Exploring CEFR classification for German based on rich linguistic modeling

Julia Hancke Detmar Meurers
Universität Tübingen

Learner Corpus Research Conference (LCR 2013)
Bergen, Norway. September 27–29, 2013

Introduction

- The Common European Framework of Reference for Languages (CEFR) is an increasingly used standard for
  - characterizing the foreign language ability of a learner
  - based on functional abilities to use language in different domains (public, private, occupational, etc.).
- But there is a lack of
  - authentic learner data illustrating CEFR levels and
  - insight into the precise linguistic characteristics correlating with the proficiency levels.

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
Eberhard Karls Universität Tübingen

Introduction

Towards addressing the desiderata

- MERLIN is creating a learner corpus with CEFR-rated essays for German, Italian & Czech (Abel et al. 2013).
- How can we explore the impact of different aspects of linguistic modeling on the CEFR classification?

⇒ Use machine learning to quantify the value of different linguistic features for automatic proficiency classification.

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
Eberhard Karls Universität Tübingen

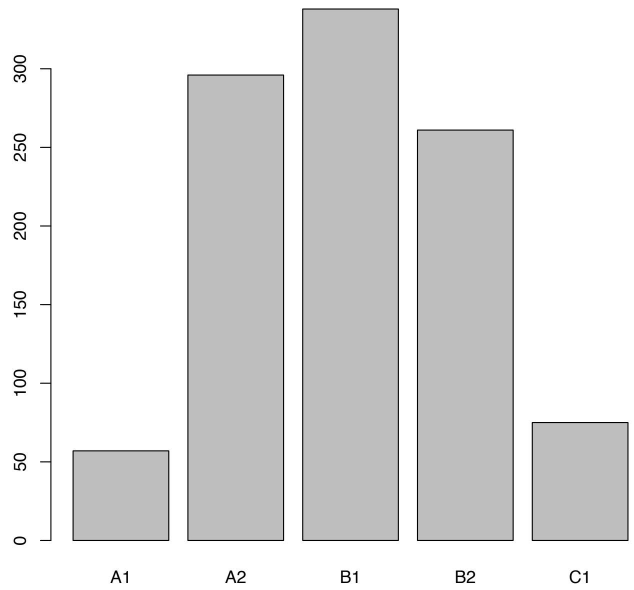
Data used: German portion of MERLIN corpus

- 1027 German learner texts
  - about 200 texts per exam type (A1–C1)
  - range of lengths (6–366 words) with average 122 words
  - texts also vary in other parameters:
    - written for different tasks (one of three tasks per level)
    - written by learners with different native languages (> 12)
- Each text was graded in terms of CEFR levels
  - by multiple trained human raters at TELC, a major language test provider in Germany
  - reliability of ratings externally validated (Univ. Leipzig)
  - most common rating: B1

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
Eberhard Karls Universität Tübingen

Distribution of Ratings over CEFR levels

Number of texts per essay rating level



Features to be investigated

- Goal: richer linguistic modeling of CEFR levels
  - ⇒ explore potentially relevant language features
  - ⇒ test their impact on predicting CEFR class of each essay
- We explored:
  - lexical features
  - syntactic features
    - statistical language model
    - constituency-based
    - dependency-based
  - morphological features

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features explored

Lexical features

- Lexical density (Lu 2012)
  - ratio of number of lexical words to total number of words
- Lexical diversity:
  - TTR variants, MTLD, lexical word variation (McCarthy & Jarvis 2010; Crossley et al. 2011a; Lu 2012)
- Depth of lexical knowledge
  - lexical frequency scores (Crossley et al. 2011b)
- Lexical relatedness
  - hypernym & polysemy scores (Crossley et al. 2009)
- Shallow measures
  - spelling errors per number of words, word length

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features explored

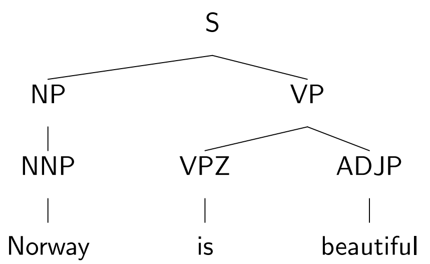
Syntactic features: 1. Statistical Language Models

- inspired by readability assessment research (Schwarm & Ostendorf 2005; Petersen & Ostendorf 2009; Feng 2010)
- used SRILM Language Modeling Toolkit (Stolcke 2002)
- trained on two data sets (Hancke, Meurers & Vajjala 2012)
  - **easy**: 2000 texts, German kid news website *News4Kids*
  - **hard**: 2000 texts, German news channel *NTV* website
- 12 features: unigram, bigram and trigram perplexity for
  - *easy* or *hard* text models based on
  - *word* or *mixed (word+POS)* representations

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features explored

Syntactic features: 2. Data-driven constituency features

- Is the frequency of common rules characteristic?
  (Briscoe et al. 2010; Yannakoudakis et al. 2011)
- Extracted all rules in the parse trees assigned by Stanford Parser in 700 articles from the *NTV* corpus
- 

  S → NP VP
  NP → NNP
  VP → VPZ ADJP
- Given a learner text, for each rule, we use as feature:
  rule frequency in text / number of words in text

Features explored

Syntactic features: 3. Theory-driven constituency features

(Hancke, Meurers & Vajjala 2012)

Syntactic properties assumed to be characteristic of complexity or difficulty in SLA proficiency and readability research:

- number and length of
  - clauses, sentences, T-units
  - NPs, VPs, PPs
- dependent clauses and coordinated phrases
  - per clause, sentence, T-unit
- interrogative, relative, conjoined clause ratios
- nonterminals per sentence
- parse tree height

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features explored

Syntactic features: 4. Theory-driven dependency features

(Vor der Brück et al. 2008; Yannakoudakis et al. 2011; Dell'Orletta et al. 2011)

Linguistic properties based on dependency analysis used in SLA proficiency and readability assessment research:

- number of words between head and dependent
  - maximum
  - average number per sentence
- avg. number of dependents per verb (in words)
- number of dependents per NP (in words)

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features explored

Morphological features

- Word Formation
  - ratios of nominal suffixes (*-ung*, *-heit*) and compounds
- Inflectional Morphology
  - of verb: person, mood, verb-form (participle, infinitive)
  - of noun: case
- Tense:
  - frequency ratios of verbal tense features
  - data-driven, based on 700 texts from *NTV* corpus

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

NLP used for automatic feature identification

- Preprocessing
  - sentence segmentation, tokenization (Apache OpenNLP)
  - spelling correction (Java API for Google Spell Check)
- Lexicon
  - lexical semantic relations (GermaNet, Hamp & Feldweg 1997)
  - lexical frequencies (dlexDB, http://dlexdb.de)
- Part-of-Speech Tagging
  - POS and lemmatization (TreeTagger, Schmid 1995)
  - fine-grained POS (RFTagger, Schmid & Laws 2008)
- Parsing
  - constituents (Stanford PCFG Parser, Rafferty & Manning 2008)
  - dependencies (*MATE*, Bohnet 2010)

Experimental Setup

- We divided the MERLIN data into
  - training set (721 essays)
  - test set (302 essays)
- We classify into five CEFR classes (A1, A2, B1, B2, C1).
- We use the WEKA machine learning toolkit (Hall et al. 2009) for classification, specifically
  - SMO to train support vector machines (linear kernel)
- Many further experiments → Hancke (2013)

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Performance of different feature groups

Name	#	Accuracy (%)
Random Baseline	-	20.0
Majority Baseline	-	33.0
TENSE	230	38.5
ParseRules	3445	49.0
LanguageModel	12	50.0
SYN	47	53.6
MORPH	41	56.8
LEX	46	60.5

- Informative – but for this data set:
  - Text Length as a single feature: 61.4% accuracy

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Feature Groups Combinations

The best two, three, and four class combinations:

Name	Accuracy
LEX_MORPH	61.1
LEX_TEN	59.8
LEX_LM	59.4
LEX_LM_MORPH	61.1
SYN_LEX_MORPH	58.5
LEX_LM_TEN	57.8
SYN_LEX_LM_MORPH	58.8
SYN_LEX_LM_PR	57.8
LEX_LM_MORPH_TEN	57.8
ALL Features	57.2

- not particularly exciting, but lexical features help

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Feature Selection

- How can we identify the best features?
- The features we use are not independent, so taking the best features using Information Gain is problematic.
- *CfsSubsetEval*: correlation-based feature selection
  - Features that correlate highest with the class but have a low inter-correlation are preferred (Witten & Frank 2005).
- Results:

Name	#	Accuracy
CfsSubsetEval(LEX_LM_MORPH)	30	61.7
CfsSubsetEval(SYN_LEX_LM_MORPH)	34	62.7
CfsSubsetEval(ALL)	88	61.8

Qualitative analysis of the 34 selected features
Syntax

- sophistication of production units
  - avg. sentence length, length of a t-unit
- embedding
  - dep. clause with conj. to dep. clause ratio
- verb phrase complexity
- coordination
- passive voice
- text length

Qualitative analysis of the 34 selected features
Lexicon

- spelling errors
- lexical richness (TTR, MTLD)
- verbal/nominal style (verb variation, noun token ratio)
- lexical sophistication (frequency, easy unigrams, length)

- but: no lexical relatedness features were selected

Qualitative analysis of the 34 selected features
Morphology

- use of derivation (derived nouns/nouns, specific suffixes)
- nominal case (genitive, nominative)
- verbal mood and person (subjunctive, 2. person forms)

Summary

- Automatic proficiency classification: a useful experimental sandbox for exploring the role of linguistic modeling
- Quantitatively difficult but possible to outperform the very high text-length baseline on the new MERLIN corpus.
- Qualitatively insightful analysis of features is feasible.
  - Feature selection helps improve classification results and identify qualitatively interpretable feature groups.
- Outlook:
  - reliable sentence segmentation for learner language needed, crucial for many complexity features
  - analyze impact of learner errors on such analyses, possible using target hypotheses
  - principled exploration of variationist linguistic features (→ talk on Saturday with Julia Krivanek)

References

Abel, A., L. Nicolas, J. Hana, B. Štindlová, S. Bykh & D. Meurers (2013). A Trilingual Learner Corpus illustrating European Reference Levels. In K. Tenfjord, A. Golden, F. Meunier & K. D. Smedt (eds.), *Learner Corpus Research 2013. Book of Abstracts*. Bergen, pp. 3–5. URL http://lcr2013.b.uib.no/files/2013/09/abstracts-book.pdf.

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., T. Salsbury & D. McNamara (2009). Measuring L2 Lexical Growth Using Hyperniymic Relationships. *Language Learning* 59, 307–334.

Crossley, S. A., T. Salsbury & D. S. McNamara (2011a). Predicting the proficiency level of language learners using lexical indices. In *Language Testing*.

Crossley, S. A., T. Salsbury, D. S. McNamara & S. Jarvis (2011b). 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.

Feng, L. (2010). Automatic Readability Assessment. Ph.D. thesis, City University of New York (CUNY). URL http://lijun.symptotic.com/files/thesis.pdf?attredirects=0.

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.

Hamp, B. & H. Feldweg (1997). GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid. URL http://aclweb.org/anthology/W97-0802.

Hancke, J. (2013). Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master's thesis, International Studies in Computational Linguistics. Seminar für Sprachwissenschaft, Universität Tübingen.

Hancke, J., D. Meurers & S. Vajjala (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080. URL http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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.

Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level assessment. *Computer Speech and Language* 23, 86–106.

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. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. URL http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf.

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.

Schwarm, S. & M. Ostendorf (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Ann Arbor, Michigan, pp. 523–530.

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

Witten, I. H. & E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam; Boston, MA: Morgan Kaufmann, 2nd ed.

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.

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Qualitative analysis of selected features

Detailed Syntax

Interpretation	*Features*
sophistication of production units	avg. sentence length, avg. length of a t-unit
embedding	dep. clauses with conj. to dep. clause ratio, avg. num. non-terminal per words
verb phrase complexity	avg. num. VZs per sentence, avg. length of a VP
coordination	avg. num. co-ordinate phrases per sentence
passive voice	passive voice to sentence ratio
script length	text length

CEFR classification for German
Julia Hancke
Detmar Meurers
Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary
merlin
Lifelong Learning Programme
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Qualitative analysis of selected features

Detailed Lexicon

Interpretation	*Features*
lexical richness	type-token ratio, root type-token ratio, corrected type-token ratio, HDD, MTLD
lexical richness w. respect to verbs	squared verb variation 1, corrected verb variation 1
nominal style	noun token ratio
word length / difficulty	avg. num. syllables per word, avg. num. characters per word
lexical sophistication	annotated type ratio, unigram plain easy ratio of words in log frequency band two, ratio of words in log frequency band four
spelling errors	ratio of lex. types not in Dlex, Google spell check error rate

Qualitative analysis of selected features

Detailed Morphology

Interpretation	*Features*
nominalization, use of derivational suffixes and words with Germanic stems	*-keit*, *-ung*, *-werk*, derived nouns to nouns ratio
nominal case	genitive-noun ratio, nominative-noun ratio
verbal mood and person	subjunctive-verb ratio, second person-verb ratio, third person-verb ratio

CEFR classification for German
Julia Hancke
Detmar Meurers

Introduction
Data
Features
Lexical
Syntactic
Language Model
Constituency
Dependency
Morphological
NLP used for feature identification
Experimental setup
Results
Individual Feature Groups
Feature Groups
Feature Selection
Qualitative feature analysis
Summary