001. Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework. In: *Applied Linguistics* 22 (1), pp. 27–57.

Robinson, Peter, Ting, Sarah Chi-Chien & Urwin, Jian Jun. 1995. Investigating Second Language Task Complexity. In: *RELC Journal* 2 (62-79).

Samuda, V. & Bygate, M. 2008. *Tasks in second language learning*. New York: Palgrave Macmillan.

Sanell, A. 2007. "Parcours acquisitionnel de la négation et de quelques particules de portée en français L2". PhD thesis. Institutionen för franska, italienska och klassiska språk.

Schlömer, Anne. 2013. *Erweiterte Nominalgruppen als Merkmal von Wissenschaftssprache. Eine Analyse in Schülertexten und Lehrbüchern*.

Schmidt, Karsten. 2016. Der graphematische Satz. In: *Zeitschrift für germanistische Linguistik* 44 (2), pp. 215–256.

Schröder, Astrid et al. 2012. German norms for semantic typicality, age of acquisition, and concept familiarity. In: *Behavior research methods* 44 (2), pp. 380–394.

Shain, Cory et al. 2016. "Memory access during incremental sentence processing causes reading time latency". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pp. 49–58.

Skehan, Peter. 1996. A Framework for the Implementation of Task-based Instruction. In: *Applied Linguistics* 17 (1), pp. 38–62.

Skehan, Peter. 1998. *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, Peter & Foster, Pauline. 1997. Task type and task processing conditions as influence on foreign language performance. In: *Language Teaching Research* 1 (3), pp. 185–211.

Taguchi, N., Crawford, W. & Wetzel, D. Z. 2013. What Linguistic Features Are Indicative of Writing Quality? A Case of Argumentative Essays in a College Composition Program. In: *Tesol Quarterly* 47 (2), pp. 420–430.

Tavakoli, Parvaneh & Foster, Pauline. 2011. Task design and second language performance: The effect of narrative type on learner output. In: *Language Learning* 61, pp. 37–72.

Tavakoli, Parvaneh & Skehan, Peter. 2005. Strategic planning, task structure, and performance testing. In: *Planning and task performance in a second language* 11, pp. 239–273.

Thomas, Margaret. 1994. Assessment of L2 proficiency in second language acquisition research. In: *Language Learning* 44, pp. 307–336.

Thorndike, E. L. 1921. Word Knowledge in the Elementary School. In: *Teachers College Record* 28 (5), pp. 334–370.

Todirascu, Amalia et al. 2013. Coherence and cohesion for the assessment of text readability. In: *Natural Language Processing and Cognitive Science* 11, pp. 11–19.

Tracy-Ventura, Nicole & Myles, Florence. 2015. The importance of task variability in the design of learner corpora for SLA research. In: *International Journal of Learner Corpus Research* 1 (1), pp. 58–95.

Vajjala, Sowmya. 2015. "Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications". PhD thesis. Eberhard Karls Universität Tübingen.

Vajjala, Sowmya & Meurers, Detmar. 2012. "On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition". In: *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*. Vol. 7. Association for Computational Linguistics. Montréal, Canada, pp. 163–173.

van Rij, Jacolien et al. 2016. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. R package version 2.2.

Vasylets, Olena, Gilabert, Roger & Manchón, Rosa M. 2017. The Effects of Mode and Task Complexity on Second Language Production. In: *Language Learning* 67 (2), pp. 394–430.

Vogel, M. & Washburne, C. 1928. An Objective Method of Determining Grade Placement of Children's Reading Material. In: *The Elementary School Journal* 28, pp. 373–381.

von der Brück, Tim & Hartrumpf, Sven. 2007. "A Semantically Oriented Readability Checker for German". In: *Proceedings of the 3rd Language & Technology Conference*, pp. 270–274.

von der Brück, Tim, Hartrumpf, Sven & Helbig, Hermann. 2008. A Readability Checker with Supervised Learning Using Deep Indicators. In: *Informatica* 32, pp. 429–435.

Vyatkina, Nina. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. In: *Modern Language Journal* 96, pp. 576–598.

Weiß, Zarah Leonie. 2015. *More Linguistically Motivated Features of Language Complexity in Readability Classification of German Textbooks: Implementation and Evaluation*. B.A. Thesis. Tübingen, Germany.

Wisniewski, Katrin et al. 2013. "MERLIN: An Online Trilingual Learner Corpus Empirically Grounding the European Reference Levels in Authentic Learner Data". In: *ICT for Language Learning 2013*. Florence, Italy.

Wolfe-Quintero, Kate, Inagaki, Shunji & Kim, Hae-Young. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center.

Wood, Simon N. 2003. Thin-plate regression splines. In: *Journal of the Royal Statistical Society (B)* 65 (1), pp. 95–114.

Wood, Simon N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. In: *Journal of the American Statistical Association* 99, pp. 673–686.

Wood, Simon N. 2006. *Generalized additive models: an introduction with R*. CRC press.

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. In: *Journal of the Royal Statistical Society* 72 (1), pp. 3–36.

Wood, Simon N., Pya, Natalya & Säfken, Benjamin. 2016. Smoothing Parameter and Model Selection for General Smooth Models. In: *Journal of the American Statistical Association* 111 (516), pp. 1548–1563.

Yoon, Hyung-Jo & Polio, Charlene. 2016. The Linguistic Development of Students of English as a Second Language in Two Written Genres. In: *Tesol Quarterly*.

Part IX.

Supplementary Material

A. Supplementary Material for Study on Merlin Data

```
gam.merlin.reference <- gam(OverallCefrScore ~
                          hasTransitionsFromSubjectToNot +
                          has3rdPersPossessivePronouns +
                          containsToInfinitives +
                          usesConjunctionalClauses +
                          halfModalClusterPerVP +
                          logSumNonTerminalNodesPerSentence +
                          logATFBand2PerTypesFoundInDlex +
                          avgVTotalIntegrationCostAtFiniteVerb +
                          lexTypesFoundInDlexPerLexType +
                          typeTokenRato +
                          s(charactersPerWord) +
                          s(sumNonTerminalNodesPerWord) +
                          s(numberOfSentencesSquared) +
                          TaskTheme,
                        data=merlin,
                        family=ocat(R=5))
```

(a) Reference model.

```
gam.merlin.complexity <- gam(OverallCefrScore ~
                               hasTransitionsFromSubjectToNot +
                               has3rdPersPossessivePronouns +
                               containsToInfinitives +
                               usesConjunctionalClauses +
                               halfModalClusterPerVP +
                               logSumNonTerminalNodesPerSentence +
                               logATFBand2PerTypesFoundInDlex +
                               avgVTotalIntegrationCostAtFiniteVerb +
                               lexTypesFoundInDlexPerLexType +
                               s(typeTokenRato) +
                               s(charactersPerWord) +
                               s(sumNonTerminalNodesPerWord) +
                               s(numberOfSentencesSquared),
                             data=merlin,
                             family=ocat(R=5))
```

(b) Complexity model.

Figure A.1.: Model formulas of *Merlin* reference and complexity model predicting overall CEFR scores from scaled and transformed complexity measures.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	8.0213	0.3139	25.5533	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.4047	0.2292	-1.7655	0.0775
has3rdPersPossessivePronouns[TRUE]	-0.8590	0.1958	-4.3863	< 0.0001
containsToInfinitives[TRUE]	-0.4944	0.2218	-2.2294	0.0258
usesConjunctionalClauses[TRUE]	-0.3950	0.2096	-1.8842	0.0595
halfModalClusterPerVP	0.1751	0.0982	1.7830	0.0746
logSumNonTerminalNodesPerSentence	1.8726	0.1688	11.0951	< 0.0001
logATFBand2PerTypesFoundInDlex	-0.1548	0.0886	-1.7464	0.0807
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1018	3.6404	0.0003
lexTypesFoundInDlexPerLexType	0.8600	0.0928	9.2695	< 0.0001
typeTokenRato	0.8820	0.1527	5.7748	< 0.0001
TaskTheme[Society]	1.5422	0.4329	3.5628	0.0004
TaskTheme[Profession]	1.0416	0.4014	2.5949	0.0095
TaskTheme[Smalltalk]	-0.6986	0.2181	-3.2026	0.0014
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord)	2.7841	3.5456	20.4366	0.0003
s(sumNonTerminalNodesPerWord)	1.7176	2.1855	37.4943	< 0.0001
s(numberOfSentencesSquared)	4.6002	5.6802	269.4891	< 0.0001

Table A.1.: Summary of *Merlin* reference model predicting overall CEFR scores from scaled and transformed complexity measures estimated on *Merlin* data without outliers ($N = 1,024$).

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	8.2553	0.2550	32.3702	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.5433	0.2259	-2.4051	0.0162
has3rdPersPossessivePronouns[TRUE]	-0.7403	0.1870	-3.9596	0.0001
containsToInfinitives[TRUE]	-0.5087	0.2178	-2.3356	0.0195
usesConjunctionalClauses[TRUE]	-0.3451	0.2102	-1.6420	0.1006
halfModalClusterPerVP	0.2182	0.0956	2.2830	0.0224
logATFBand2PerTypesFoundInDlex	-0.1203	0.0871	-1.3819	0.1670
avgVTotalIntegrationCostAtFiniteVerb	0.4112	0.0977	4.2065	< 0.0001
lexTypesFoundInDlexPerLexType	0.8591	0.0906	9.4837	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(logSumNonTerminalNodesPerSentence)	2.6186	3.3618	145.3155	< 0.0001
s(typeTokenRatos)	2.2110	2.8305	21.1931	0.0001
s(sumNonTerminalNodesPerWord)	1.7536	2.2256	34.5858	< 0.0001
s(charactersPerWord)	3.3771	4.2638	73.0793	< 0.0001
s(numberOfSentencesSquared)	4.6104	5.6851	305.2893	< 0.0001

Table A.2.: Summary of *Merlin* complexity model predicting overall CEFR scores from scaled and transformed complexity measures estimated on *Merlin* data without outliers ($N = 1,024$).

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	8.8680	0.6361	13.9421	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.5349	0.2387	-2.2408	0.0250
has3rdPersPossessivePronouns[TRUE]	-0.8906	0.2030	-4.3873	< 0.0001
containsToInfinitives[TRUE]	-0.5541	0.2282	-2.4284	0.0152
usesConjunctionalClauses[TRUE]	1.5788	0.9115	1.7322	0.0832
halfModalClusterPerVP	0.1831	0.1011	1.8113	0.0701
logSumNonTerminalNodesPerSentence	1.9714	0.1785	11.0435	< 0.0001
logATFBand2PerTypesFoundInDlex	-0.4830	0.4051	-1.1923	0.2331
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1059	3.4968	0.0005
lexTypesFoundInDlexPerLexType	0.8840	0.0942	9.3858	< 0.0001
typeTokenRato	0.8103	0.3181	2.5475	0.0108
logSumNonTerminalNodesPerWord	-1.6499	0.3970	-4.1559	< 0.0001
TaskTheme[Demand]	-0.4921	0.7085	-0.6947	0.4873
TaskTheme[Profession]	0.5853	0.6101	0.9593	0.3374
TaskTheme[Smalltalk]	-1.3039	0.6616	-1.9710	0.0487
usesConjunctionalClauses:TaskTheme[Demand]	-2.1839	0.9603	-2.2742	0.0230
usesConjunctionalClauses:TaskTheme[Profession]	-2.6025	1.0103	-2.5759	0.0100
usesConjunctionalClauses:TaskTheme[Smalltalk]	-1.6684	0.9694	-1.7211	0.0852
logATFBand2PerTypesFoundInDlex:TaskTheme[Demand]	0.1827	0.4194	0.4357	0.6631
logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]	0.7344	0.5233	1.4033	0.1605
logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]	0.7219	0.4477	1.6122	0.1069
typeTokenRato:TaskTheme[Demand]	0.4750	0.3634	1.3072	0.1912
typeTokenRato:TaskTheme[Profession]	-0.1225	0.4431	-0.2765	0.7822
typeTokenRato:TaskTheme[Smalltalk]	-0.3585	0.3846	-0.9321	0.3513
sumNonTerminalNodesPerWord:TaskTheme[Demand]	0.9369	0.4216	2.2224	0.0263
sumNonTerminalNodesPerWord:TaskTheme[Profession]	0.7847	0.4893	1.6037	0.1088
sumNonTerminalNodesPerWord:TaskTheme[Smalltalk]	1.2049	0.4284	2.8122	0.0049
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord)	2.7714	3.5484	18.5670	0.0007
s(numberOfSentencesSquared)	4.6262	5.7193	254.0399	< 0.0001

Table A.3.: Interaction model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses 'society' as reference level.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	9.4533	0.4789	19.7408	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.5349	0.2387	-2.2408	0.0250
has3rdPersPossessivePronouns[TRUE]	-0.8906	0.2030	-4.3873	< 0.0001
containsToInfinitives[TRUE]	-0.5541	0.2282	-2.4284	0.0152
usesConjunctionalClauses[TRUE]	-1.0237	0.4497	-2.2764	0.0228
halfModalClusterPerVP	0.1831	0.1011	1.8113	0.0701
logSumNonTerminalNodesPerSentence	1.9714	0.1785	11.0435	< 0.0001
logATFBand2PerTypesFoundInDlex	0.2513	0.3359	0.7482	0.4544
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1059	3.4968	0.0005
lexTypesFoundInDlexPerLexType	0.8840	0.0942	9.3858	< 0.0001
typeTokenRato	0.6877	0.3684	1.8668	0.0619
logSumNonTerminalNodesPerWord	-0.8652	0.3104	-2.7872	0.0053
TaskTheme[Smalltalk]	-1.8891	0.5035	-3.7518	0.0002
TaskTheme[Society]	-0.5853	0.6101	-0.9593	0.3374
TaskTheme[Demand]	-1.0774	0.5508	-1.9560	0.0505
usesConjunctionalClauses:TaskTheme[Smalltalk]	0.9340	0.5671	1.6469	0.0996
usesConjunctionalClauses:TaskTheme[Society]	2.6025	1.0103	2.5759	0.0100
usesConjunctionalClauses:TaskTheme[Demand]	0.4185	0.5417	0.7726	0.4398
logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]	-0.0125	0.3871	-0.0323	0.9742
logATFBand2PerTypesFoundInDlex:TaskTheme[Society]	-0.7344	0.5233	-1.4033	0.1605
logATFBand2PerTypesFoundInDlex:TaskTheme[Demand]	-0.5517	0.3530	-1.5628	0.1181
typeTokenRato:TaskTheme[Smalltalk]	-0.2360	0.4205	-0.5611	0.5747
typeTokenRato:TaskTheme[Society]	0.1225	0.4431	0.2765	0.7822
typeTokenRato:TaskTheme[Demand]	0.5975	0.3998	1.4947	0.1350
sumNonTerminalNodesPerWord:TaskTheme[Smalltalk]	0.4202	0.3525	1.1920	0.2333
sumNonTerminalNodesPerWord:TaskTheme[Society]	-0.7847	0.4893	-1.6037	0.1088
sumNonTerminalNodesPerWord:TaskTheme[Demand]	0.1522	0.3409	0.4465	0.6552
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord)	2.7714	3.5484	18.5670	0.0007
s(numberOfSentencesSquared)	4.6262	5.7193	254.0399	< 0.0001

Table A.4.: Full interaction model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses 'profession' as reference level.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	7.5642	0.3454	21.9003	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.5349	0.2387	-2.2408	0.0250
has3rdPersPossessivePronouns[TRUE]	-0.8906	0.2030	-4.3873	< 0.0001
containsToInfinitives[TRUE]	-0.5541	0.2282	-2.4284	0.0152
usesConjunctionalClauses[TRUE]	-0.0896	0.3622	-0.2474	0.8046
halfModalClusterPerVP	0.1831	0.1011	1.8113	0.0701
logSumNonTerminalNodesPerSentence	1.9714	0.1785	11.0435	< 0.0001
logATFBand2PerTypesFoundInDlex	0.2388	0.1933	1.2358	0.2165
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1059	3.4968	0.0005
lexTypesFoundInDlexPerLexType	0.8840	0.0942	9.3858	< 0.0001
typeTokenRato	0.4518	0.2569	1.7583	0.0787
logSumNonTerminalNodesPerWord	-0.4450	0.1876	-2.3722	0.0177
TaskTheme[Society]	1.3039	0.6616	1.9710	0.0487
TaskTheme[Demand]	0.8117	0.3529	2.3004	0.0214
TaskTheme[Profession]	1.8891	0.5035	3.7518	0.0002
usesConjunctionalClauses:TaskTheme[Society]	1.6684	0.9694	1.7211	0.0852
usesConjunctionalClauses:TaskTheme[Demand]	-0.5155	0.4714	-1.0937	0.2741
usesConjunctionalClauses:TaskTheme[Profession]	-0.9340	0.5671	-1.6469	0.0996
logATFBand2PerTypesFoundInDlex:TaskTheme[Society]	-0.7219	0.4477	-1.6122	0.1069
logATFBand2PerTypesFoundInDlex:TaskTheme[Demand]	-0.5392	0.2197	-2.4539	0.0141
logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]	0.0125	0.3871	0.0323	0.9742
typeTokenRato:TaskTheme[Society]	0.3585	0.3846	0.9321	0.3513
typeTokenRato:TaskTheme[Demand]	0.8335	0.2925	2.8494	0.0044
typeTokenRato:TaskTheme[Profession]	0.2360	0.4205	0.5611	0.5747
sumNonTerminalNodesPerWord:TaskTheme[Society]	-1.2049	0.4284	-2.8122	0.0049
sumNonTerminalNodesPerWord:TaskTheme[Demand]	-0.2680	0.2344	-1.1429	0.2531
sumNonTerminalNodesPerWord:TaskTheme[Profession]	-0.4202	0.3525	-1.1920	0.2333
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(charactersPerWord)	2.7714	3.5484	18.5670	0.0007
s(numberOfSentencesSquared)	4.6262	5.7193	254.0399	< 0.0001

Table A.5.: Full interaction model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses 'smalltalk' as reference level.

Model	AIC	REML	REML diff. (Ref.)	REML diff. (-4)
Full data	1327.07	652.34		
-4 worst	1281.00	628.84	-23.50	
-6 worst	1263.23	619.74	-32.60	-9.10

Table A.6.: Model comparison for *Merlin* interaction model when training on i) full data, ii) data without the four outliers, iii) data without the six data points with the highest residual errors.

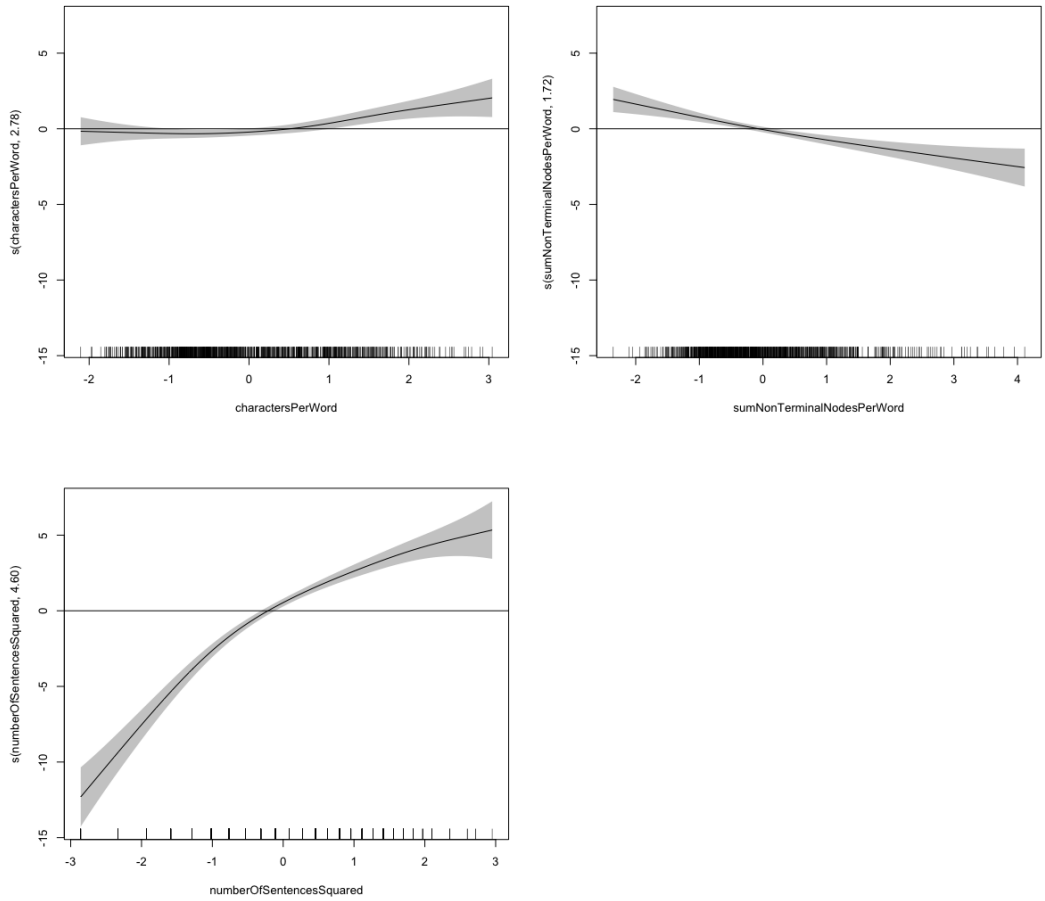


Figure A.2.: Smooths of *Merlin* reference model.

Model	AIC	REML	REML diff. (Ref.)	REML diff. (-4)
Full data	1333.26	662.78		
-4 worst	1287.08	642.77	-23.48	
-6 worst	1268.55	629.91	-32.87	-9.39

Table A.7.: Model comparison for *Merlin* reference model when training on i) full data, ii) data without the four outliers, iii) data without the six data points with the highest residual errors.

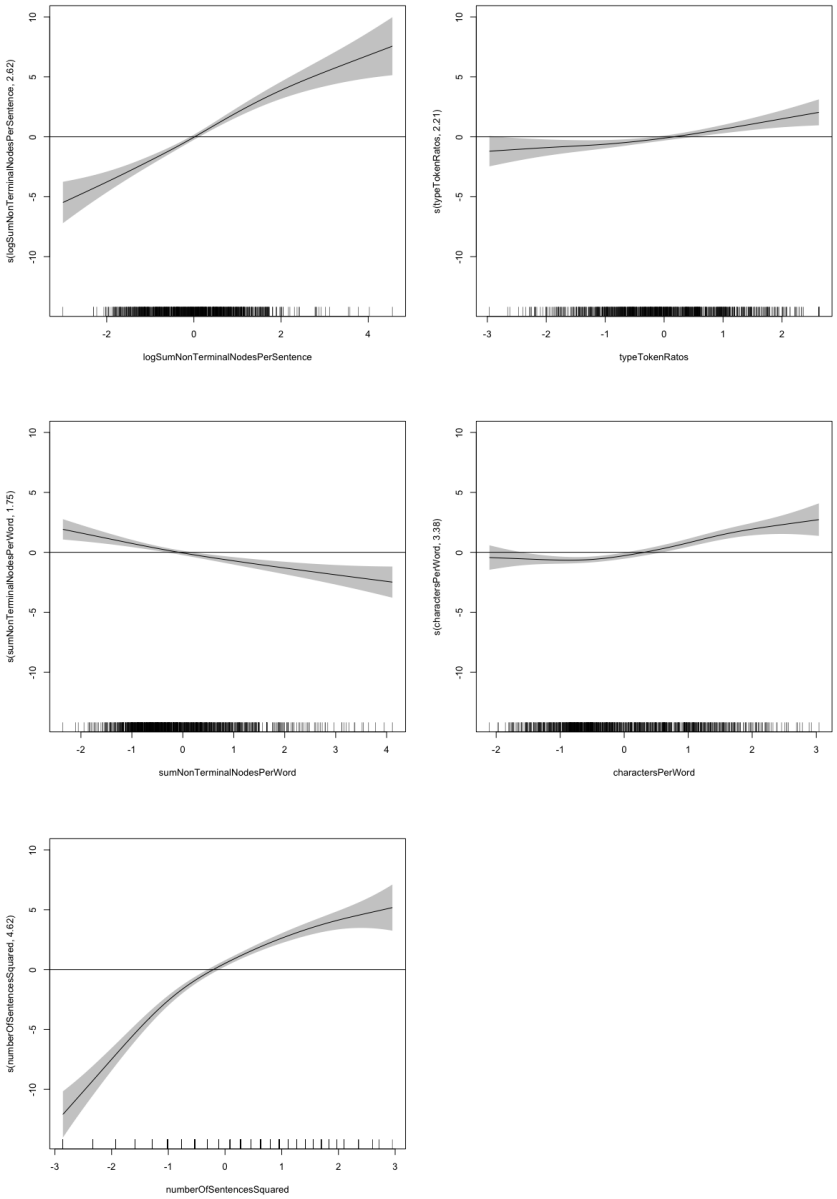
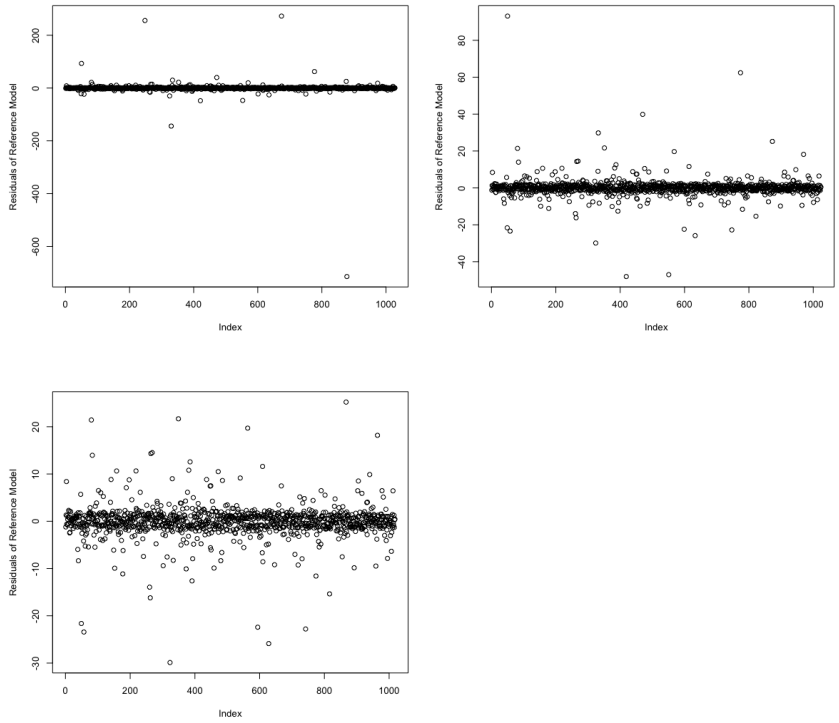
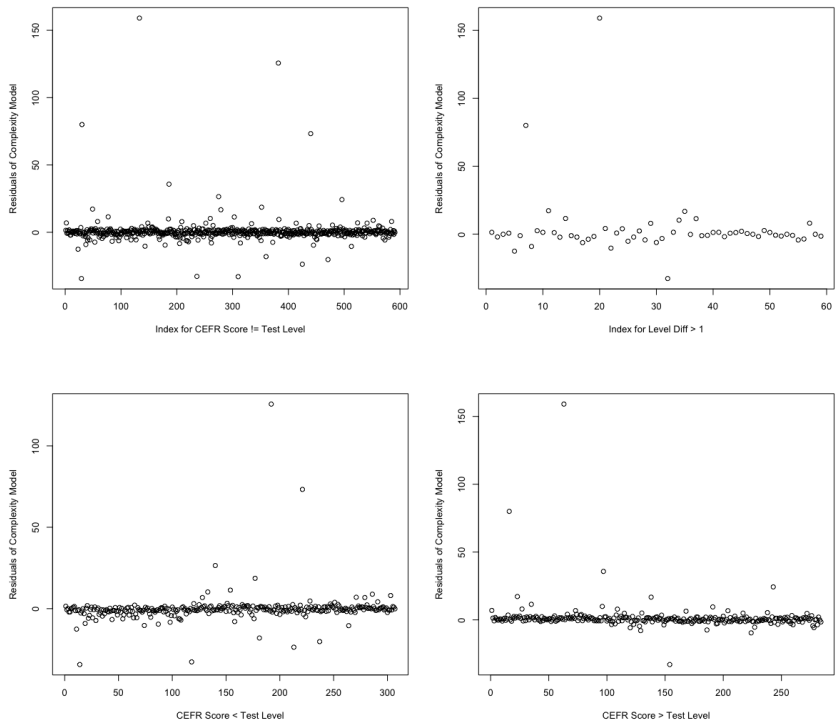


Figure A.3.: Smooths of *Merlin* complexity model.



(a) Reference model.



(b) Complexity model.

Figure A.4.: Residuals of *Merlin* reference and complexity models on i) full data (upper left); ii) data without 4 most severe outliers (upper right); iii) data without any outliers (lower left).

Model	AIC	REML	REML diff. (Ref.)	REML diff. (-4)
Full data	1359.80	681.26		
-4 worst	1315.05	658.56	22.70	
-6 worst	1300.52	651.25	30.01	7.31

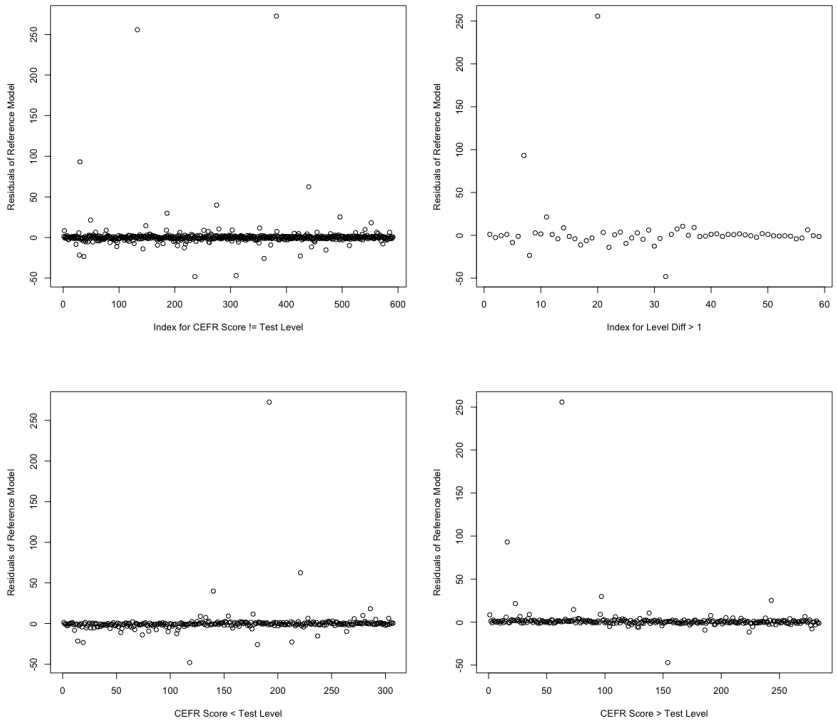
Table A.8.: Model comparison for *Merlin* complexity model when training on i) full data, ii) data without the four outliers, iii) data without the six data points with the highest residual errors.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-6.7060	1.8434	-3.6378	0.0003
numberOfWords	0.0648	0.0070	9.2459	$<$ 0.0001
typeTokenRato	5.8119	1.9158	3.0337	0.0024
TaskTheme[Society]	7.7312	3.0671	2.5207	0.0117
TaskTheme[Profession]	3.0859	3.8470	0.8022	0.4225
TaskTheme[Smalltalk]	3.6202	2.7037	1.3390	0.1806
typeTokenRato:TaskTheme[Society]	3.0952	3.5527	0.8712	0.3836
typeTokenRato:TaskTheme[Profession]	3.5934	4.3085	0.8340	0.4043
typeTokenRato:TaskTheme[Smalltalk]	-4.8271	2.9204	-1.6529	0.0984
numberOfWords:TaskTheme[Society]	-0.0467	0.0082	-5.7117	$<$ 0.0001
numberOfWords:TaskTheme[Profession]	-0.0305	0.0090	-3.3997	0.0007
numberOfWords:TaskTheme[Smalltalk]	-0.0066	0.0081	-0.8179	0.4134

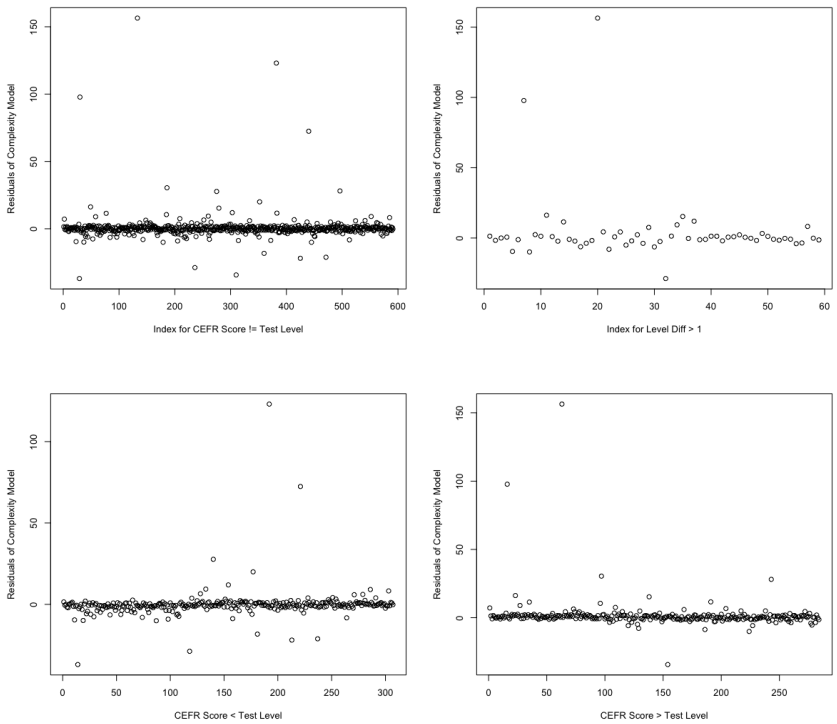
Table A.9.: Summary of model that inspects interaction of task theme with standard type token ratio and number of words as predictors on the *Merlin* data. Uses 'demand' as reference level.

Model	AIC	Df	REML	Edf	Compared with	χ^2	Edf difference	$Pr(>\chi^2)$
Only nWord	1620.32	12	801.15	9				
+TTR	1579.25	15	789.91	12	Only nWord	11.244	3	$5.163e-05$

Table A.10.: Model comparison for model with only a task theme interaction for number of words and the model including also an interaction with the standard type token ratio on the *Merlin* data.



(a) Reference model.



(b) Complexity model.

Figure A.5.: Residuals of *Merlin* reference and complexity models for test level and CEFR score mismatches: i) CEFR score and test level differ (upper left); ii) CEFR score and test level differ more than 1 level (upper right); iii) CEFR score lower than test level (lower left); iv) CEFR score higher than test level (lower right).

```
762 gam.merlin.success.extended <- gam(OverallCefrScore ~
763                                     hasTransitionsFromSubjectToNot +
764                                     has3rdPersPossessivePronouns +
765                                     containsToInfinitives +
766                                     usesConjunctionalClauses +
767                                     logSumNonTerminalNodesPerSentence +
768                                     logATFBand2PerTypesFoundInDlex +
769                                     avgVTotalIntegrationCostAtFiniteVerb +
770                                     lexTypesFoundInDlexPerLexType +
771                                     typeTokenRato +
772                                     logSumNonTerminalNodesPerWord +
773                                     halfModalClusterPerVP +
774                                     logATFBand2PerTypesFoundInDlex:TaskTheme +
775                                     usesConjunctionalClauses:TaskTheme +
776                                     logSumNonTerminalNodesPerWord:TaskTheme +
777                                     typeTokenRato:TaskTheme +
778                                     halfModalClusterPerVP:TaskTheme +
779                                     s(charactersPerWord, by = Passed, k = 6) +
780                                     s(numberOfSentencesSquared, by = Passed) +
781                                     Passed +
782                                     TaskTheme,
783                                   data=merlin,
784                                   family=ocat(R=n_cat))
```

Figure A.6.: Model formula of *Merlin* extenced success model predicting overall CEFR scores from scaled and transformed complexity measures.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	3.8610	0.5186	7.4455	< 0.0001
hasTransitionsFromSubjectToNot[TRUE]	-0.8565	0.2732	-3.1350	0.0017
has3rdPersPossessivePronouns[TRUE]	-1.3267	0.2556	-5.1903	< 0.0001
containsToInfinitives[TRUE]	-0.7246	0.2701	-2.6825	0.0073
usesConjunctionalClauses[TRUE]	-0.5514	0.2698	-2.0434	0.0410
logATFBand2PerTypesFoundInDlex	-0.3972	0.1202	-3.3054	0.0009
avgVTotalIntegrationCostAtFiniteVerb	0.4765	0.1380	3.4522	0.0006
lexTypesFoundInDlexPerLexType	0.9649	0.1132	8.5218	< 0.0001
typeTokenRato	1.2797	0.1877	6.8176	< 0.0001
sumNonTerminalNodesPerWord	-0.8316	0.1398	-5.9464	< 0.0001
logSumNonTerminalNodesPerSentence	2.4829	0.2093	11.8655	< 0.0001
Passed[TRUE]	6.6843	0.3018	22.1510	< 0.0001
TaskTheme[Society]	11.5649	0.6668	17.3437	< 0.0001
TaskTheme[Profession]	7.2479	0.5982	12.1158	< 0.0001
TaskTheme[Smalltalk]	0.9101	0.2796	3.2550	0.0011
logATFBand2PerTypesFoundInDlex:TaskTheme[Society]	0.4881	0.5980	0.8163	0.4144
logATFBand2PerTypesFoundInDlex:TaskTheme[Profession]	1.1930	0.4812	2.4795	0.0132
logATFBand2PerTypesFoundInDlex:TaskTheme[Smalltalk]	0.3248	0.2428	1.3376	0.1810
s(charactersPerWord):Passed[FALSE]	2.5964	3.2239	5.3214	0.1503
s(charactersPerWord):Passed[TRUE]	1.3498	1.6284	6.5613	0.0297
s(numberOfSentencesSquared):Passed[FALSE]	3.6322	4.5433	69.6021	< 0.0001
s(numberOfSentencesSquared):Passed[TRUE]	4.3517	5.3657	306.2779	< 0.0001

Table A.11.: Summary of success model predicting Merlin overall CEFR scores from scaled and transformed complexity measures in Merlin. Uses 'demand' as reference level.

B. Supplementary Material for Study on Falko Georgetown Data

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	40.6731	1.3562	29.9894	< 0.0001
codeComplexity[HIGH]	-5.1897	2.0234	-2.5648	0.0103
ungDerivationPerTokenSquared[LOW]	2.0616	1.1848	1.7401	0.0818
ungDerivationPerTokenSquared[HIGH]	12.0748	0.7673	15.7368	< 0.0001
logLexTypesNotFoundInKCTPerLexType[LOW]	2.9956	1.4298	2.0952	0.0362
logLexTypesNotFoundInKCTPerLexType[HIGH]	22.6254	1.9759	11.4505	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(3rdPersPossessivePronounsPerToken):codeComplexity[LOW]	1.3283	1.5699	0.4051	0.6550
s(3rdPersPossessivePronounsPerToken):codeComplexity[HIGH]	2.3056	2.6049	106.1385	< 0.0001
s(wordsPerClause)	2.7580	2.9459	74.5939	< 0.0001

Table B.1.: Summary of *Falko* GAM with code complexity interactions. Predicts course level from scaled and transformed complexity measures estimated on longitudinal *Falko Georgetown L2* data.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	2.9803	0.3418	8.7190	< 0.0001
codeComplexity[HIGH]	-0.2110	0.4458	-0.4732	0.6361
s(ungDerivationPerTokenSquared):codeComplexity[LOW]	1.0000	1.0001	1.6442	0.1998
s(ungDerivationPerTokenSquared):codeComplexity[HIGH]	2.0260	2.4917	38.3181	< 0.0001
s(logLexTypesNotFoundInKCTPerLexType):codeComplexity[LOW]	1.1954	1.3639	3.1236	0.0825
s(logLexTypesNotFoundInKCTPerLexType):codeComplexity[HIGH]	3.1145	3.8709	22.8255	0.0001
s(3rdPersPossessivePronounsPerToken):codeComplexity[LOW]	1.0000	1.0001	0.4400	0.5072
s(3rdPersPossessivePronounsPerToken):codeComplexity[HIGH]	1.3764	1.6588	26.8120	< 0.0001
s(wordsPerClause)	1.9340	2.4769	31.9694	< 0.0001

Table B.2.: Summary of *Falko* GAM trained on full data set including all three code complexity interactions that were significant on longitudinal *Falko Georgetown L2* data. Note that on the full data set only 3rdPersPossessivePronounsPerToken interaction is significantly improving model fit.

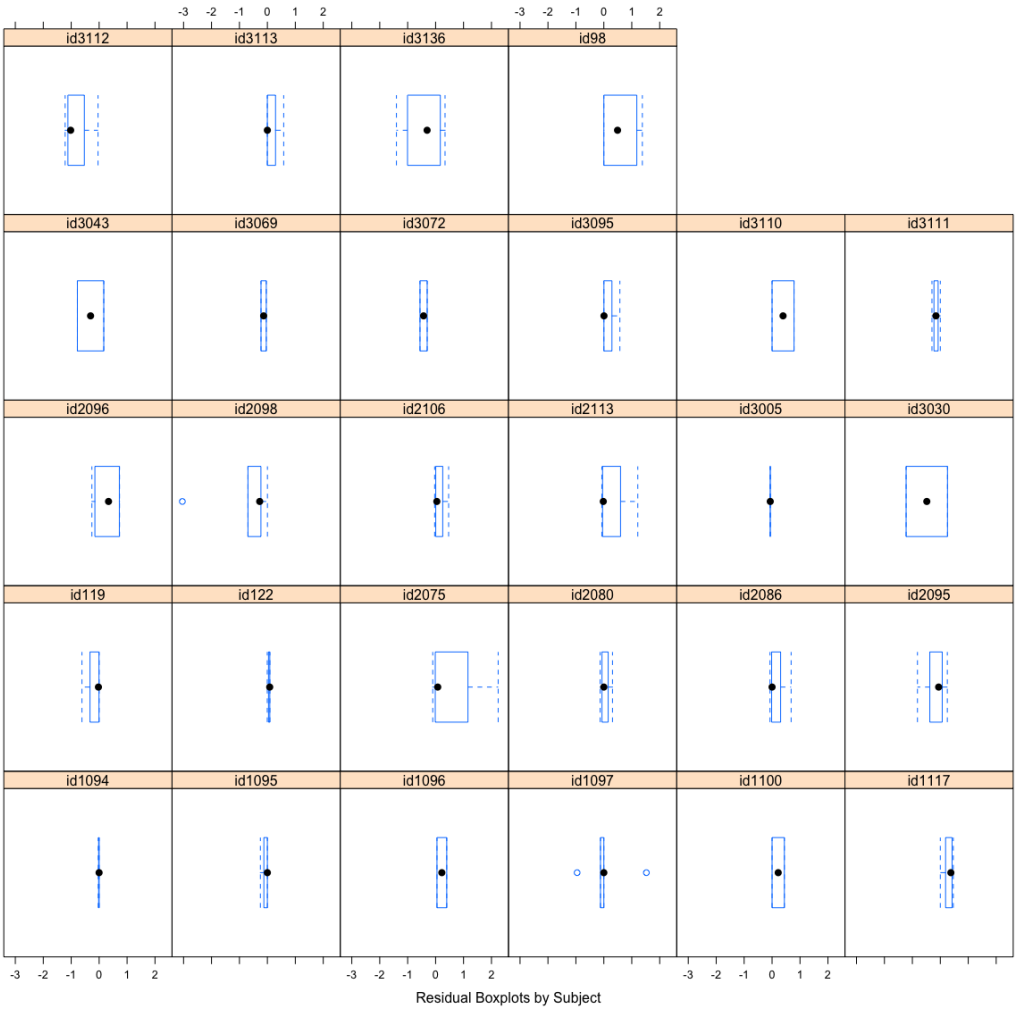


Figure B.1.: By-subject residuals of GAM with code complexity interactions fitted on longitudinal *Falko Georgetown L2* data.

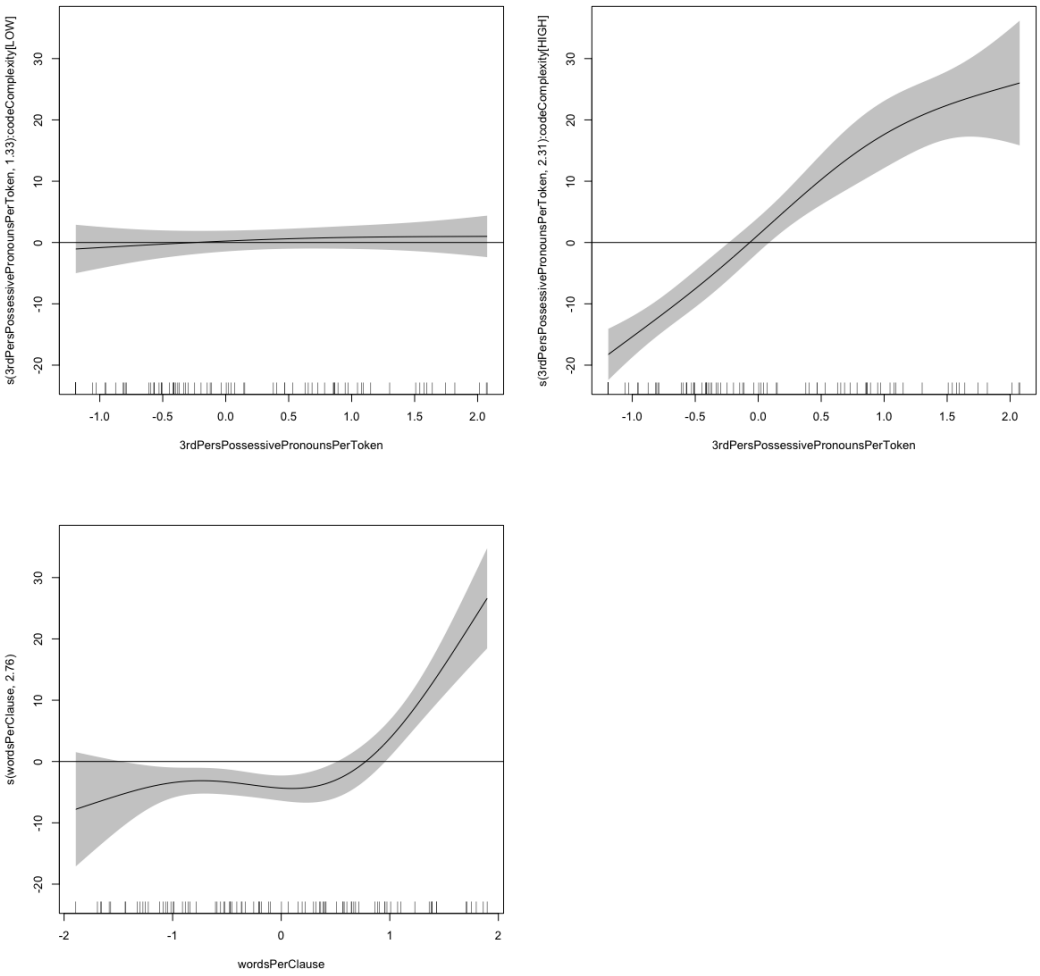


Figure B.2.: Smooths of GAM with three code complexity interactions fitted on longitudinal *Falko Georgetown L2* data.