Predicting CEFRL levels in learners of English on the basis of metrics and full texts

Taylor Arnold[1], Nicolas Ballier[2], Thomas Gaillat[3], Paula Lissón[2]
[1]University of Richmond, [2]Université Paris Diderot, [3]Insight Centre for Data Analytics, NUI Galway

Insight Centre for Data Analytics
A world leading SFI Research Centre

Introduction

Motivation:
Requirements in language learning:
- to provide regular level evaluations to both learners and teachers
- to group learners homogeneously in order to set adequate teaching objectives and methods.

Our proposal: error-independent approach to compute levels

Task:
To classify French learners of English according to levels of the Common European Framework of Reference for Languages (CEFRL)

Purpose:
To build a model for the prediction of learner levels as a function of language complexity features.

Data

Corpus:
- French-L1 subset of EFCAMDAT (Ref)
- Manually annotated with the six CEFRL levels (A1 to C2)
- 41,626 texts (approx. 3,298,343 tokens),
- 128 essay topics
- 7,695 French learners

Dataset:
- Automatically computed language metrics based on PoS-tagged and parsed texts (REf to Liu and Korpus)
- Matrix of complexity metrics and CEFRL-related classes:

	A1	A2	B1	B2	C1	C2
Train set	8,436	5,489	3,819	1,715	330	33
Test set	9,169	6,095	4,286	1,799	412	43

fulltext	sentences	words	letters.all	syllables	punct	avg.sentc.length	avg.word.length	avg.syll.word	sntc.per.word
Hi!My name's Karine. I'm from France. I live in Les Ar…	9	56	208	74	23	6.222222	3.714286	1.321429	0.16071429

fulltext	sentences	words	letters.all	syllables	punct	avg.sentc.length	avg.word.length	avg.syll.word	sntc.per.word
The Eiffel Tower is the symbol of France and one of t…	8	155	732	248	13	19.375000	4.722581	1.600000	0.05161290

2.Features

Lexical diversity metrics

- Type token relationship, e.g. TTR
- Type factor ratios, e.g. MTLD

Readability metrics

- Text difficulty
- Average number of words/sentence combined with average word length, e.g. ARI, LIX, Flesh-Kincaid

Morpho-syntactic features

- Bag of words with word counts
- PoS frequency per text
- Syntactic dependency relations
- Zipf-scale category

Syntactic complexity metrics

- Syntactic measures : # of T-Units, Words, Sentences, VP, NP etc. (Lu, 2011)
- Indices (based on T-Unit): MLS, MLT

Machine learning models

Supervised-learning approach
Adjacent classes - 5 pairwise models
Tested models:
- Gradient Boosted Trees for model explanation
- Elastic Net
- LSTM

Discussion

- Modeling issues: skewed distribution of the data

- Coarse-grained vs. fine-grained POS tags:

More subtle tagsets (i.e. Penn Treebank tagset, punctuation) possibly yield more fine-grained results at the expense of precision: more relevant but more error-prone

- L1 and level-specific features:

Features based on potential errors made by a given population of learners

- Lexically based metrics:

Variations in word spelling impact metrics based on number of tokens

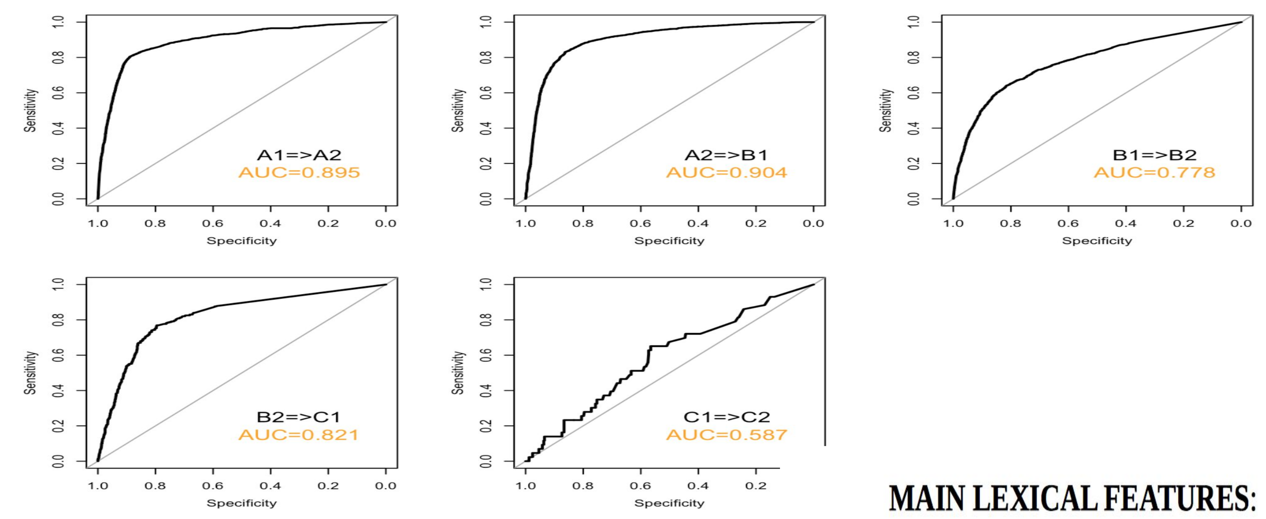
Conclusions and Future Directions

- Binary classification: [A1,B1] & [B2,C2] to reflect main distinction in third level education requirements
- Use of Penn Treebank Tagset
- Semantic and discourse-based metrics

Results

Features	Model	Partition	A1=>A2	A2=>B1	B1=>B2	B2=>C1	C1=>C2
Metrics	GBT	train	0.897	0.888	0.903	0.854	0.948
Metrics	GBT	test	0.895	0.897	**0.778**	**0.821**	**0.587**
Term Freq.	Elastic Net	train	0.972	0.977	0.895	0.998	0.949
Term Freq.	Elastic Net	test	0.865	0.847	0.686	0.629	0.550
Word Seq.	LSTM	train	0.985	0.985	0.942	0.777	0.634
Word Seq.	LSTM	test	0.863	0.824	0.548	0.525	0.535
Metrics+	GBT	train	0.931	0.899	0.927	0.860	0.853
Metrics+	GBT	test	**0.916**	**0.904**	0.753	0.746	0.558

TABLE 1 – AUC Metrics for binary classification models across four methods.



A1=>A2	A2=>B1	B1=>B2	B2=>C1	C1=>C2
wordtokens	wordtypes	wordtokens	wordtokens	ndwerz
W	W	W	adjv	DC.C
svv1	ls2	DC.C	vs1	slextypes
wordtypes	MLC	vs2	swordtypes	MLC
MLS	CN	lextokens	W	lv
DC.T	CN.T	ttr	CN.T	modv

TABLE 2 – Most relevant metrics for pairwise level distinctions

MAIN LEXICAL FEATURES:

The CAp2018 challenge metrics + Lexical Complexity Analyzer (Lu 2012)
adjv: index measuring variation for adjectives
ls2: proportion of lexemes among the most frequent words (Laufer & Nation 1994)
lv: index of lexical variation
ndwerz: index based on number of tokens in 50 token samples
swordtype: index based on grammatical categories
ttr: type to token ratio
vs1: ratio of words that are not among the most frequent

Contact

taylor.arnold@acm.org
nicolas.ballier@univ-paris-diderot.fr
thomas.gaillat@insight-centre.org
paula.lisson@uni-potsdam.de

References

Ballier, N., & Gaillat, T. (2016). Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique (Vol. 9, pp. 1–14). Proceedings of the JEP-TALN-RECITAL 2016
Bentz, C., & Buttery, P. (2014). Towards a computational model of grammaticalization and lexical diversity (pp. 38–42). Presented at the Proc. of 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)@ EACL.
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 Language Journal*, *96*(2), 190–208.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.
Vajjala, S. (2018). Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features. *International Journal of Artificial Intelligence in Education*, *28*(1), 79–105. https://doi.org/10.1007/s40593-017-0142-3