Exploring CEFR classification for German based on rich linguistic modeling

Julia Hancke, Detmar Meurers
Universität Tübingen
{jhancke,dm}@sfs.uni-tuebingen.de

The issue The Common European Framework of Reference for Languages (CEFR) has gained a leading role as an instrument of reference for the certification of language proficiency. At the same time, there is increasing interest in a more comprehensive empirical characterization of the relevant linguistic properties of the CEFR levels.

The research reported on in this paper approaches this issue by studying which linguistic properties reliably support the classification of short essays in terms of CEFR levels. Complementing the work on English criterial features and learner language characteristics that is starting to emerge (Hawkins & Buttery 2010; Yannakoudakis et al. 2011), we focus on identifying learner language characteristics of different levels of German proficiency.

Corpus used The empirical basis of our research consists of 1027 professionally rated free text essays from CEFR exams taken by second language learners of German. Each exam level (A1 to C1) is represented by about 200 texts, varying between 8 and 366 words in length (mean length of 121 words). The data has been collected by the project *MERLIN – Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context* (http://merlin-platform.eu).

Features explored We defined a broad set of 3821 features which can be automatically identified using current NLP tools. We primarily use complexity measures from Second Language Acquisition research to model lexical and morphological richness and syntactic sophistication:

At the ***lexical level***, we started by adapting the features discussed for English by Lu (2012) and McCarthy & Jarvis (2010) for German. To measure the depth of lexical knowledge, we implemented a number of features suggested by Crossley et al. (2011). We extracted frequency scores from the lexical database dlexDB (http://dlexdb.de). We computed features of lexical relatedness using GermaNet 7.0 (http://www.sfs.uni-tuebingen.de/lsd), a lexical-semantic resource for German, similar to WordNet (Miller 1995) for English. We added shallow measures of spelling errors in terms of the number of content word types not found in dlexDB and the misspelled words found by Google Spell Check (version 1.1, https://code.google.com/p/google-api-spelling-java).

Our ***morphological features*** for German capture the learner's use of mood, case, and word formation. We automatically extracted tense patterns from the RFTagger (Schmid & Laws 2008) output and included frequency ratios of these pattern as features for our classifier. The tense features might allow more detailed insights into the tenses the learners used at each of the levels.

At the ***syntactic level***, our features are mostly inspired by the measures used for the analysis of syntactic complexity in English (Lu 2010). However, German syntactically differs from English in several relevant respects. For example, German allows subjectless sentences. Thus, while in general the intention behind the English SLA complexity measures can be expressed in terms of the German syntactic structure and categories, the process of adapting and defining syntactic complexity features for German is far from trivial. As basic syntactic vocabulary for German, we made use of the Negra treebank annotation scheme (Skut et al. 1997).

We added ***dependency-based features*** of syntactic complexity that were previously used in second language writing assessment (Yannakoudakis et al. 2011) and readability assessment (Vor der Brück & Hartrumpf 2007; Vor der Brück et al. 2008; Dell'Orletta et al. 2011).

We automatically extracted parse tree rules from the parse trees produced by the Stanford Parser, following Briscoe et al. (2010) and Yannakoudakis et al. (2011), who used a similar feature based on the output of the RASP parser. We used frequency ratios of these parse tree rules as features for our classifier.

Complementing the linguistic syntactic analysis, we also implemented a number of ***shallower language features***. Unigram, bigram and trigram language model scores provide statistical comparisons to a linguistically simpler model based on a news website for children (http://news4kids.de) and a more complex model based on a news website for adults (http://www.n-tv.de).

NLP tools used To automatically identify the lexical, morphological, and syntactic features, we employ a range of NLP tools and resources including Apache OpenNLP (http://opennlp.apache.org), RFTagger (Schmid & Laws 2008), the Stanford Parser (Rafferty & Manning 2008) with the standard German model trained on the NEGRA corpus (http://coli.uni-saarland.de/projects/sfb378/negra-corpus), the SRILM Language Modeling Toolkit (Stolcke 2002), and the lexical database dlexDB (http://dlexdb.de). For dependency parsing we used the *MATE* dependency parser (Bohnet 2010), with the standard model for German (Seeker & Kuhn 2012) trained on the *TIGER* corpus. Before tagging and parsing, a Java API for Google Spell Check was used to reduce problems caused by spelling errors.

Experimental setup On the basis of the 3821 automatically derived features, we trained a classifier using the Sequential Minimal Optimization (SMO) Algorithm as implemented in the *WEKA* toolkit (Hall et al. 2009). We split the dataset into a training and test set by randomly assigning 2/3 of the samples from each class to the training set (721 samples) and 1/3 to the test set (306 samples). As an additional method for evaluation we used ten-fold cross-validation on the whole dataset.

Results The following table provides an overview of the performance of the classifier for the five level (A1–C1) CEFR classification task:

	Accuracy on test set	Crossvalid. on all data
Random baseline	20%	
Majority baseline	32.9%	33.0%
SMO (all features)	57.2%	64.5%
SMO (best features)	62.7%	

The classifier trained with all features achieves an accuracy of 57.2% with the separate training and test set and an accuracy of 64.5% when using cross-validation on all data. Compared to a majority baseline of classifying all samples as the largest class, this is an improvement of 24.3% and 31.5% respectively.

Investigating the notable difference between the test set and the cross-validation results, we identified two issues. Looking at the results of each individual cross-validation fold revealed that there is considerable variance in the results (10.7% between the best and worst performing fold). However, the worst cross-validation fold still had a better result than our test set. This could be an effect to the slightly larger amount of training data available in the cross validation procedure. Another reason for the comparably poor performance on our test set could lie in the uneven distribution of exam types (as opposed to essay grades) across the different CEFR levels.

Examining the performance of individual feature groups with holdout estimation revealed that the lexical (60.5%) and morphological (56.8%) features were the most successful predictors of the

CEFR level. The syntactic features and language modeling scores were not very successful predictors taken on their own (53.6% and 50.0%), but the syntactic features clearly improved the classification in combination with other features groups. Parse rule features and tense features were the least predictive feature groups (49.0% and 38.5%), however, further experiments showed that their indicative power improves when they are encoded as binary instead of as frequency-based features.

The best model was obtained by combining all feature groups and using WEKA's *CfsSubsetEval*, a correlation-based method for feature selection. It included a set of 34 features consisting of syntactic, lexical, language model and morphological indicators and resulted in a classification accuracy of 62.7% on the test set.

References

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Beijing, China, pp. 89–97.

Briscoe, T., B. Medlock & O. Andersen (2010). *Automated assessment of ESOL free text examinations*. Tech. rep., University of Cambridge Computer Laboratory.

Crossley, S. A., T. Salsbury, D. S. McNamara & S. Jarvis (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing* 28, 561–580.

Dell'Orletta, F., S. Montemagni & G. Venturi (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. pp. 73–83.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Hawkins, J. A. & P. Buttery (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* .

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

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

McCarthy, P. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392. URL https://serifos.sfs.uni-tuebingen.de/svn/resources/trunk/papers/McCarthy.Jarvis-10.pdf.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM* 38(11), 39–41. URL http://aclweb.org/anthology/H94-1111.

Rafferty, A. N. & C. D. Manning (2008). Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*. Stroudsburg, PA, USA: Association for Computational Linguistics, PaGe '08, pp. 40–46. URL http://dl.acm.org/citation.cfm?id=1621401.1621407.

Schmid, H. & F. Laws (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *COLING '08 Proceedings of the 22nd International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, vol. 1, pp. 777–784. URL http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/COLING08/Schmid-Laws.pdf.

Seeker, W. & J. Kuhn (2012). Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation, 3132–3139. Istanbul, Turkey: European Language Resources Association (ELRA)*.

Skut, W., B. Kreen, T. Brants & H. Uszkoreit (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fith Conference on Applied Natural Language*. Washington, D.C. URL http://www.coli.uni-sb.de/publikationen/softcopies/Skut:1997:ASF.pdf.

Stolcke, A. (2002). SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*. Denver, USA, vol. 2, pp. 901–904. URL http://www.speech.sri.com/cgi-bin/run-distill?papers/icslp2002-srilm.ps.gz.

Vor der Brück, T. & S. Hartrumpf (2007). A semantically oriented readability checker for German. In Z. Vetulani (ed.), *Proceedings of the 3rd Language & Technology Conference*. Poznań, Poland: Wydawnictwo Poznańskie, pp. 270–274. URL http://pi7.fernuni-hagen.de/papers/brueck_hartrumpf07_online.pdf.

Vor der Brück, T., S. Hartrumpf & H. Helbig (2008). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. *Informatica* 32(4), 429—-435.

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, HLT '11, pp. 180–189. URL http://aclweb.org/anthology/P11-1019.pdf. Corpus available: http://ilexir.co.uk/applications/clc-fce-dataset.