Weiss, Z. & Meurers, D. (2019). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In Andrea Abel, Aivars Glaznieks, Verena Lyding & Lionel Nicolas (eds) *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Corpora and Language in Use – Proceedings 5, Louvain-la-Neuve: Presses universitaires de Louvain, 419-435.

# Broad linguistic modeling is beneficial for German L2 proficiency assessment

Zarah Weiss & Detmar Meurers

University of Tübingen[1]

**Abstract**

We investigate the applicability of a broad range of language features to German second language proficiency assessment by comparing the performance of classification models based on linguistically diverse vs. homogeneous feature groups in terms of their overall performance and their success at individual proficiency levels (A1 to C1/C2). For this, we extract 400 measures of linguistic complexity from the domains of syntax, lexicon, morphology, discourse, language use, and human language processing. Overall, our results show that a broad feature set integrating aspects of language as a system, language use, and human sentence processing costs results in higher classification performance on language learner data. At individual proficiency levels, lexical complexity in particular, but also clausal and phrasal complexities as well as discourse measures successfully distinguish several proficiency levels. Morphological complexity is particularly important for more advanced learners.

## 1. Introduction

This study investigates the applicability of a broad range of language features to German second language (L2) proficiency assessment. We focus on aspects

1 http://icall-research.de

of *Complexity*, a core component of the triad *Complexity, Accuracy, and Fluency* (CAF) that is used in Second Language Acquisition (SLA) research to characterize language performance (Housen *et al.* 2012). In recent years, diverse features have been proposed to measure language proficiency, readability, and writing skills (Bulté & Housen 2014; Ortega 2012). They differ in terms of the nature of the language characteristics they measure, their specificity, sensitivity to task-effects, and how difficult it is to extract the information. To which extent the combination of diverse features is beneficial, as far as we are aware, has not been systematically investigated. Furthermore, while it is common to introduce the benefits of so far under-researched domains of linguistic complexity, such as morphology or SLA based features, a detailed comparison of which domains of linguistic complexity discriminate best at certain levels of proficiency has less often been attempted. We address this by comparing (1) performance differences between German L2 proficiency classifiers based on either broad, linguistically diverse or homogeneous feature groups; and (2) performance differences of linguistically homogeneous classifiers at individual proficiency levels (A1 to C1/C2). We find that linguistically diverse proficiency models that combine features from various linguistic domains systematically outperform those informed by individual linguistic domains. Regarding the informativeness of these linguistic domains, we find that single features from all linguistic domains that we measured are highly informative. Regarding the contribution of feature groups comprised from one linguistic domain to the identification of individual proficiency levels, we find in particular lexical, but also clausal and phrasal complexities as well as discourse measures to be highly successful across levels.

The remainder of the article is structured as follows. We briefly review previous work on the link between linguistic complexity and different levels of L2 proficiency, before outlining our automatic complexity analysis approach. We then introduce the data from the Merlin corpus which we use for our analyses. This is followed by our classification study and an outlook on current and future work, before we conclude with some final remarks.

## 2. Related work

Automatic complexity analyses for proficiency assessment often focus on longitudinal English L2 data elicited in University contexts for a certain group of L2 learners, such as intermediate or advanced learners, rather than distinguishing multiple proficiency levels at once. Thus, they focus more on developmental patterns that may be observed within a group of learners.

Comparisons of proficiency levels are mostly based on the collective evaluation of multiple studies targeting different learner groups, which may be potentially problematic due to different study set-ups or operationalizations of complexity.

Overall, such studies have found that learners at low and intermediate proficiency levels predominantly develop in terms of sentence length and clausal elaborateness (Lu 2010; Norris & Ortega 2009; Ortega 2003), the latter of which has been shown to be correlated with human proficiency ratings (Crossley & McNamara 2014). For more advanced English L2 learners, research indicates a stronger development of the phrasal domain, in particular regarding noun phrases (Crossley & McNamara 2014; Taguchi *et al.* 2013). Other complexity measures, such as lexical and clausal complexity, were found to be less informative to distinguish between advanced-intermediate and advanced learners (Paquot 2017; Ortega 2012; Biber *et al.* 2011). Some studies indicate that advanced learners also develop in terms of lexical abstractness, lexical familiarity, and semantic inter-relatedness (Crossley *et al.* 2014; Crossley & McNamara 2012), but that this development is not necessarily considered for advanced proficiency ratings (Crossley & McNamara 2014). As for discourse measures, studies for more advanced learners have found that more proficient learners use more implicit cohesion markers and less explicit markers, such as connectives (Crossley *et al.* 2014; Crossley & McNamara 2012; McNamara *et al.* 2009).

While there is an extensive body of research on English L2 development, there is overall less research on German complexity assessment, most of which focuses on German readability assessment (Hancke *et al.* 2012; vor der Brück *et al.* 2008). Hancke & Meurers (2013) investigate how measures of clausal, lexical, and morphological complexity as well as language model features relate to CEFR ratings. They find lexical and morphological complexities to be most informative, and clausal complexity, while less informative on its own, to boost classification performance when being combined with the other measures, and that a combined model of features overall performs best with accuracy values of 62.7%. In this study, we follow up on these results on the same corpus with a broader set of features and analytical methods.

## 3. Automatic complexity analysis

For our automatic analysis of German linguistic complexity, the elaborateness and variation in the different domains of linguistic modeling (Ellis & Barkhuizen 2005), we extracted 400 features using an elaborate NLP tool chain.

### 3.1. Complexity measures

Our features may broadly be grouped into two categories: those targeting dimensions of the theoretical linguistic system (syntax, lexicon, morphology, and discourse) and those targeting the cognitive or psycholinguistic dimension of language productions (language use and human language processing). Additionally, we calculate two descriptive, superficial features of text length in words and sentences.

**Clausal and phrasal complexities** assess syntactic complexity development on two levels: clausal complexity is associated with phenomena such as clausal subordination and the use of syndetic and asyndetic constructions. Our system measures in particular various types of subordination and clausal structure (t-units per sentence, dependent clauses per t-unit, etc.). Phrasal complexity measures aspects of phrasal modification and coordination. We assess these in terms of various modifier ratios and coverage of modifier measures with a particular focus on the nominal domain and verb clusters.

**Lexico-semantic complexity** is typically associated with vocabulary range (lexical density and variation) and size (lexical sophistication), but also lexical relatedness. We measure lexical diversity using raw type token ratio as well as its length normalized variants. Lexical density and variation are assessed for various Parts-of-Speech (PoS), including for example verb and noun variations. To measure lexical relatedness, we assess the number of semantic relations between words (hyponymy, synonymy, etc.) using GermaNet 9.0.1 (Henrich & Hinrichs 2010).

**Morphological complexity** has shown to be particularly interesting for languages that exhibit richer morphology than English, such as German or French (François & Fairon 2012; Hancke *et al.* 2012). We measure features of inflection (tenses, person, number, etc.), derivation (in particular nominalizations), and composition.

**Discourse measures** assess textual cohesion, i.e. the linguistic items that link propositions or idea units, which has been shown to be highly informative in previous work on complexity assessment among others by the *CohMetrix* project. Following them, we measure co-referential cohesion in terms of noun, argument, stem, and content-word overlaps, pronoun ratios, and various types of connectives. We also adopted transitional features from Barzilay & Lapata (2008) that assess changes in grammatical functions (subject, object, other complement, not present) that are assigned to repeated linguistic material in adjacent sentences.

**Language use measures** origin from psycho- and corpus-linguistic research and include, for example, word frequencies from representative language samples or approximations of age of acquisition. Word frequency measures are well established in complexity research, although they are often listed among features of lexical complexity. We calculate a series of word frequency measures based on frequency data bases Subtlex-DE and Google Books 2000-2009 (Brysbaert *et al.* 2011), as well as dlexDB (Heister *et al.* 2011). These features include absolute and log transformed frequencies. We also approximate average and first age of active use through the KCT corpus (Lavalley *et al.* 2015).

**Human language processing measures** are based on research in cognitive science and information theory. They evaluate complexity in terms of processing costs associated with linguistic material, for example in terms of surprisal or cognitive load. We measure cognitive load in terms of integration costs based on dependency lengths and Gibson's (2000) Dependency Locality Theory (DLT). For the latter, we follow Shain *et al.*'s (2016) dependency parse based operationalization including variants which feature their suggested weight modifications for verbs, coordination, and modifiers.

### 3.2. System description

We extract our complexity features based on a three-step procedure. First, each text is linguistically annotated by applying a series of NLP tools and consulting external linguistic resources. In particular, texts are tokenized, segmented into sentences using OpenNLP 1.6.0 (http://opennlp.apache.org). Then, we perform PoS tagging, lemmatization, morphological analysis, and dependency parsing using the Mate tools 3.6.0 (Bohnet & Nivre 2012). We perform compound analysis using the JWordSplitter 3.4.0. (http://github.com/danielnaber/jwordsplitter). Finally, we obtain constituency parses using the Stanford PCFG parser 3.6.0 (Rafferty & Manning 2008) and topological field parses using the Berkeley parser 1.7.0 (Petrov & Klein 2007). While for many of these tasks, other NLP tools could also be employed, the mentioned tools, as far as we are aware, perform close to the state of the art in terms of quality and speed so that an exploration of alternatives is beyond the scope of this paper. In general, we use the default models provided by the respective tools for German analyses, except for topological field parsing, for which we used the model trained by Ziai & Meurers (2018) because the default topological field model is not compatible with the latest version of the parser.

These linguistic annotations are used in the second step to identify all instances of linguistic constructions that are relevant for our complexity analysis. This step relies on the identification of different linguistic units, some of which have different justifiable operationalizations, such as t-units or lexical words. To allow for comparability across complexity studies, it is crucial to make the underlying definitions of these units explicit (Bulté & Housen 2014; Housen *et al.* 2012). An elaborate documentation of the units underlying our system may be found in Weiss (2017: 82f). In the final step of the analysis, ratios and features are calculated to approximate the complexity of each document by means of a feature vector.[2] These are exported into a CSV table including all documents, which may then be used for further statistical evaluation.

To the best of our knowledge, this is currently the most extensive feature set for German complexity assessment. We are in the final stages of making the system publicly accessible via the Common Text Analysis Platform (CTAP) by Chen & Meurers (2016), which originally only facilitated English complexity analyses.

## 4. Merlin data

We analyze the non-normalized German section of the Merlin corpus (Abel *et al.* 2014) to assess German L2 writing proficiency. It is comprised of 1,033 texts written by the same number of adult learners of German, which have been elicited in official standardized language certification tests for the five CEFR test levels A1 to C1. With this, it is not only to our knowledge the largest freely available German L2 corpus, it also features text from an extraordinarily broad variety of thoroughly and transparently established proficiency levels. The corpus consists of approximately 200 texts per test level, which were prompted by overall 15 different tasks (three tasks per level). All texts are rated based on the CEFR scale from levels A1 to C2 by human experts for various performance categories as well as a holistic overall proficiency rating (Abel *et al.* 2014). Since learners achieved not only proficiency scores at the level of the test they took, but also scores above or below, the uniform distribution of test levels in Merlin does not translate to a uniform distribution of proficiency scores. Due to the negligible number of C2 rated texts (4 in total), we combined C1 and C2 texts to a single C1/C2 level rating for the purposes of our statistical analyses.

2 All features and formulas available at http://www.sfs.uni-tuebingen.de/~zweiss/rsrc/feat.pdf

# 5. Classification study

## 5.1. Set-up

As classification algorithm for our assessment of overall L2 proficiency, we chose the Sequential Minimal Optimization (SMO) support vector classifier by Platt (1998) with a linear kernel. This state-of-the-art algorithm is known to be relatively robust for many, potentially correlated measures and thus particularly suited for our large sets of complexity measures. We applied the SMO algorithm to varying combinations of complexity features. First, we grouped features together that assess the same theoretical or psycholinguistic linguistic domain. This resulted in seven linguistically homogeneous, theory-based feature groups: clausal, phrasal, lexico-semantic, and morphological complexity, discourse, human language processing (HLP), and language use (LU). Second, we performed information gain ranking to identify data-driven the most 50, 100, 150, and 200 informative features across all seven linguistic domains. We then discarded all but the most successful classifier, *IG 150*, which uses the 150 most informative complexity measures. Finally, we trained a classifier using all features. All classifiers were trained and tested using 10-fold cross-validation. The information gain ranking, too, was performed with this method. For further comparison, we also obtained a majority baseline by automatically assigning the most frequent proficiency level (B2) to all texts. We used the WEKA machine learning toolkit (Hall *et al.* 2009) for all analyses.

## 5.2. Results and discussion

### *5.2.1. Overall classification performance*

Table 1 shows the overall performance of the different proficiency classifiers. All of them clearly outperform the majority baseline (32.0%). The best performing classifier is *IG 150*. Compared to the model using all features, removing less relevant features seems to slightly increase classification performance. More importantly, however, *IG 150* clearly outperforms the linguistically homogeneous classifiers yielding accuracies between 53.7% (HLP) to 67.6% (lexico-semantic).

| Model | Num. features | Accuracy |
|---|---|---|
| Majority baseline | 1 | 32.0 |
| All features | 400 | 68.1 |
| IG 150 | 150 | 70.0 |
| Discourse | 84 | 64.7 |
| Clausal | 110 | 63.8 |
| Phrasal | 41 | 62.1 |
| Lexico-semantic | 38 | 67.6 |
| Morphological | 39 | 59.7 |
| HLP | 32 | 53.7 |
| Language use | 54 | 59.3 |

*Table 1. Classification performance of SMO models in 10-fold cross-validation*

Table 2 shows the confusion matrix for *IG 150*. Columns represent the proficiency scores predicted by the model, rows the actual proficiency scores assigned to a text in the Merlin corpus. For each observed score the most often predicted score was marked with bold font.

| Obs.\Pred. | A1 | A2 | B1 | B2 | C1/C2 | ∑ Obs. |
|---|---|---|---|---|---|---|
| A1 | 21 | **35** | 1 | 0 | 0 | 57 |
| A2 | 13 | **231** | 62 | 0 | 0 | 306 |
| B1 | 1 | 50 | **218** | 62 | 0 | 331 |
| B2 | 0 | 3 | 37 | **252** | 1 | 293 |
| C1/C2 | 0 | 0 | 1 | **44** | 1 | 46 |
| ∑ Pred. | 35 | 319 | 319 | 358 | 2 | 1,033 |

*Table 2. Confusion matrix for IG 150*

CEFR levels A2, B1, and B2 show favorable classification results: most predictions for texts from these levels are correct. For levels A1 and C1/C2, however, miss-classifications with their adjacent level are more common than correct classifications. This issue is particularly severe for level C1/C2. There are several potential explanations for this issue: partially it might be an artifact of the skewed distribution of Merlin overall CEFR scores and the resulting under-representation of these two levels: less than 10% of the corpus contains data with overall CEFR scores at levels A1 and C1/C2. For level A1, the classification might also suffer from the highly non-standard language of beginning learners, which impairs the automatic NLP annotations on which the

complexity features are based. Tono (2013) observes a risk-taking phase reaching into the intermediate level, where the number of errors increases together with complexity. While this could also cause problems for the NLP analysis, there is no indication for this in our results given the high classification accuracy for intermediate learners. For level C1/C2 another plausible explanation would be that the differences between B2 and C1 learners relate more to phraseological and stylistic writing aspects (Paquot 2017; Biber *et al.* 2011), which are not sufficiently captured in the current set of complexity features.

| Rank | Avg. merit | Feature | Group |
|---|---|---|---|
| 1 | 0.889 ± 0.010 | Number of tokens | Descriptive |
| 2 | 0.827 ± 0.019 | Corrected type token ratio | Lexical |
| 7 | 0.466 ± 0.009 | Longest word in syllables | Lexical |
| 8 | 0.432 ± 0.015 | Sum of longest dependencies per sentence | HLP |
| 13 | 0.391 ± 0.011 | Dep. clauses with conjunction per dep. clause | Clausal |
| 14 | 0.391 ± 0.006 | Coverage of NP modifier types | Phrasal |
| 16 | 0.387 ± 0.009 | Dependent clauses per sentence | Clausal |
| 22 | 0.372 ± 0.009 | P(not-not) per transition | Cohesion |
| 25 | 0.369 ± 0.012 | Verbs per sentence | Phrasal |
| 27 | 0.359 ± 0.025 | VP modifiers per VP | Phrasal |
| 29 | 0.358 ± 0.015 | Words per t-unit | Clausal |
| 31 | 0.355 ± 0.007 | Sum non-terminal nodes per word | Clausal |
| 35 | 0.354 ± 0.013 | Standard deviation of verb cluster sizes | Phrasal |
| 36 | 0.350 ± 0.007 | P(not-object) per transition | Cohesion |
| 37 | 0.350 ± 0.005 | To-infinitives per sentence | Phrasal |
| 39 | 0.346 ± 0.009 | Total integration cost at finite verb per finite verb (with additional verb weight) | HLP |
| 43 | 0.344 ± 0.006 | HDD | Lexical |
| 44 | 0.341 ± 0.011 | Syllables per word | Lexical |
| 50 | 0.326 ± 0.008 | Temporal (Eisenberg) connectives per sentence | Cohesion |
| 52 | 0.324 ± 0.013 | Coverage of verb cluster sizes | Phrasal |

*Table 3. Top 20 complexity measures based on 10-fold cross-validated information gain with Pearson correlation less extreme than r ± 0.7*

To examine closer the best performing classifier, *IG 150*, Table 3 shows the 20 most informative features included in the model. To allow for a broader view on the features represented in the model, we excluded measures which showed an extremely high Pearson correlation with higher ranking measures, i.e. that were more extremely correlated than ±0.7. The table shows the original total rank of each feature and their average merit in the 10-fold cross validation to allow for a more informed comparison of their overall informativeness. It also includes a reference to the feature group each feature is attributed to.

The results confirm that our data-driven feature selection approach in fact yields a highly diverse collection: the features include measures from nearly all feature groups and include operationalizations of the elaborateness and variation of these domains. The elaborateness of clausal subordination, the elaborateness and variation of nominal and verbal modification, lexical diversity and sophistication, transitions of grammatical roles and temporal connectives, and dependency-length based cognitive integration costs are particularly informative. Features of language use and morphological complexity are not represented in Table 3. However, they are repeatedly represented among the most informative 50 not extremely correlated features. Furthermore, morphological complexity features are represented among the higher-ranking measures, but highly correlated with word length and corrected type token ratio thus not eligible for Table 3. This holds in particular for derivational measures indicating nominalizations. The informativeness of language use measures is partially impaired by the type of data that is being analyzed: since we do not investigate the normalized texts and misspelled words will not be found in any of our word frequency data bases.

### *5.2.2. Classification performance by proficiency level*

In the last step of our analysis, we investigated the relevance of certain linguistic domains and feature combinations for the identification of individual proficiency level changes with increasing proficiency. For this, we compared the performance of all classifiers for each individual proficiency level in terms of precision, recall, and f1 score as displayed in Table 4.

| | A1 | | | A2 | | | B1 | | | B2 | | | C1/C2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 | Pre | Rec | F1 |
| Maj | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 32.0 | 100 | 48.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| All | 45.7 | 36.8 | 40.8 | 68.9 | 71.6 | 70.2 | 66.3 | 65.3 | 65.8 | 72.1 | 76.8 | 74.4 | 38.7 | 26.1 | 31.2 |
| 150 | 60.0 | 36.8 | 45.7 | 72.4 | 75.5 | 73.9 | 68.3 | 65.9 | 67.1 | 70.4 | 86.0 | 77.4 | 50.0 | 2.2 | 4.2 |
| LU | 100 | 1.8 | 3.4 | 63.5 | 63.7 | 63.6 | 52.7 | 56.2 | 54.4 | 62.1 | 78.8 | 69.5 | 0.0 | 0.0 | 0.0 |
| HLP | 0.0 | 0.0 | 0.0 | 58.5 | 61.8 | 60.1 | 47.2 | 51.1 | 49.1 | 56.0 | 67.2 | 61.1 | 0.0 | 0.0 | 0.0 |
| Disc | 66.7 | 3.5 | 6.7 | 66.1 | 73.2 | 69.5 | 63.0 | 54.1 | 58.2 | 64.6 | 89.8 | 75.1 | 0.0 | 0.0 | 0.0 |
| Cla | 50.0 | 7.0 | 12.3 | 63.5 | 75.2 | 68.9 | 59.7 | 57.4 | 58.6 | 68.1 | 80.2 | 73.7 | 0.0 | 0.0 | 0.0 |
| Phr | 33.3 | 1.8 | 3.3 | 62.7 | 74.8 | 68.3 | 56.9 | 55.9 | 56.4 | 66.8 | 77.5 | 71.7 | 0.0 | 0.0 | 0.0 |
| Lex | 100 | 15.8 | 27.3 | 68.6 | 78.4 | 73.2 | 65.3 | 59.2 | 62.1 | 67.6 | 86.3 | 75.9 | 0.0 | 0.0 | 0.0 |
| Mor | 0.0 | 0.0 | 0.0 | 58.7 | 68.3 | 63.1 | 53.6 | 54.5 | 54.0 | 67.1 | 77.8 | 72.0 | 0.0 | 0.0 | 0.0 |

*Table 4. Precision, recall, and f1 score for homogeneous SMO models*

Since the majority baseline labels all texts as level B2, its performance scores are zero for all but this level. More interestingly, none of the linguistically homogeneous models correctly identifies a single instance of C1/C2 level writings. Only the linguistically diverse models correctly classify some of these texts. The classifier using all measures even outperforms *IG 150* for this level, but worse on all others. This might indicate that there are some relevant features for the distinction of B2 and C1/C2 level writing, which were not elicited by the data-driven feature selection due to the under-representation of C1/C2 level texts. Potentially for similar reasons, for level A1 only very few texts are correctly classified, which results in extremely high precision scores with very low recall. The highest f1 (= 27.3%) score is achieved by the lexical complexity model. In contrast, HLP and morphological measures do not classify any text as A1.

For levels A2, B1, and B2 classification performance is better, although levels A2 and B2 exhibit systematically higher recall than precision values. An investigation of the misclassifications confirmed that these are predominantly due to the incorrect labeling of A1 and C1/C2 texts. Across these three levels, the lexical complexity model again performs best in terms of f1 score. Furthermore, clausal and phrasal complexities as well as discourse are considerably more successful at identifying texts at these levels than the other feature groups. For level B2 texts, the discourse model performs comparable to the lexical model. In contrast, HLP is the least informative feature group across all levels. This might be due to the limited diversity within this feature group, which predominantly consists of differently weighted instances of DLT integration cost measures. Language use also performs relatively poorly across

these levels. Morphological complexity shows little discriminatory power for most proficiency levels. Interestingly, however, it is as successful as phrasal and clausal complexities for the B2 level. The high relevance of discourse and morphological measures for B2 level texts is remarkable. Since the distinction between B2 and C1/C2 fails for the homogeneous models, this shows that the morphological and discourse models are able to learn a systematic distinction between B2 and the lower levels, while the discourse, clausal, phrasal, and lexical models identify systematic differences across levels A2 to B2.

## 6. Outlook

Our study clearly demonstrates the benefits of modeling L2 proficiency using broad, linguistically diverse feature selections and yields interesting insights regarding performance differences of linguistically homogeneous classifiers at individual proficiency levels. Yet, it also raised some issues that could only briefly addressed in our current study and that need further investigation. In particular, correctly identifying level A1 and level C1/C2 texts remains a challenge to all presented classifiers. This issue is partially due to the lack of data support for these proficiency levels in our data set. In a follow-up study, the extent to which our results are influenced by this artifact of the data distribution in Merlin needs to be investigated by additional analyses on more balanced data subsets.

Furthermore, there are two additional potential influences on our results that might also contribute to this issue and which we are currently addressing in ongoing studies. On the one hand, our models do not account for nonlinear development of individual measures. This work presented a view on various feature groups to illustrate the relevance of broad language modeling for proficiency assessment and to identify differences in the overall impact of linguistic domains on proficiency assessment. Building on this, we have moved on to the analysis of individual, potentially nonlinear measures by training Generative Additive Models (GAMs) on the Merlin data. In Weiss (2017), we present first results of this approach, where we closely model 13 complexity measures from all our feature groups including nonlinear developments. The results show an improvement in the classification of levels A1 and C1/C2 compared to the models presented here.

On the other hand, the Merlin data entails a particularly broad task background with three elicitation tasks per test level. These may cause task effects in the linguistic properties of the learner texts, in particular with regard to their

complexity, as earlier research on task effects and language performance has shown (Alexopoulou *et al.* 2017; Yoon & Polio 2016; Tracy-Ventura & Myles 2015). Thus, we broaden our investigation of the applicability of diverse complexity features to their sensitivity to task effects in learner corpora, thus shifting the focus of our analysis from learners to tasks. We have analyzed Merlin's elicitation tasks for various functional and cognitive task factors and performed first analyses on the effect these factors have on individual complexity measures as well as feature groups (Weiss 2017). Our preliminary results show that some complexity measures and groups seem to be sensitive to task effects to varying degrees: morphological complexity, for example, is particularly susceptible to task effects, while human language processing features seem to be remarkably robust (see Weiss 2017 for details). Both of these analysis strands have already yielded promising results and are currently pursued further.

## 7. Conclusion

We investigated to which extent broad linguistic modeling is beneficial for German L2 proficiency assessment. For this, we automatically extracted 400 measures of linguistic complexity from various linguistic domains with an elaborate NLP pipeline. We focused on comparing feature groups. On the one hand, we combined features from various linguistic domains in a data-driven approach. On the other, we grouped features together from the same linguistic domain. We compared them in terms of their ability to successfully distinguish between five holistic CEFR proficiency scores assigned to German L2 writings (A1 to C1/C2) when employed in SMO classifiers. Our results show that a broad selection of features that integrates aspects of language as a system, language use, and human sentence processing costs, results in higher classification performance on language learner data. In particular, lexical variation, sentential elaboration, phrasal elaboration and variation, and discourse elaboration are highly beneficial, as an analysis of the overall most informative measures in terms of information gain showed.

In a second step, we investigated to which extent the relevance of certain linguistic domains for the identification of individual proficiency levels changes with increasing proficiency. For this, we compared the performance of the classifiers assessing certain linguistic domains for identifying each individual proficiency level. This showed that lexical, clausal, and phrasal complexity are informative for the identification of several proficiency levels. In contrast, morphological and discourse measures are mostly relevant for distinguishing

B2 from lower proficiency levels. Human language processing and language use features are less successful, although we found individual measures from both groups to be highly informative and included in the classifier using features from various domains. In this analysis, too, the combination of features outperformed all linguistically homogeneous models across individual proficiency levels. Overall, our results show that broad linguistic modelling is beneficial and feasible for German L2 proficiency assessment, even when applied to non-normalized data.

## References

Abel, A., Wisniewski, K., Nicolas, L., Boyd, A., Hana, J. & Meurers, D. (2014). A trilingual learner corpus illustrating european reference levels. *Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne* 2(1), 111-126.

Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1), 180-208.

Barzilay, R. & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics* 34, 1-34.

Biber, D., Gray, B. & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1), 5-35.

Bohnet, B. & Nivre, J. (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*, 1455-1465.

Bulté, B. & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26, 42-65.

Chen, X. & Meurers, D. (2016). CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113-119.

Crossley, S., Kyle, K., Allen, L., Guo, L. & McNamara, D. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment* 7(1).

Crossley, S. & McNamara, D. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2), 115-135.

Crossley, S. & McNamara, D. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26, 66-79.

Ellis, R. & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

François, T. & Fairon, C. (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*, 466-477.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita & W. O'Neil (eds) *Image, Language, Brain*. Cambridge: MIT Press, 95-12.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* 11(1), 10-18.

Hancke, J., Vajjala, S. & Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. *Proceedings of COLING 2012: Technical Papers*, 1063-1080.

Hancke, J. & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *Book of Abstracts of the Learner Corpus Research Conference 2013, Universitetet i Bergen, Norway*, 54-56.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A. & Kliegl, R. (2011). dlexDB - Eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau* 62(1), 10-20.

Henrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 2228-2235.

Housen, A., Vedder, I. & Kuiken, F. (eds). (2012). *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Amsterdam & Philadelphia: John Benjamins.

Lavalley, R., Berkling, K. & Stüker, S. (2015). Preparing children's writing database for automated processing. In *Language Teaching, Learning and Technology (LTLT-2015), Leipzig, 4 September, 2015*, 9-15.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474-496.

McNamara, D., Crossley, S. & McCarthy, P. (2009). Linguistic features of writing quality. *Written Communication* 27(1), 58-86.

Norris, J. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555-578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492-518.

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (eds) *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: de Gruyter, 127-155.

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 1-25.

Petrov, S. & Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL-HLT 2007, Rochester, 22-27 April, 2007*, 404-411.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Reports MSR-TR-98-14*.

Rafferty, A. & Manning, C. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL Workshop on Parsing German (PaGe-08), Columbus, 20 June, 2008*, 47-54.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E. & Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 49-58

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. *TESOL Quarterly* 42(2), 420-430.

Tono, Y. (2013). Criterial feature extraction using parallel learner corpora and machine learning. In A. Díaz-Negrillo, N. Ballier & P. Thompson (eds)

*Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam & Philadelphia: John Benjamins, 169-204.

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

vor der Brück, T., Hartrumpf, S. & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica* 32, 429-435.

Weiss, Z. (2017). *Using measures of linguistic complexity to assess German L2 proficiency in learner corpora under consideration of task-effects.* Master thesis, University of Tübingen. http://www.sfs.uni-tuebingen.de/~zweiss/ma-thesis/ weiss2017-distr.pdf (last accessed on 26 September, 2018).

Yoon, H.-J. & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly* 51(2), 275-301.

Ziai, R. & Meurers, D. (2018). Automatic focus annotation: Bringing formal pragmatics alive in analyzing the information structure of authentic data. In *Proceedings of NAACL-HLT 2018*, 117-128.