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

Zarah Weiß
Eberhard Karls Universität Tübingen

Kolloquium Korpuslinguistik und Phonetik (SS17)
Humboldt Universität zu Berlin

June 7th, 2017

Table of Contents

Introduction

Overview

- Complexity analysis productive approach in SLA - CL intersection
- Automatic assessment of proficiency, readability, essay scoring, etc.
- Started in 1930s with superficial measures (text length in words, word length in characters) (Frogner 1933; Thorndike 1921)
- Development towards more sophisticated easily obtainable feature sets with progress in NLP

→ Increasing number of diverse complexity measures

→ Increasing availability of text analysis systems facilitating complexity analysis

→ Lack of consistency in findings and interpretation of measures

Introduction

Criticism

- Lack of theoretical foundation of feature implementation, selection, and interpretation (Bulté and Housen 2014; Housen, Vedder, and Kuiken 2012; Pallotti 2009)
- Assumption of homogeneous learner profiles (Crossley and McNamara 2011; Jarvis et al. 2003)
- Disregard for task differences and task effects (Alexopoulou et al. 2017; Tracy-Ventura and Myles 2015)

Solutions

- Modeling of heterogeneous learner profiles; e.g. L1 backgrounds (Crossley and McNamara 2011), learning strategies (Jarvis et al. 2003)
- Modeling of task effects (Alexopoulou et al. 2017; Polio and Park 2016; Tracy-Ventura and Myles 2015)

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?

Procedure

- Selection of 2 German learner corpora with diverse task backgrounds: Merlin (Abel et al. 2013), Falko Georgetown (*Falko Georgetown Dokumentation* 2007)
- Automatic extraction of 398 measures of complexity
- Extract data from close transcription of learner data
- Manual extraction of cognitive and functional task factors from task descriptions provided in corpus documentations
- Descriptive cross-corpus analysis for +100 measures
- Inferential GAM regression analysis with task interactions for data-driven feature selection
- Within-corpus classification experiment on Merlin

Complexity in SLA

Overview

- Assessment of language performance to measure
  - Text readability
  - L2 and L1 proficiency
  - L2 and L1 development
  - Writing performance
  - Task performance
  - ...
- Dimensions of language performance
  1. Complexity
  2. Accuracy
  3. Fluency

→ CAF

Complexity in SLA

CAF Definitions

Fluency: native-like production speed (Pallotti 2009; Wolfe-Quintero, Inagaki, and Kim 1998)

- e.g. production rate units in time (if available); sometimes: frequency or length of sentences, t-units, utterances, clauses, and phrases

Accuracy: native-like production error rate (Housen, Vedder, and Kuiken 2012; Wolfe-Quintero, Inagaki, and Kim 1998)

- e.g. error-free t-units, overall error count

Complexity: elaborateness, variedness, and inter-relatedness of a system (Ellis and Barkhuizen 2005; Rescher 1998)

- e.g. type token ratio, word frequency, etc. for lexical complexity
- e.g. modifiers per phrase, words per clause, clauses per t-unit, etc. for syntactic complexity

Complexity in SLA

CAF Criticism

Vagueness of definitions used in empirical studies

- Does *sentence length* measure fluency or complexity? (Wolfe-Quintero, Inagaki, and Kim 1998)
- What is more complex: *adäquat* or *angemessen*?

Lack of explicit norm in complexity definition

- Is the observed complexity *adequate*? (Pallotti and Ferrari 2008; Pallotti 2015)
- High complexity in simple tasks indicates lack of socio-linguistic awareness Ortega 2003; Pallotti 2015

Complexity in SLA

CAF Criticism

- Lack of theoretical underpinnings for measures
- Complexity $\neq$ continuous, temporal progression
- Complexity $\neq$ competence
- Complexity $\neq$ difficulty

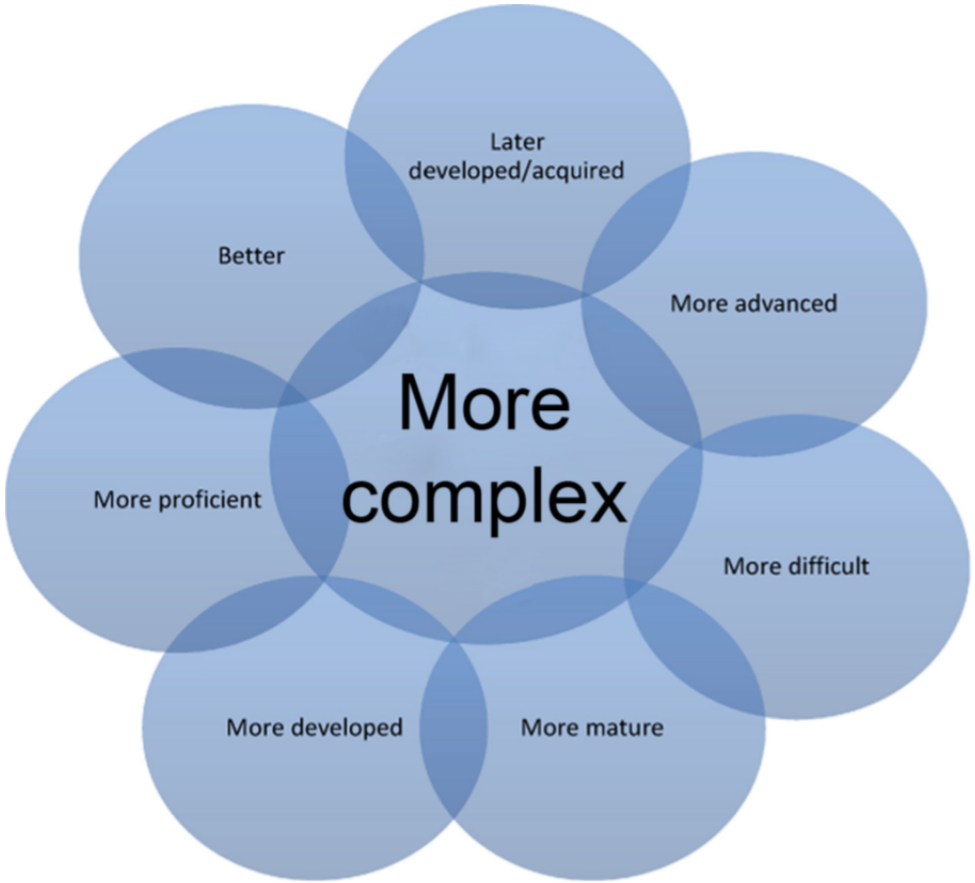

Figure: L2 complexity and related constructs (Bulté and Housen 2014)

Taxonomy

Overview

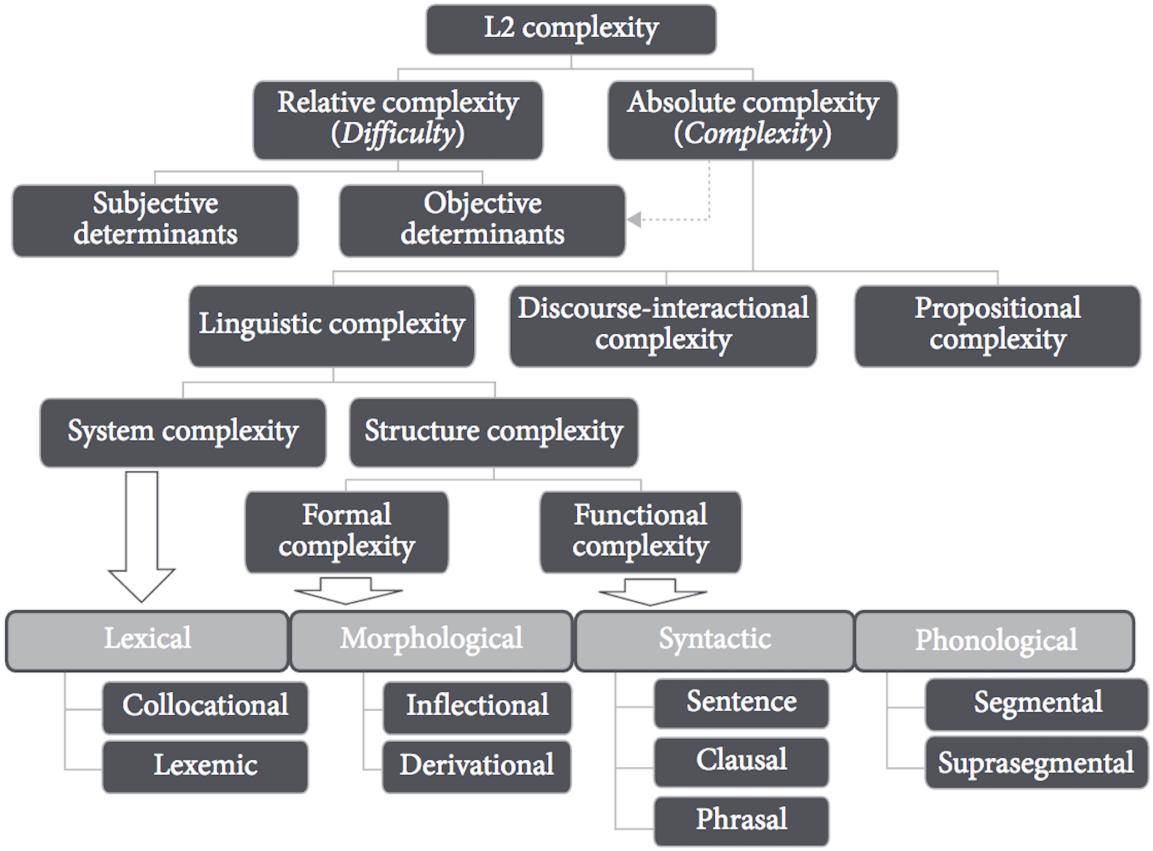

Figure: Taxonomy by Housen, Vedder, and Kuiken 2012

Taxonomy

Difficulty vs. Complexity

- Complexity:
  general validity across groups
- Difficulty:
  valid for specific groups (ability dependent)
- Cf. task complexity vs. task difficulty by Robinson 2001

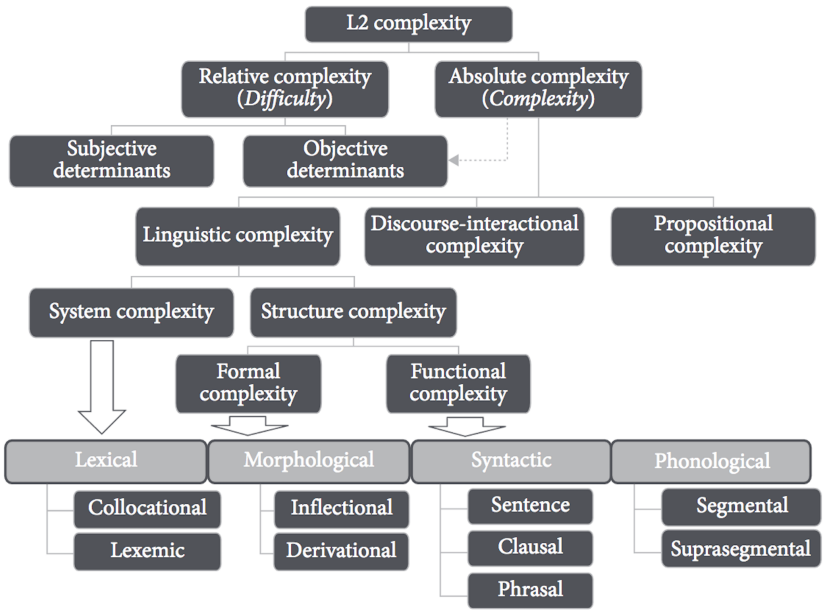

Figure: Taxonomy by ibid.

Taxonomy

System vs. Structure

(1) The small, happy bear from Peru ate all the orange marmalade.

- Structure complexity: prenominal modifiers, postnominal modifiers
- System complexity: NP length, number of modifiers

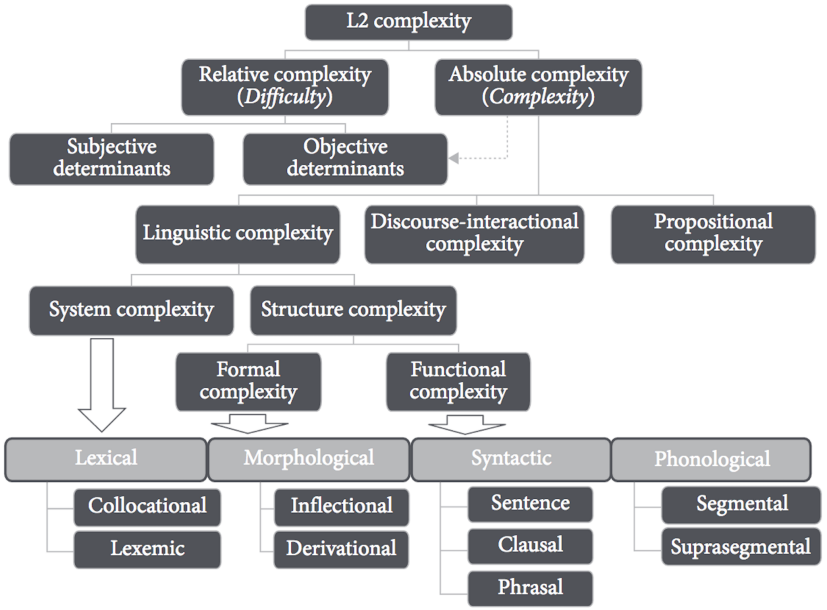

Figure: Taxonomy by Housen, Vedder, and Kuiken 2012

Taxonomy

Formal vs. Functional

(2) She carrie-$\mathbf{s}_1$ her dog-$\mathbf{s}_2$

- 3rd P. Sg. Pres. (s_1) vs. plural marker (s_2)
- Same formal complexity
- s_1 functionally more complex

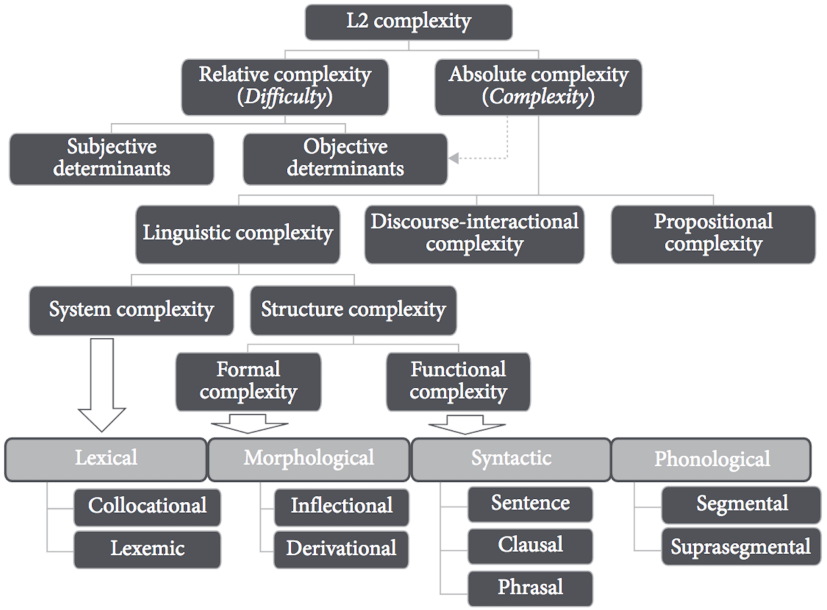

Figure: Taxonomy by Housen, Vedder, and Kuiken 2012

Taxonomy

Benefits

- Taxonomy is not exhaustive
- Helps to clarify how measures relate to complexity
- Helps to interpret findings
- Helps to assess coverage of measures included in studies

Overview

- In task-based language teaching CAF as dimensions of task performance
- Task factors known to influence CAF measures from research on TBLT (Polio and Park 2016; Robinson 2001; Skehan 1996)
- Task effects rarely analyzed in corpus-based studies (Tracy-Ventura and Myles 2015)
- Functional task factors vs. cognitive task factors

→ Cognitive task factors assumed to influence attentional resources dedicated to CAF

→ Functional task factors assumed to functionally require or inhibit aspects of CAF

Cognitive Task Factors

Skehan's Limited Attentional Capacity Model

- Differentiates
  1. Code complexity
  2. Cognitive complexity (processing cost, familiarity)
  3. Communicative stress (mode, time, stakes, etc.)

→ Deplete attentional resources from CAF

Robinson's Cognition Hypothesis

- Differentiates
  1. Task complexity
  2. Task conditions
  3. Task difficulty

→ more elements, tempo-spatial disclocation, reasoning demands **direct** resources to complexity and accuracy

→ more planning time, num. tasks, prior knowledge **depletes** resources from complexity and accuracy

Functional Task Factors

- LACM and CH make contradicting predictions about task effects on CAF
- Findings are inconsistent (Tracy-Ventura and Myles 2015; Yoon and Polio 2016)
- Functional task effects argued to be stronger (Biber, Gray, and Poonpon 2011; Yoon and Polio 2016)
- Assumed to functionally require or inhibit aspects of CAF
- Examples: discourse type, text genre, topic
- Most often comparison of argumentative, narrative, and descriptive texts

Merlin

In a Nutshell

- Cross-sectional corpus of 1,033 German L2 writings
- Elicited in official standardized language certification tests by Abel et al. 2013
- Test levels ranging from A1 to C1
- Proficiency scores assigned by 2 expert raters ranging from A1 to C2

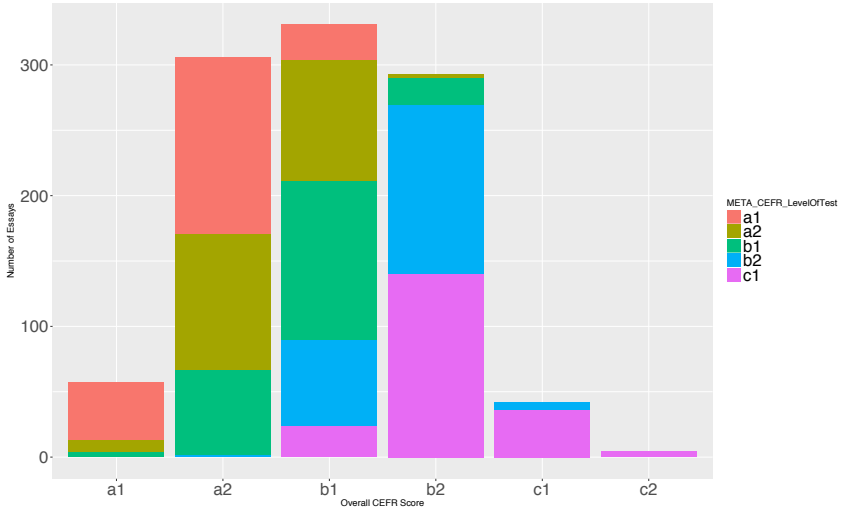

Figure: Distribution of proficiency scores grouped by test level.

Task	Test	$\sum$	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: Mapping of tasks to test levels, task frequency, and their distribution across overall proficiency scores (A1 to C2).

Falko Georgetown

In a Nutshell

- Partially longitudinal corpus of 209 German L2 writings by 123 students
- Elicited in curricular writing courses at Georgetown University by *Falko Georgetown Dokumentation* 2007; Reznicek et al. 2007
- Course levels 1 to 4 for intermediate to advanced learners of German
- No external validation of proficiency besides course levels

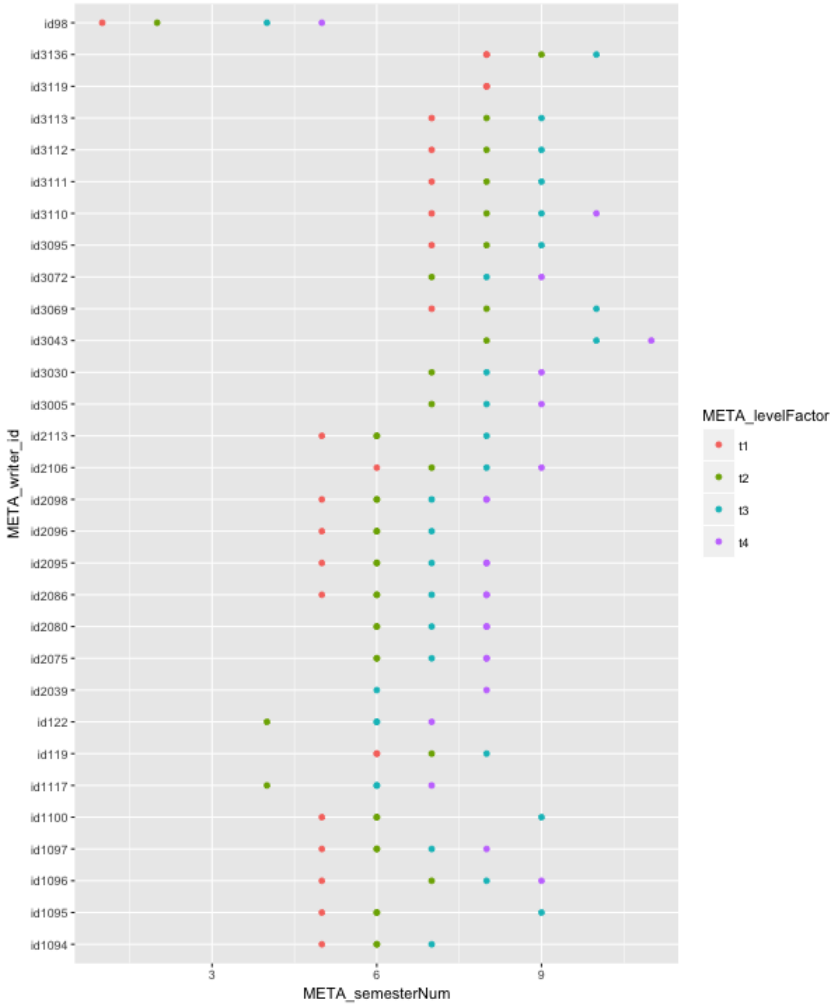

Figure: Texts written by learners who

Falko Georgetown

Tasks

Task	$\sum$	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: Frequency of tasks across course levels in the *Falko Georgetown L2* corpus.

Overview

- Annotate 15 cognitive and functional task factors
- Goal: disentangle correlation of course / test level and task
- Based on task descriptions in supplementary material *Falko Georgetown Dokumentation* 2007; Merlin project 2014a,b,c,d,e,f,g,h,i,j,k,l,m,n,o
- Follow approach by Alexopoulou et al. 2017

Operationalizations

Cognitive Factors

Code complexity: instructions provided no, few, or detailed language material to draw from

Cognitive complexity: require reasoning about writing structure vs. outline or refer to known structure

Shared context temporal and spatial dislocation (here/there; now/then)

Reasoning demands: quantity and elaborateness of spatial reasoning, i.e. referencing a location without extra-linguistic support, and reasoning about other people's intentions, beliefs, desires, or relationships

Referenced elements: number of discourse referents minimally required in solution

Perspective requires perspective of i) self, ii) psomeone else; iii) multiple other people.

Operationalizations

Functional Factors

Genre	text category; cf. task descriptions
Audience	recipient; cf. task descriptions (partially grouped)
Formality	tone; cf. task descriptions or inferred from genre and audience
Task theme	general topic; professional/occupational interests, public social affairs, small talk, or (by extension) goal-oriented personal matters (*demand*)
Task type	determined by a combination of functional needs and genre; argumentative, narrative, descriptive/expositional, and instructional

Tasks in Merlin

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

Tasks in Falko Georgetown

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

Overview

- 398 measures of elaborateness and variedness of various domains
- Extracted automatically using elaborate NLP tool chain
- Written by Galasso 2014; Hancke 2013; Weiß 2015, 2017
- Domains:
  1. Language use
  2. Human language processing
  3. Discourse & encoding of meaning
  4. Theoretical linguistics (lexico-semantics, syntax, morphology)

Pipeline

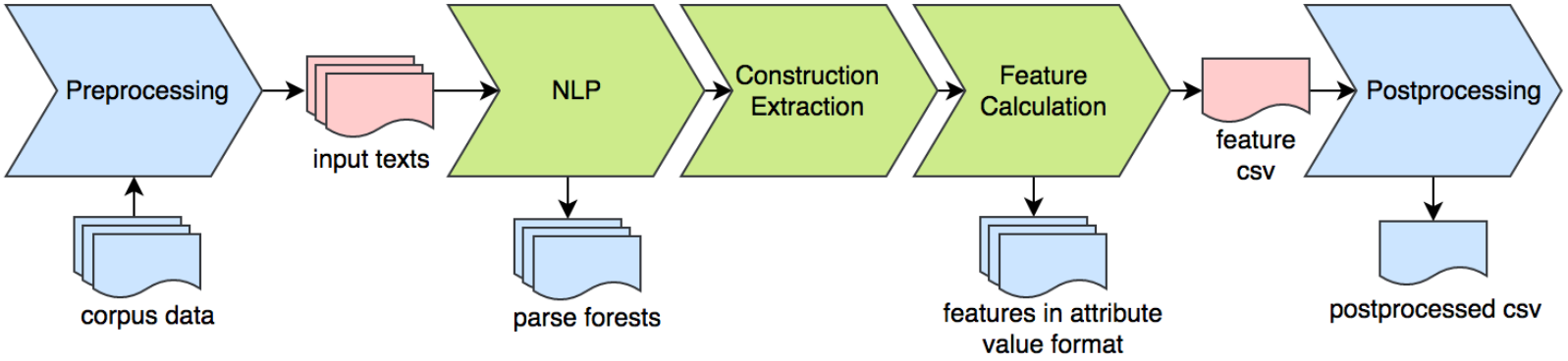

Figure: System pipeline from plain text corpus to feature analysis.

Resources

Task	Component	Version	Model
Tokenization and sentence segmentation	OpenNLP	1.6.0	default
POS tagging	Mate tools	3.6.0	default
Lemmatization	Mate tools	3.6.0	default
Morphological analysis	Mate tools	3.6.0	default
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: NLP components used in the complexity analysis pipeline.

Language Use

Domains:

- Corpus and psycho-linguistics

Measures:

- Word frequencies: less frequent = more sophisticated/complex
- Age of acquisition: later AoA = more sophisticated/complex

Implemented:

- Frequency data bases: dlexDB, SUBTLEX-DE, Google Books 2000
- AoA approximation based on KCT (Lavalley, Berkling, and Stüker 2015)

Human Language Processing

Domains:

- Cognitive science, psycho-linguistics and information theory

Measures:

- Cognitive processing costs as identified by processing time, reading time, etc.
- Storage and integration of discourse referents consumes cognitive resources
- Long distances between referents increase these costs

Implemented:

- Dependency Locality Theory (DLT) by Gibson 2000; Shain et al. 2016
- Verb argument distances in syllables (Weiß 2015)

Discourse & Encoding of Meaning

Domains:

- Psychology, psycho-linguistics

Measures:

- Propositional idea density (Brown et al. 2008): more propositions = more complex encoding of meaning
- Connectives
- Co-referential expressions
- Grammatical transitions

→ cause more cohesive writing and complex discourse

Implemented:

- PID (Louwerse et al. 2004)
- Connectives as listed by Duden (Gr) 2009
- Local and global overlap of linguistic material
- Co-referential expressions (pronouns, articles, etc.)
- Local transitions of grammatical roles (Barzilay and Lapata 2008; Galasso 2014; Todirascu et al. 2013)

Theoretical Linguistics

Lexio-Semantic

- Measures: concreteness, relatedness, diversity, and variation
- Implemented: TTR, lexical TTR, GermaNet semantic relations

Syntax:

- Measures: clausal complexity (subordination), phrasal complexity (modification)
- Implemented: dependent clause ratios, modifier ratios, complex NP, periphrastic constructions, etc.

Morphology:

- Measures: inflection, derivation, composition
- Implemented: nominalization, tense, compound depth, etc.

Overview

- Plots of measures with 95% confidence intervals
- Compare proficiency trajectory across corpora and task profiles
- Sample of over 100 complexity measures
- Grouped under theoretical considerations into concepts
- Selected to represent at least one concept per domain
- All 398 measures at `http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/supplementary-material/complexity-plots/`

Human Language Processing

DLT-V and syllable distance measures

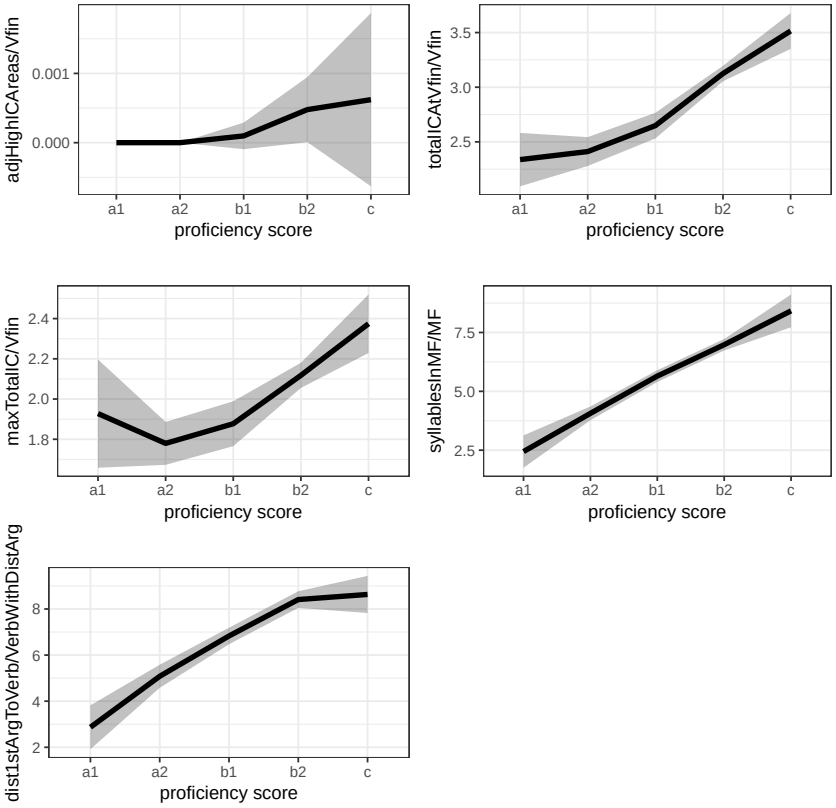

Figure: *Merlin*.

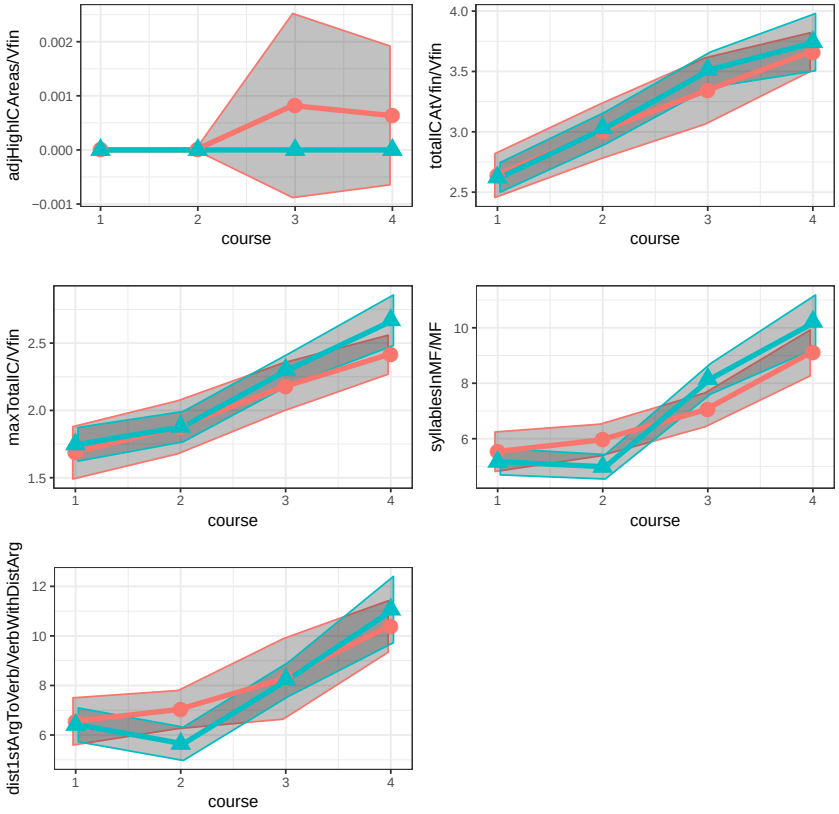

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Discourse & Encoding of Meaning

Overlap of linguistic material

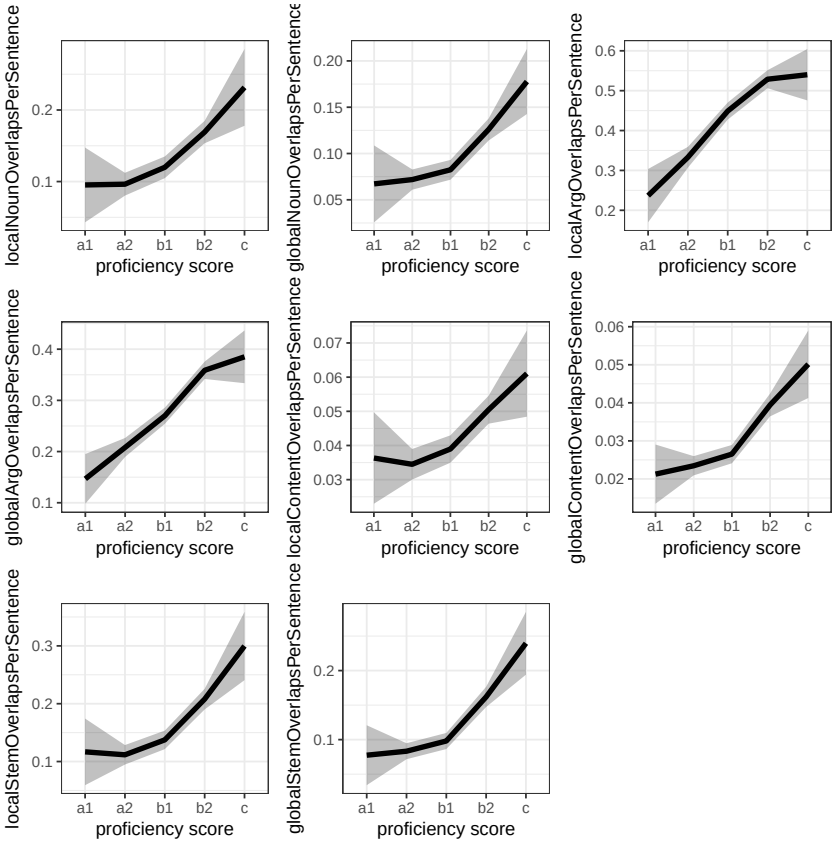

Figure: *Merlin*.

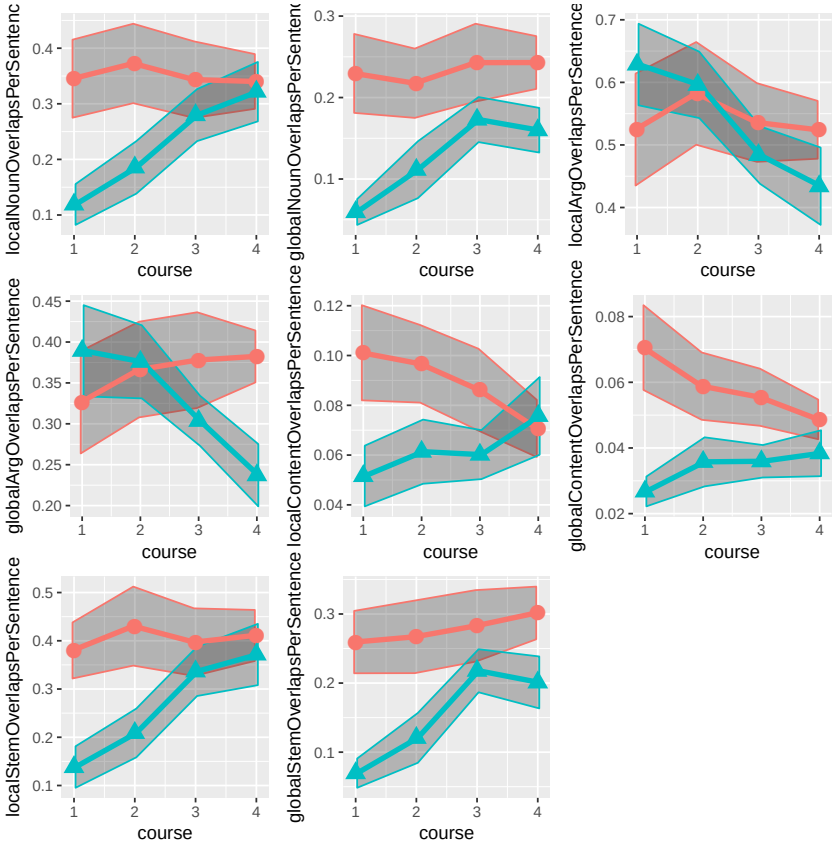

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Syntactic Complexity

Complex NPs

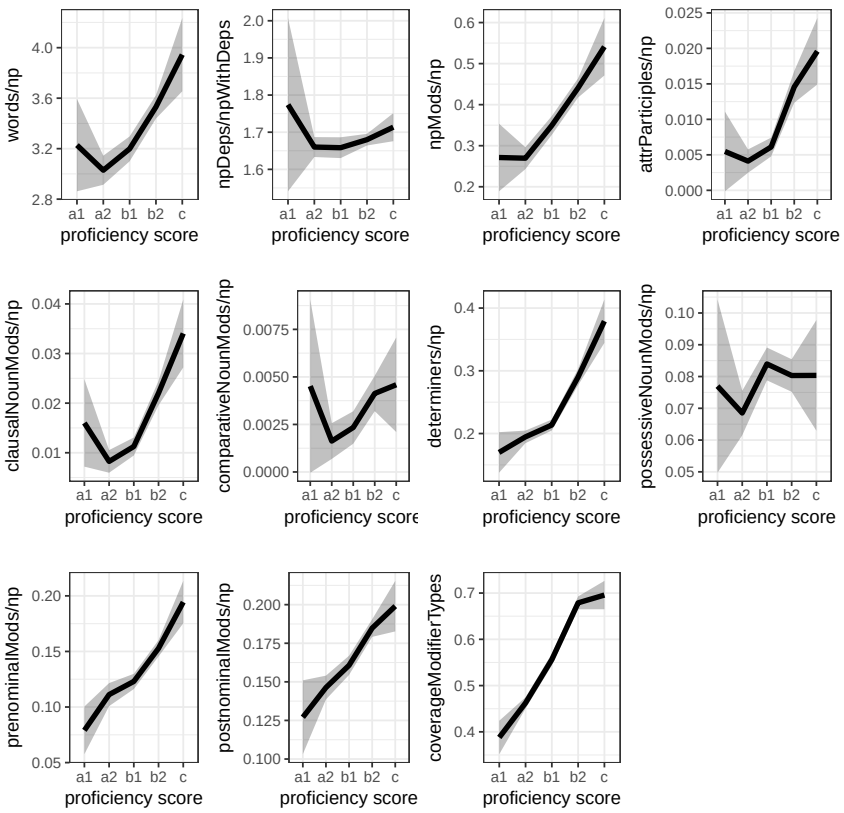

Figure: *Merlin*.

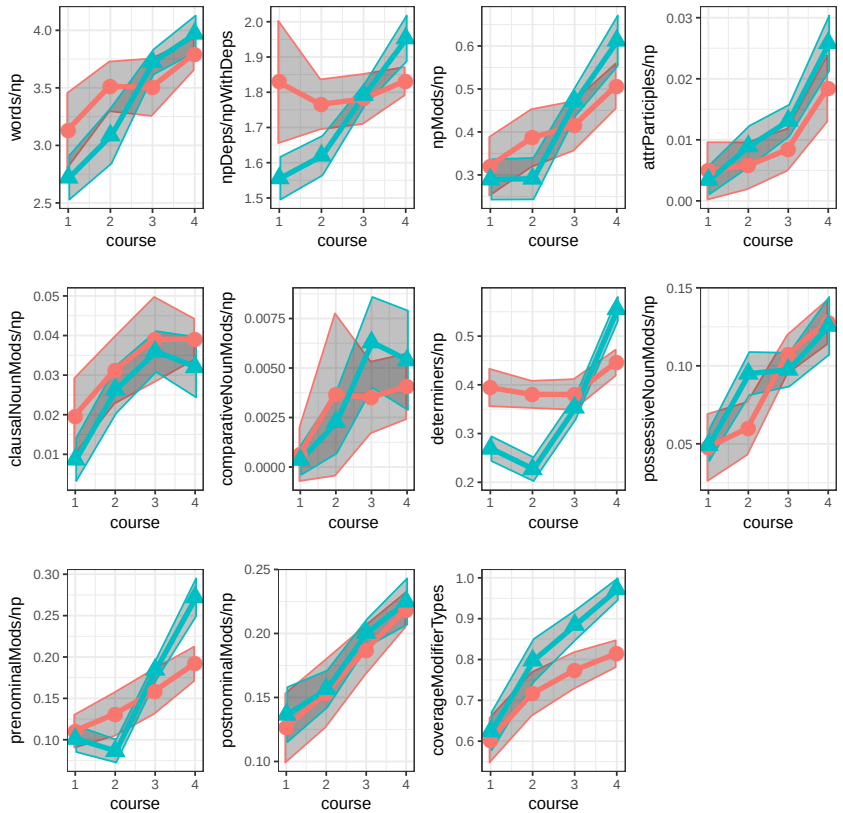

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Morphological Complexity

Inflection measures

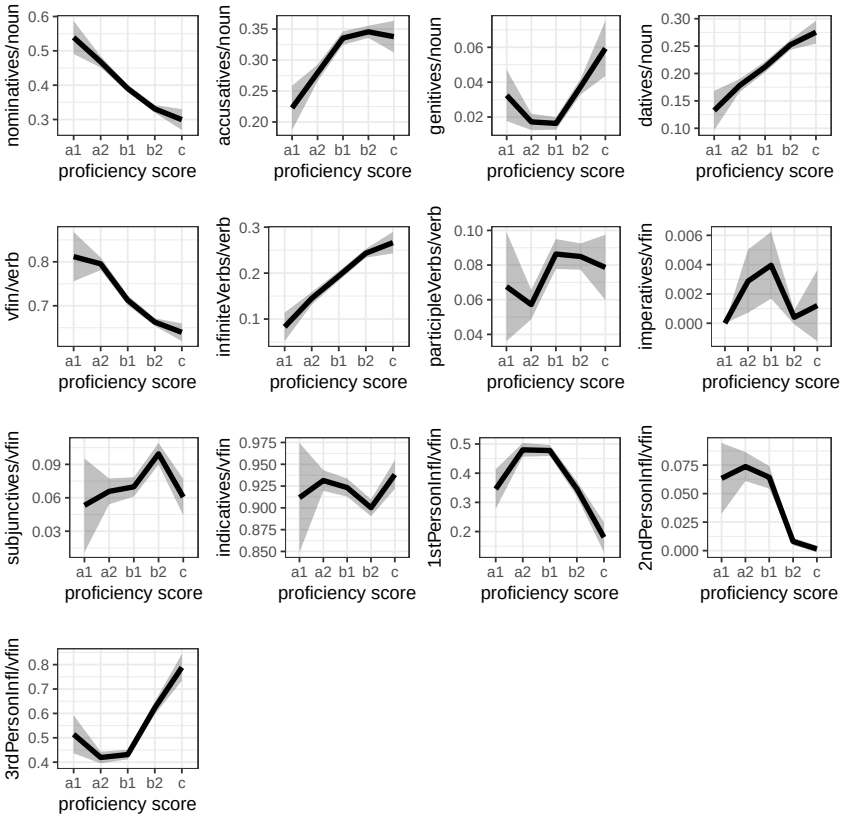

Figure: *Merlin*.

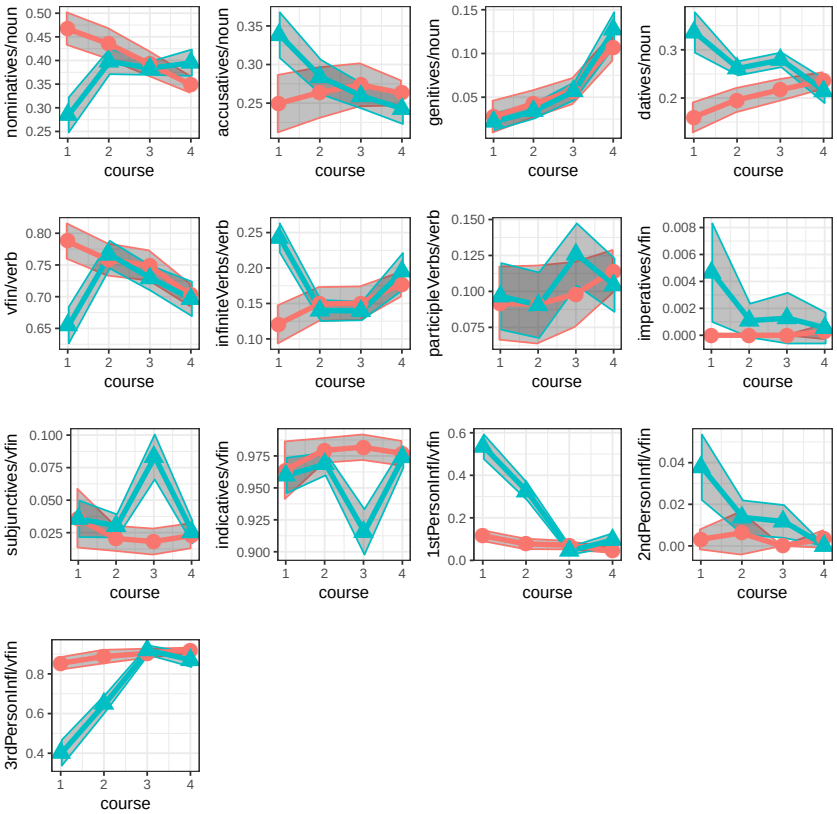

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Overview

- Regression analysis to predict proficiency from complexity and task factors
- Use ordinal generative additive regression models (GAMs)
- 2 studies on Merlin: i) task effects; ii) performance effects
- Studies on Falko Georgetown not reported here

Ordinal Generative Additive Regression Models

Overview

GAMs

- Extension of linear regression models
- Use splines as smooths for controlled introduction of non-linear relations
- Highly interpretable, yet similar predictive power as ML techniques like SVM
- Share requirements of regression models: normal, uncorrelated predictors
- Support 1 predictor per 15 to 20 data points

Ordinal Regression

- Link function to non-exponential distribution by Wood 2006
- Estimates boundaries between classes

→ keeps precedence without introducing quantity

Model Design

Iterative, data-driven model approach

1. Rank measures by information gain using WEKA
2. Test most informative measure for normality; normalize if necessary
3. Test for correlation of predictors
   a. If $< \pm 0.70$ Pearson correlation: add measure to model
   b. Else: remove correlated measures, add measure to model
4. Smooth measures unless they are linear
5. If changes lead to significant model improvement (χ^2 test), keep them
6. Do until 20 iterations did not yield better model or model contains 15/n measures

Study 1: Task Effects

Set Up

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

Study 1: Task Effects

Model Fit

- $R^2 = 0.7660$
- Approximately homescedastic residual errors with $\mu = 0.04; sd = 7.26$ after outlier removal
- Severe outliers across all model variants: Assumption of idiosyncratic properties in texts
- Outliers systematically include learners who performed above or below test level

$\rightarrow$ Prompted performance effect analysis in Study 2

Study 1: Task Effects

Model Fit

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

Study 1: Task Effects

Model Discussion

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
halfModalClusterPerVP	0.1831	0.1011	1.8113	0.0701
logSumNonTerminalNodesPerSentence	1.9714	0.1785	11.0435	< 0.0001
avgVTotalIntegrationCostAtFiniteVerb	0.3705	0.1059	3.4968	0.0005
lexTypesFoundInDlexPerLexType	0.8840	0.0942	9.3858	< 0.0001

Table: Interaction model: linear measures.

Study 1: Task Effects

Model Discussion

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	8.3759	0.3833	21.8509	< 0.0001
usesConjunctionalClauses[TRUE]	-0.6051	0.3173	-1.9074	0.0565
logATFBand2PerTypesFoundInDlex	-0.3003	0.1091	-2.7528	0.0059
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

Table: Interaction model: interactions measures.

Study 1: Task Effects

Model Discussion

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: Interaction model: smoothed measures.

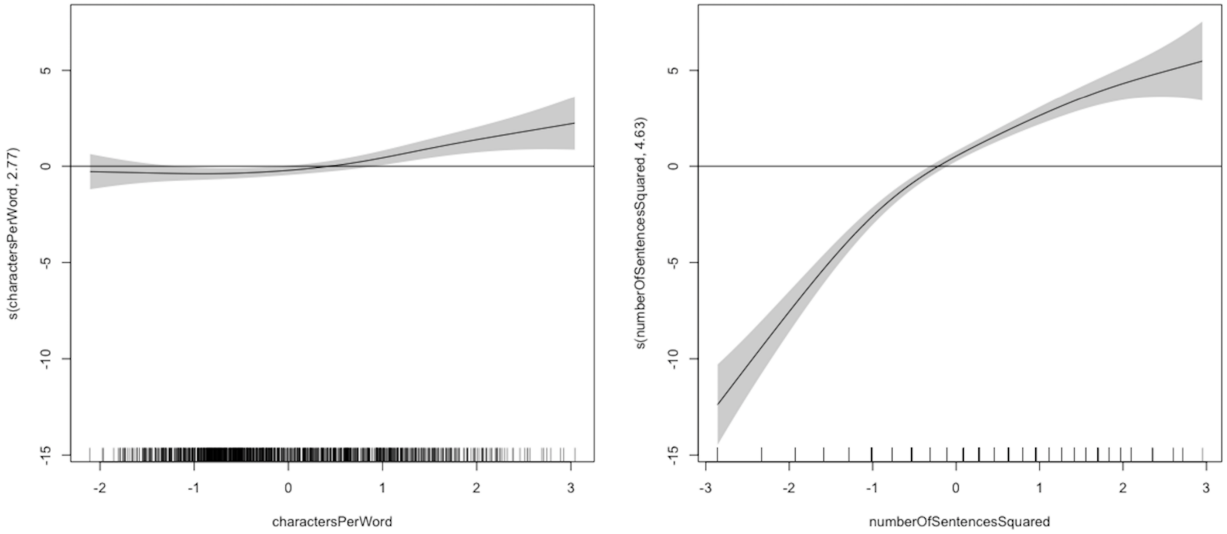

Figure: Smooths of *Merlin* interaction model.

Study 1: Task Effects

Classification Experiment

Model	μ F1	$\pm$SD	μ Recall	$\pm$SD	μ Precision	$\pm$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: Weighted average precision, recall, and f1 score for complexity, reference, and interaction model for 10 iterations of 10-folds cross-validation.

Study 1: Task Effects

Classification Experiment

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

Study 2: Performance Effects

Set Up

```
gam.merlin.success <- gam(OverallCefrScore ~
                            hasTransitionsFromSubjectToNot +
                            has3rdPersPossessivePronouns +
                            containsToInfinitives +
                            usesConjunctionalClauses +
                            logSumNonTerminalNodesPerSentence +
                            logATFBand2PerTypesFoundInDlex +
                            avgVTotalIntegrationCostAtFiniteVerb +
                            lexTypesFoundInDlexPerLexType +
                            typeTokenRato +
                            logSumNonTerminalNodesPerWord +
                            logATFBand2PerTypesFoundInDlex:TaskTheme +
                            s(charactersPerWord, by = Passed, k = 6) +
                            s(numberOfSentencesSquared, by = Passed) +
                            Passed +
                            TaskTheme,
                          data=merlin,
                          family=ocat(R=n_cat))
```

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

Study 2: Performance Effects

Model Fit

- $R^2 = 0.9000$
- Approximately homescedastic residual errors with $\mu = -0.14; sd = 12.91$ after outlier removal
- Same outliers as before, except for under-performing learners
- Outliers still systematically include learners who performed above or below test level

Study 2: Performance Effects

Model Fit

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

Study 2: Performance Effects

Model Discussion

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

Study 2: Performance Effects

Model Discussion

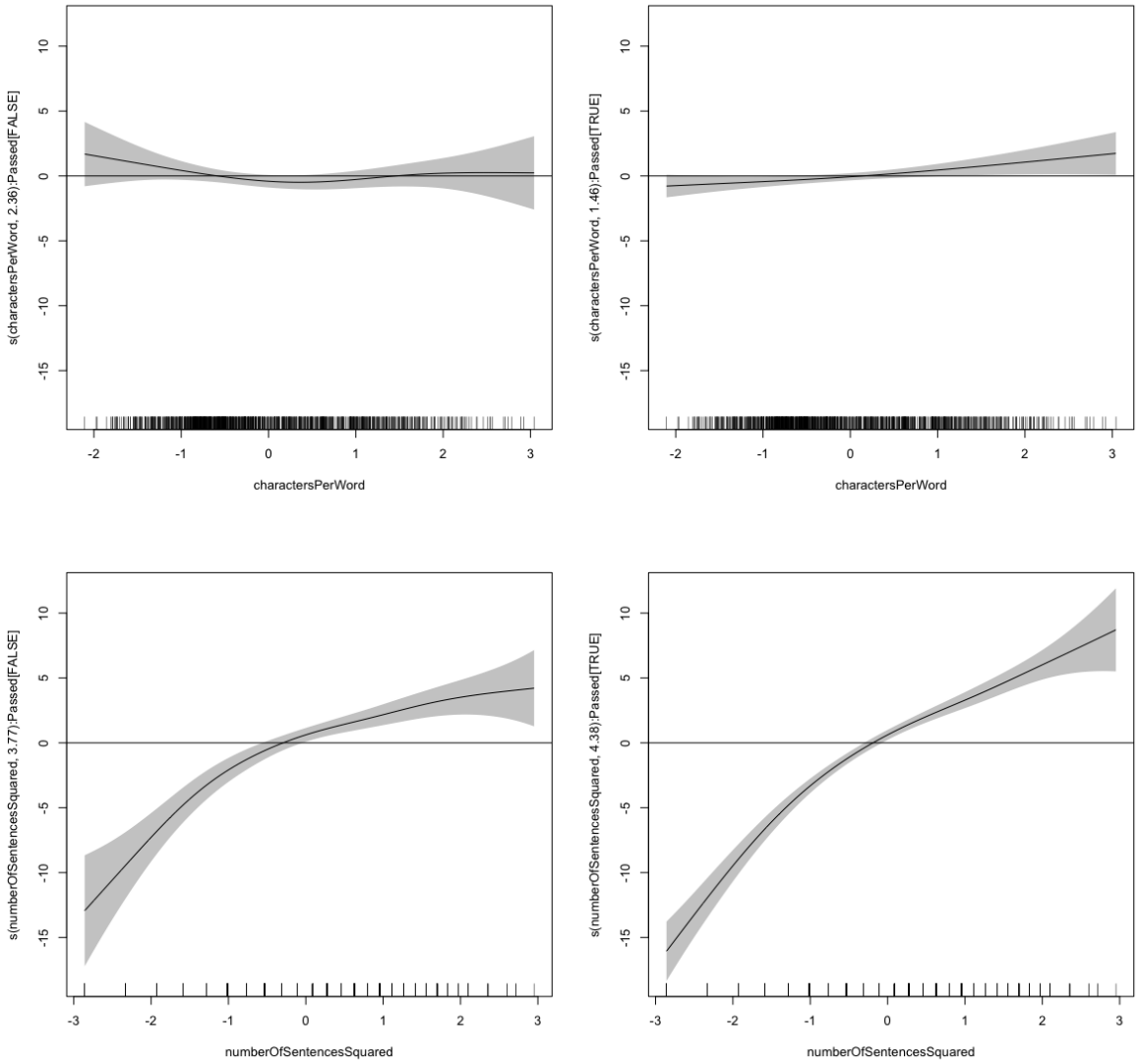

Figure: Smooths of *Merlin* success model.

Study 2: Performance Effects

Model Discussion

- Most task theme interactions become uninformative
- Still significantly different slopes, but not enough new variance explained
- Especially: texts about society heavily confounded with failed tests
- Unclear relationship between performance and task theme

Study 2: Performance Effects

Classification Experiment

Model	μ F1	$\pm$SD	μ Recall	$\pm$SD	μ Precision	$\pm$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: Weighted average precision, recall, and F1 score for complexity, reference, interaction and success model for 10 iterations of 10-folds cross-validation.

Study 2: Performance Effects

Classification Experiment

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

Conclusion

Findings

How do measures of complexity model German L2 proficiency?

- Most indices of the same concept tend to develop homogeneously and stable across corpora
- Most indices develop homogeneously across corpora
- Data-driven feature selection approaches yield diverse set of measures
- GAMs are highly interpretable, yet show considerable predictive power

Conclusion

Findings

To which extend is this influenced by cognitive or functional task-effects?

- Some measures are more stable across heterogeneous task backgrounds (human language processing, complex NPs)
- Other measures are less stable
- Stable measures especially promising for systems evaluating diverse task backgrounds
- Task factors seem to predominantly effect local measures of structural complexity
- Further research on this required

Conclusion

Findings

Does a retrospective analysis of German learner corpora with diverse task backgrounds improve complexity-based L2 proficiency modeling?

- Post-hoc annotation straight forward if task documentation available
- Suited to decrease confound of tasks and course levels
- Task factors improve model fit significantly and decrease non-linearity
- Interactions seem unstable, models suffer from wide standard deviation
- Results lack interpretability due to skewed distribution
- Analysis improves situation, but idiosyncratic distributional properties of data remain problematic

Future Work

Next:

1. Investigation of task and performance effects on more balanced data sets
2. Study of adequacy of L2 complexity by comparing results on comparable L1 productions (Falko)
3. Make complexity code used here publicly available in CTAP (Chen and Meurers 2016)

Also interesting:

- Analysis of task type interactions in data
- Cross-corpus testing of success model
- Systematically assess sensitivity of system and structure complexity to task effects
- Systematic validation of measure validity on L2 data

Thank you for your attention! Questions?

References I

- 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.
- Barzilay, Regina and Mirella Lapata (2008). "Modeling local coherence: An entity-based approach". In: *Computational Linguistics* 34, pp. 1–34.
- Biber, Douglas, Bethany Gray, and Kornwepa Poonpon (2011). "Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?" In: *Tesol Quarterly* 45, pp. 5–35.
- 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.

References II

- Bulté, Bram and Alex Housen (2014). "Conceptualizing and measuring short-term changes in L2 writing complexity". In: *Journal of Second Language Writing* 26, pp. 42–65.
- Chen, Xiaobin and Detmar Meurers (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. and Danielle S. McNamara (2011). "Shared features of L2 writing: Intergroup homogeneity and text classification". In: *Journal of Second Language Writing* 20, pp. 271–285.
- Duden (Gr) (2009). *Deutsche Grammatik*. Ed. by Ursula Hoberg and Rudolf Hoberg. 4th ed. Vol. 4. Der kleine Duden. Berlin, Germany: Dudenverlag.
- Ellis, R. and G. Barkhuizen (2005). *Analysing learner language*. Oxford: Oxford University Press.
- *Falko Georgetown Dokumentation* (2007). Humnoldt-Universität zu Berlin.

References III

- 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.
- Gibson, Edward (2000). "The dependency locality theory: A distance-based theory of linguistic complexity". In: *Image, language, brain*, pp. 95–126.
- Hancke, Julia (2013). "Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language". MA thesis. Eberhard Karls Universität Tübingen.
- Hancke, Julia, Sowmya Vajjala, and Detmar Meurers (2012). "Readability Classification for German using lexical, syntactic and morphological features". In: *Proceedings of COLING*. Mumbai, pp. 1063–1080.
- Housen, Alex, Ineke Vedder, and Folkert Kuiken (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.

References IV

- Jarvis, Scott et al. (2003). "Exploring multiple profiles of highly rated learner compositions". In: *Journal of Second Language Writing* 12, pp. 377–403.
- 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.
- Lavalley, Rémi, Kay Berkling, and Sebastian Stüker (2015). "Preparing Children's Writing Database for Automated Processing". In: *Workshop on L1 Teaching, Learning and Technology (L1TLT)*. Leipzig, Germany, pp. 9–15.
- 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.

References V

- 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/.

References VI

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/.

References VII

- 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. and S. Ferrari (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.
- Polio, Charlene and J.-H. Park (2016). "Language development in second language writing". In: *Handbook of second and foreign language writing*. Ed. by R. Manchón and P. K. Matsuda. Mouton de Gruyter.
- Rescher, Nicholas (1998). *Complexity: A philosophical overview*. Transaction Publishers.
- Reznicek, Marc et al. *Das Falko-Handbuch Korpusaufbau und Annotationen*. Humnoldt-Universität zu Berlin.

References VIII

- Robinson, Peter (2001). "Task Complexity, Task Difficulty, and Task Production: Exploring Interactions in a Componential Framework". In: *Applied Linguistics* 22.1, pp. 27–57.
- 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.
- 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 and Florence Myles (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.

References IX

- von der Brück, Tim and Sven Hartrumpf (2007). "A Semantically Oriented Readability Checker for German". In: *Proceedings of the 3rd Language & Technology Conference*, pp. 270–274.
- von der Brück, Tim, Sven Hartrumpf, and Hermann Helbig (2008). "A Readability Checker with Supervised Learning Using Deep Indicators". In: *Informatica* 32, pp. 429–435.
- 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.
- Weiß, Zarah Leonie (2017). "Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects". MA thesis. Eberhard Karls Universität Tübingen.
- Wolfe-Quintero, Kate, Shunji Inagaki, and Hae-Young Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center.

References X

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

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

Robinson's Cognition Hypothesis

Task complexity	Task conditions	Task difficulty
(cognitive factors)	(interactional factors)	(learner factors)
(a) resource-directing	(a) participation variables	(a) affective variables
e.g., +/– few elements	e.g., open/closed	e.g., motivation
+/– here-and-now	one-way/two-way	anxiety
+/– no reasoning demands	convergent/divergent	confidence
(b) resource-depleting	(b) participant variables	(b) ability variables
e.g., +/– planning	e.g., gender	e.g., aptitude
+/– single task	familiarity	proficiency
+/– prior knowledge	power/solidarity	intelligence

Sequencing criteria
Prospective decisions
about task units

Methodological influences
On-line decisions
about pairs and groups
(based on Robinson, in press*a*)

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

Language Use

DlexDB frequencies

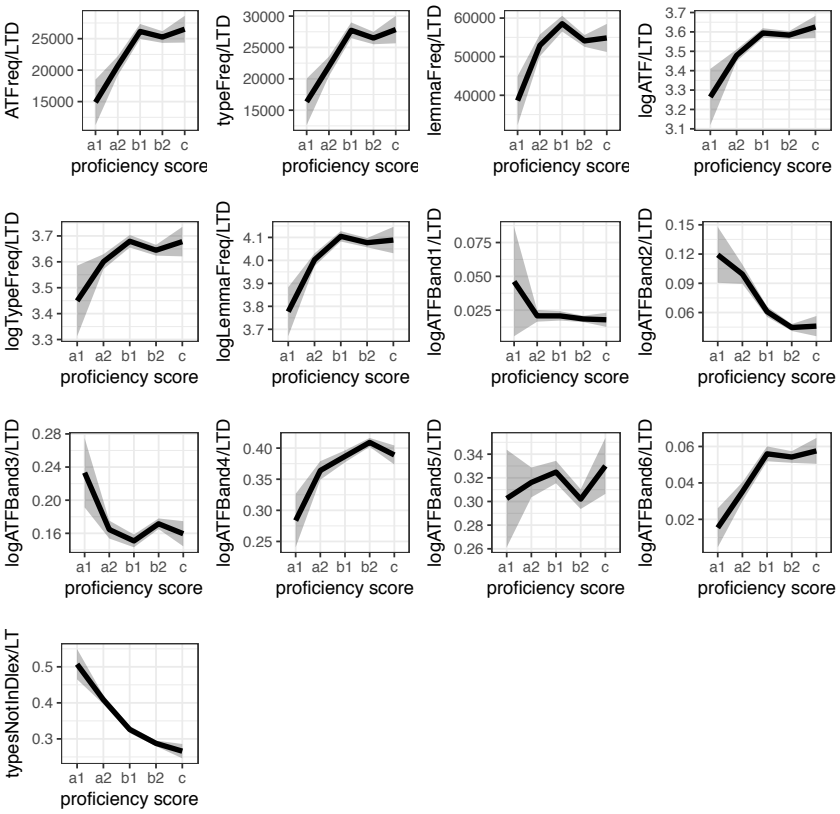

Figure: *Merlin*.

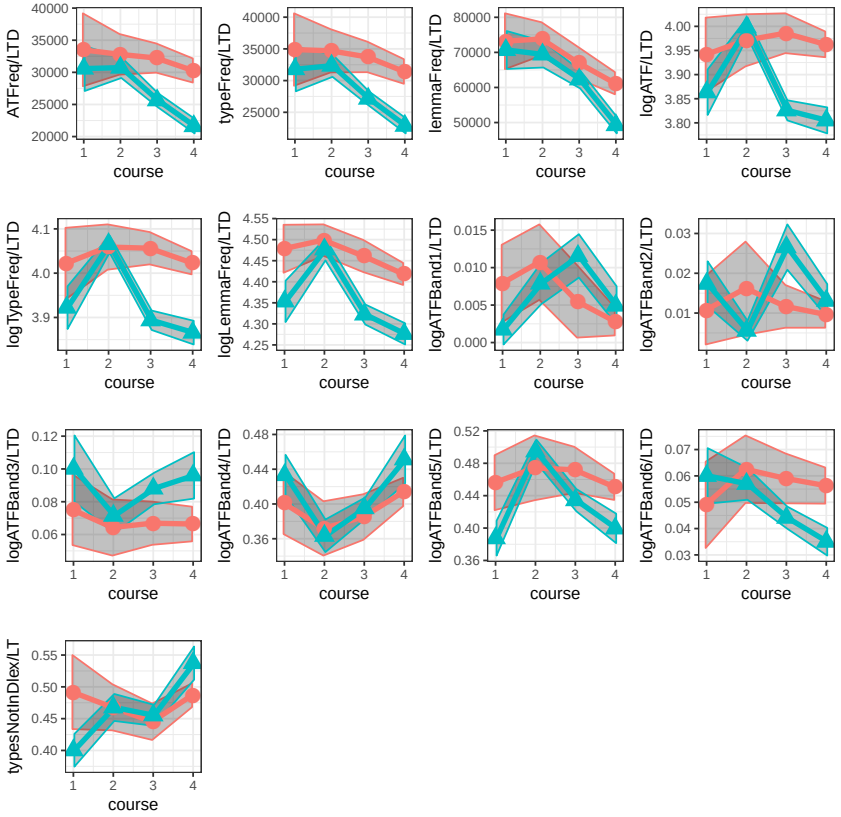

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Discourse & Encoding of Meaning

Pronouns, articles and names

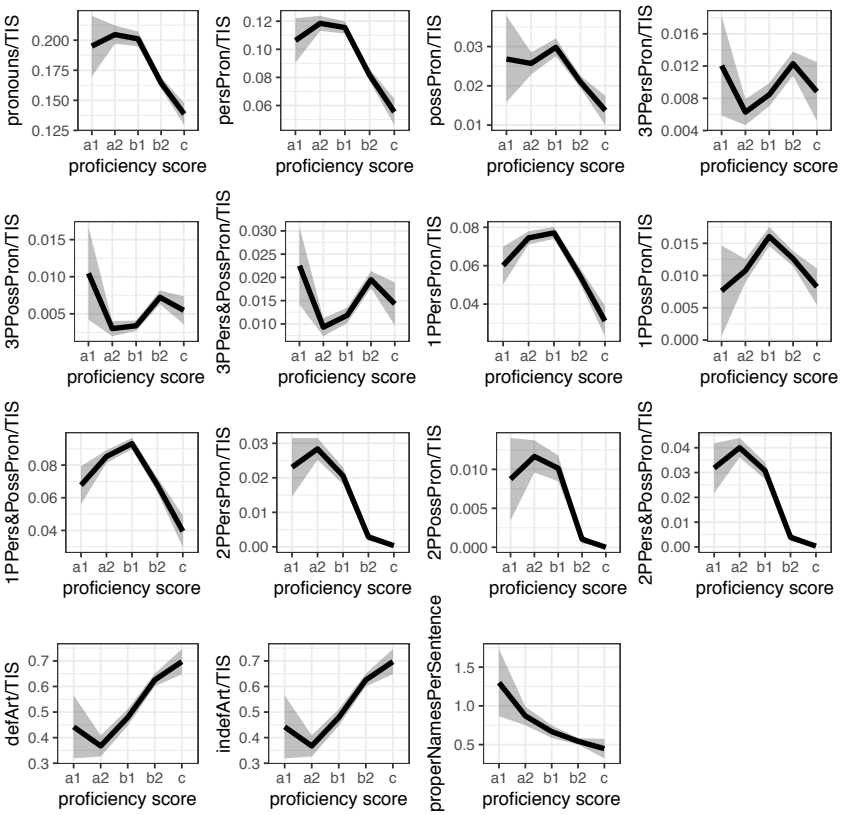

Figure: *Merlin*.

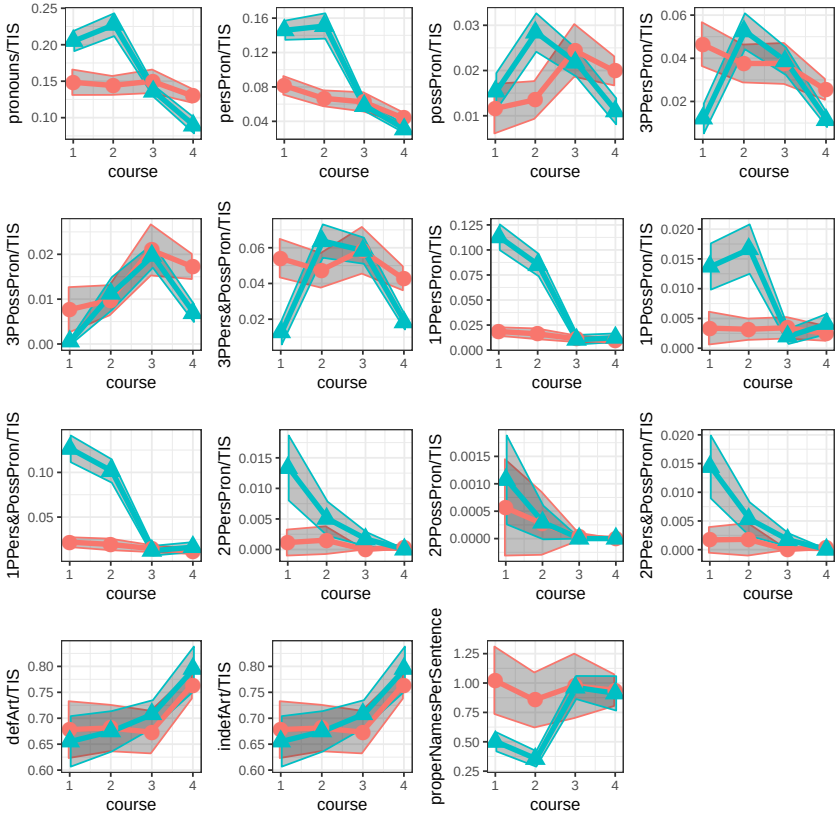

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Lexical Complexity

Lexical Variation

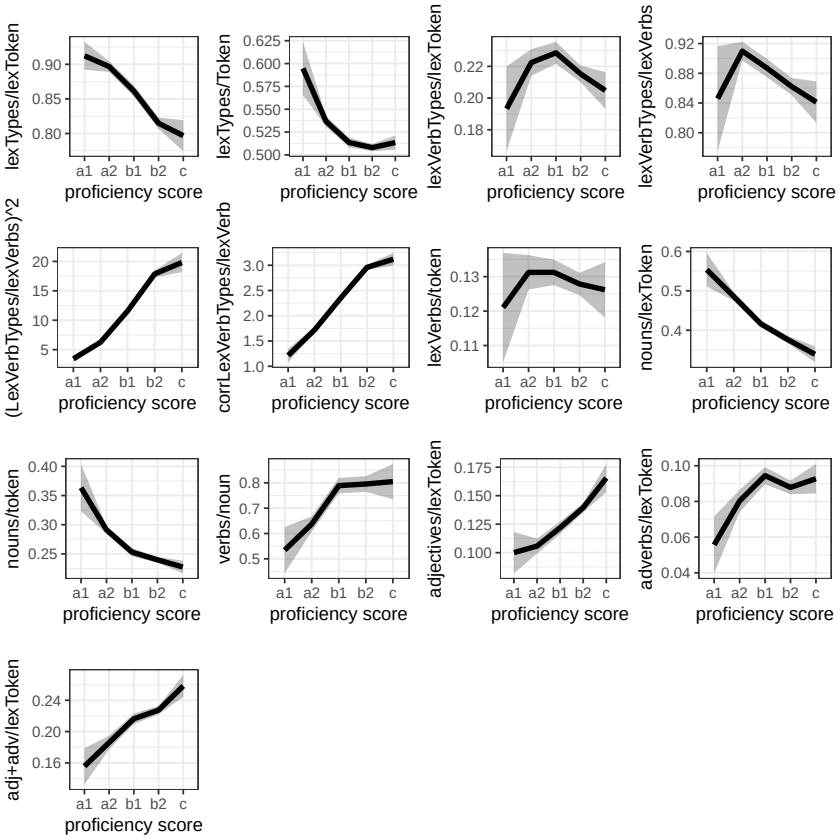

Figure: *Merlin*.

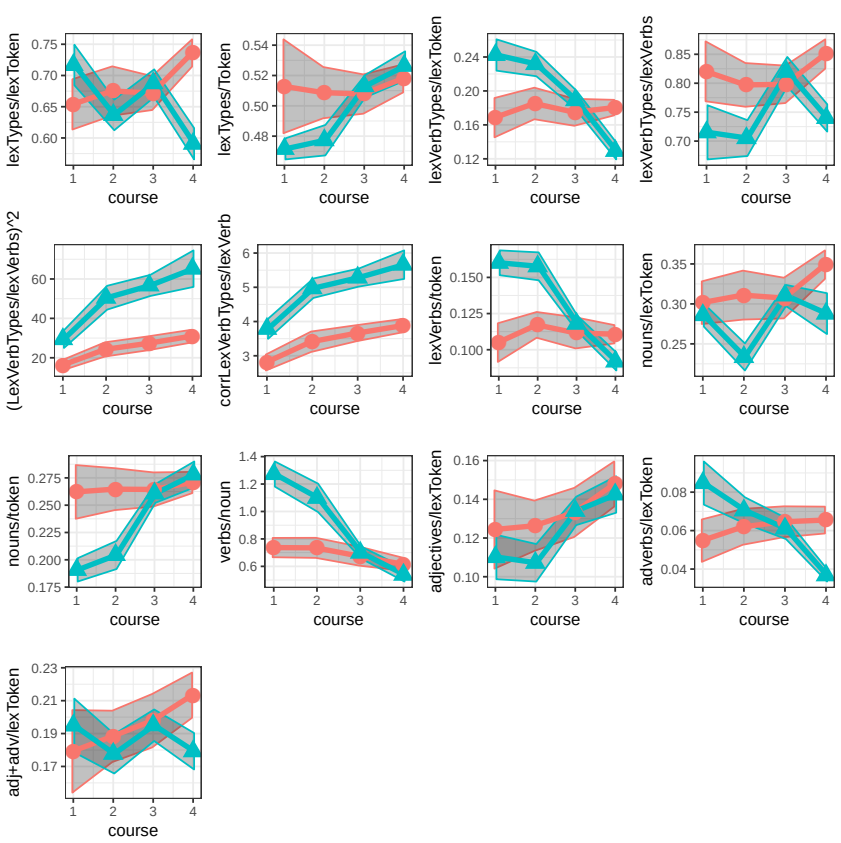

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Syntactic Complexity

Periphrastic grammatical measures

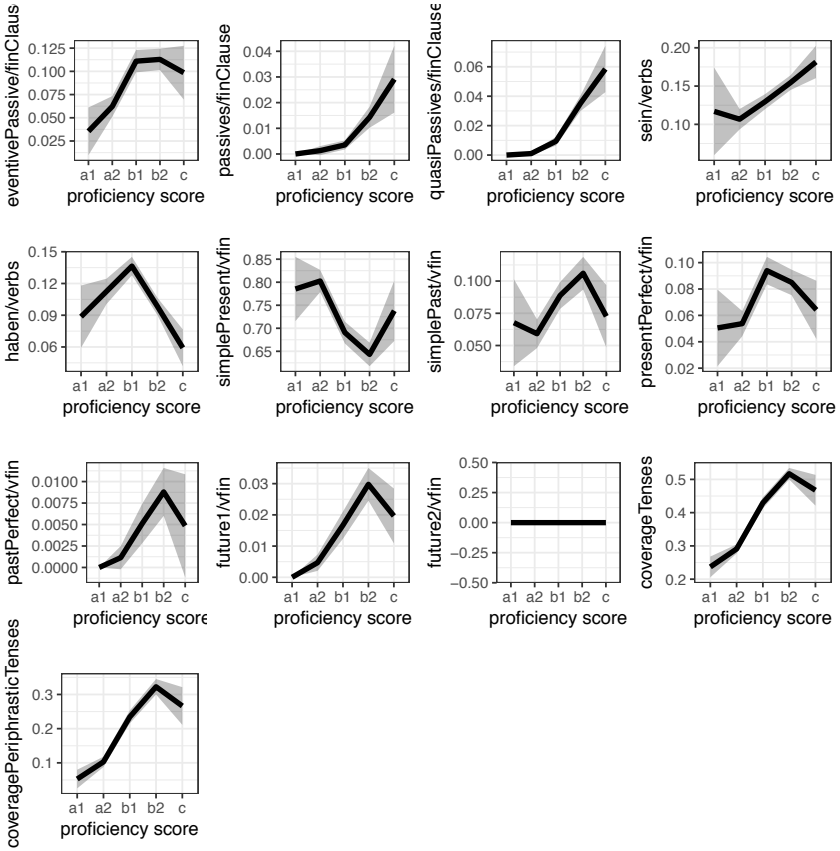

Figure: *Merlin*.

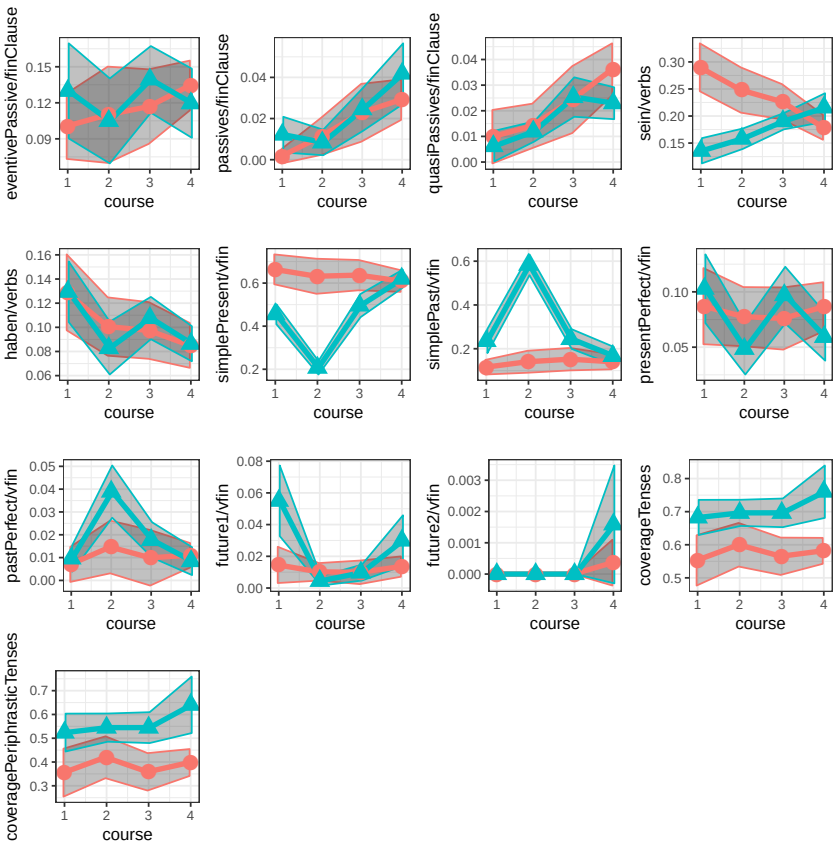

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Syntactic Complexity

Dependent clause measures

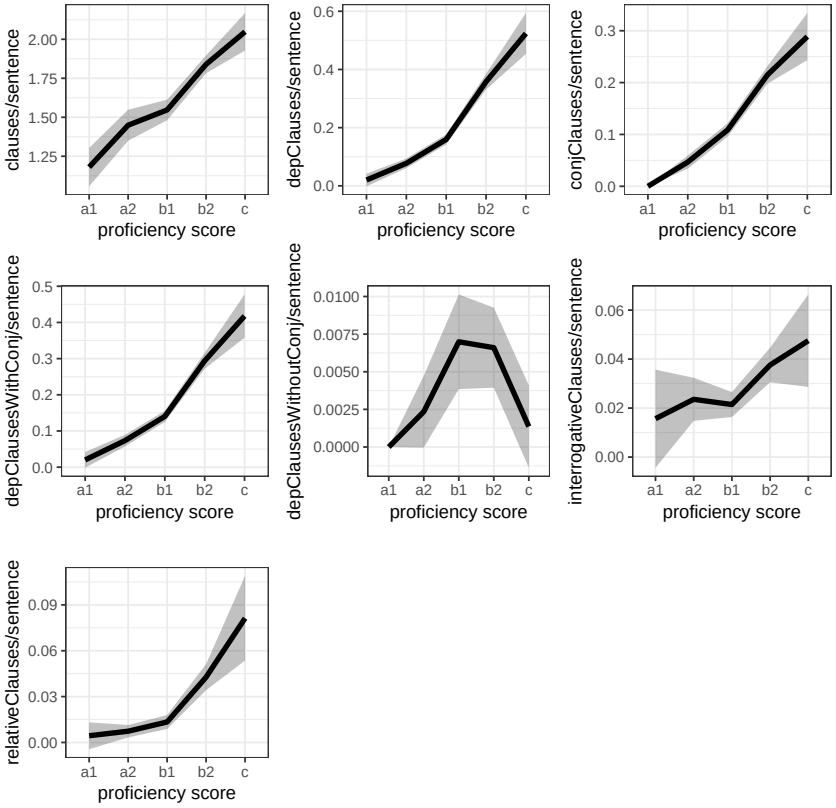

Figure: *Merlin*.

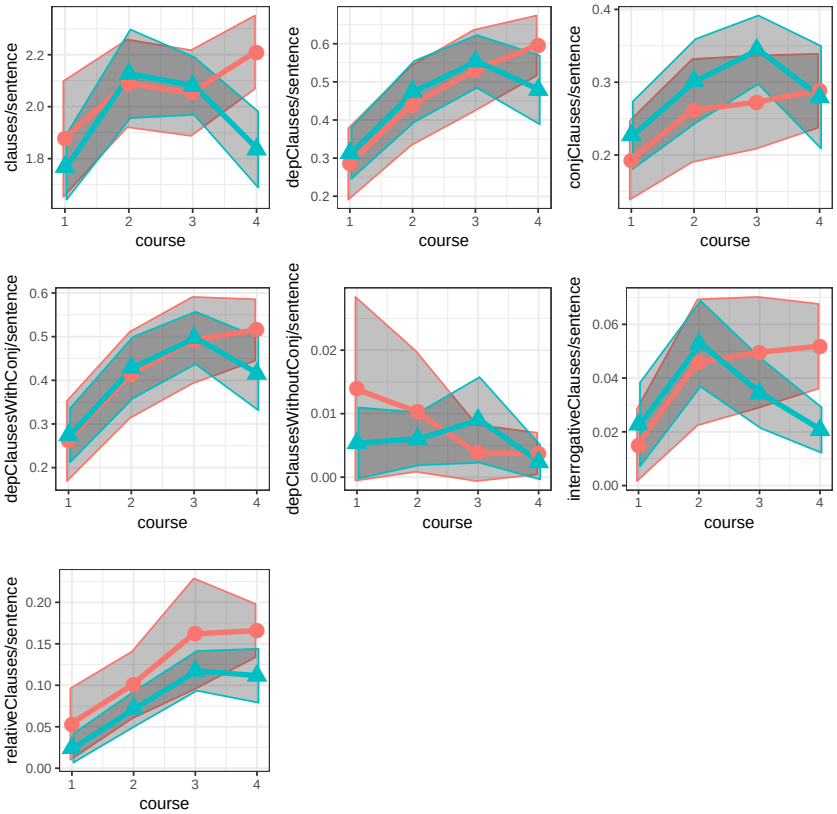

Figure: *Falko GT L2*; △: curricular tasks, ○: book reviews.

Generative Additive Regression Models

From Linear to Additive Models

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

$$g(\hat{y}) = \eta + \epsilon, \ \textit{where } \eta = \beta_0 + \sum_{i=1}^{I} x_i \beta_i \tag{2}$$

$$g(\hat{y}) = \eta + \epsilon, \ \textit{where } \eta = \beta_0 + \sum_{i=1}^{I} s_i(x_i) \tag{3}$$

Generative Additive Regression Models

From Linear to Additive Models

$$g(\hat{y}) = \eta + \epsilon, \ \textit{where } \eta = \beta_0 + \sum_{i=1}^{I} s_i(x_i) \tag{4}$$

$$s(x) = \sum_{k=1}^{K} b_k(x)\beta_k, \tag{5}$$

$$s(x) = \sum_{c=1}^{C+1} x^{c-1}\beta_c \tag{6}$$

Generative Additive Regression Models

Regression Splines

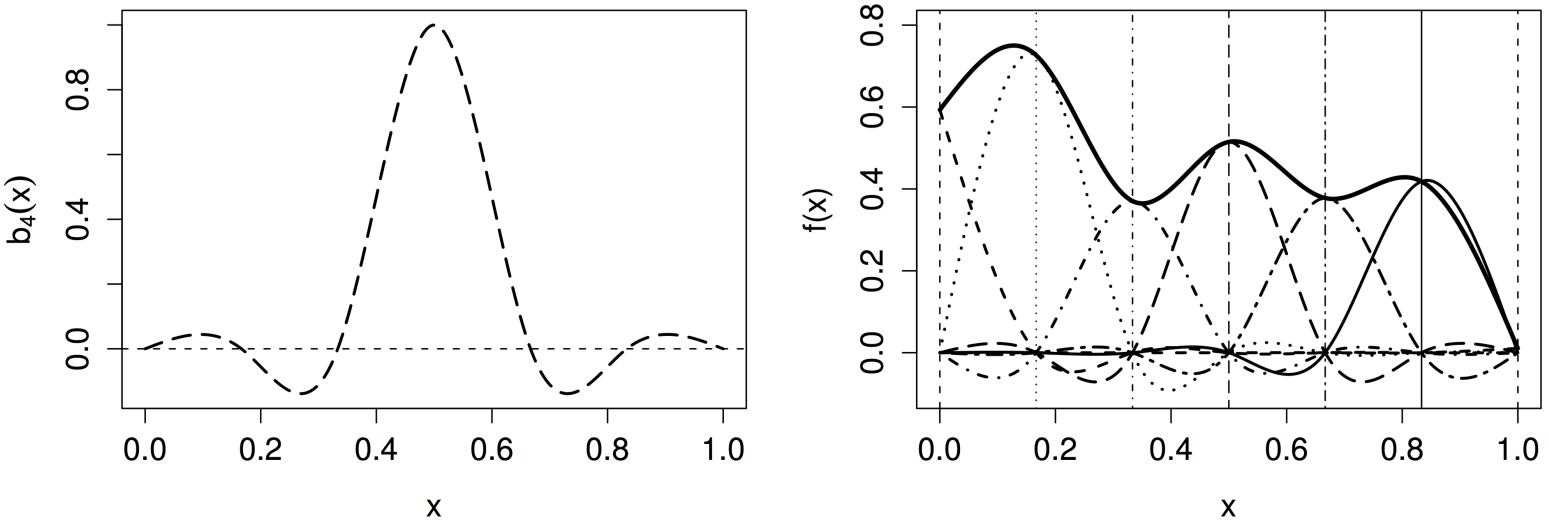

Figure: Single cubic basis function (left) and full cubic regression spline (right), cf. Wood 2006, p. 147, Figure 4.1.

Generative Additive Regression Models

Regression Splines

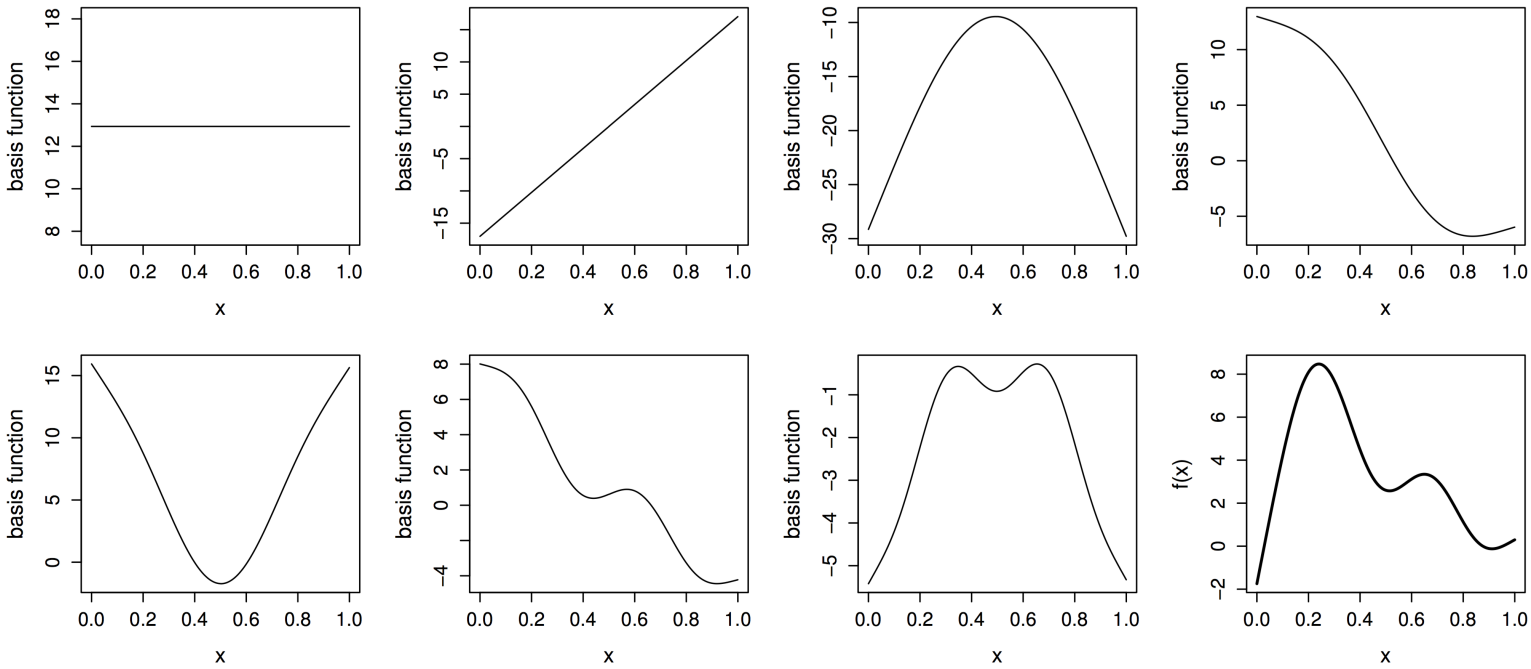

Figure: A rank 7 thin plate regression spline preceded by its weighted basis functions, cf. Wood 2006, p. 153, Figure 4.5.

Generative Additive Regression Models

Ordinal Models

$$u = \eta + \epsilon, \ \textit{where} \ \eta = \beta_0 + \sum_{i=1}^{I} s_i(x_i), \ \textit{and} \ u \ \in \ [\pm\infty] \tag{7}$$

- Ordinal data neither numeric nor nominal
- Ordinal distribution not covered in exponential link functions $g()$
- Solution by Wood 2006: partition $\pm\infty$ into K bins using K-1 boundaries
- Estimate latent variable u with regression model
- Assign ordinal category based on interval in which u falls

Generative Additive Regression Models

Ordinal Models

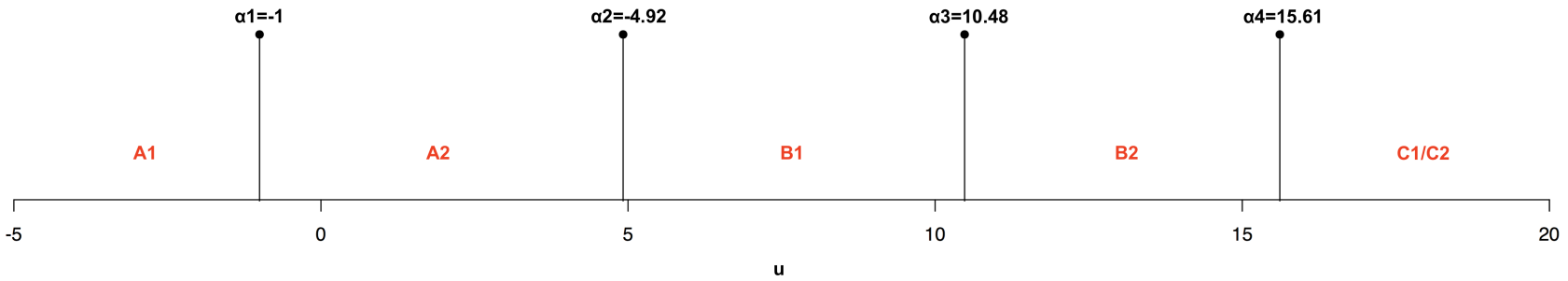

Figure: Mapping of latent variable u to CEFR levels using estimated boundaries from the *interaction* model of Study 1.

Automatic System

Overview

English Systems

- Coh-Metrix Web Interface (McNamara et al. 2014)
- Syntactic and Lexical Complexity analyzer (Lu 2010)
- Linguistic Analysis Tool (Kyle 2016)
- Common Text Analysis Platform (Chen and Meurers 2016)

German Systems

- DeLite Readability Checker (von der Brück and Hartrumpf 2007; von der Brück, Hartrumpf, and Helbig 2008)
- Tübinger complexity code (Hancke, Vajjala, and Meurers 2012; Weiß 2015)