Automatic CEFR Level Prediction for Estonian Learner Text

Sowmya Vajjala[1], *Kaidi Lõo*[2]

(1) LEAD Graduate School, University of Tübingen, Germany
(2) Department of Linguistics, University of Alberta, Canada

`sowmya@sfs.uni-tuebingen.de`, `kloo@ualberta.ca`

Abstract

This paper reports on approaches for automatically predicting a learner's language proficiency in Estonian according to the European CEFR scale. We used the morphological and POS tag information extracted from the texts written by learners. We compared classification and regression modeling for this task. Our models achieve a classification accuracy of 79% and a correlation of 0.85 when modeled as regression. After a comparison between them, we concluded that classification is more effective than regression in terms of exact error and the direction of error. Apart from this, we investigated the most predictive features for both multi-class and binary classification between groups and also explored the nature of the correlations between highly predictive features. Our results show considerable improvement in classification accuracy over previously reported results and take us a step closer towards the automated assessment of Estonian learner text.

KEYWORDS: Estonian, Proficiency Classification, CEFR, Morphological Features, Machine Learning.

Sowmya Vajjala and Kaidi Lõo 2014. Automatic CEFR level prediction for Estonian learner text. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 113–127.

1 Introduction

People learn a foreign language for many reasons like: living in a new country, having a general interest in the language etc., In many of these scenarios, language learners also undertake exams to get certified for their proficiency in a foreign language. Language proficiency is typically measured using some standardized scale like the CEFR (Council of Europe, 2001) in European nations. Evaluating free text responses like essays is one of the standard ways of assessing the language proficiency of a learner. Traditionally, these student essays were evaluated by experienced human graders trained for doing the task. With the ever increasing number of people taking language tests and with the advent of computational tools that can process language, automatic approaches that reduce human grading effort became a standard way to assess language proficiency. Automated essay grading is already being used along with human grading in several assessment exams like Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT). It can also be useful in a placement test that one may take at a language teaching institute before starting to learn a language at a certain level or serve as a guiding tool for language learners in self-assessment. Apart from this, automated approaches can also enable us to identify distinctive features at a proficiency level, thereby providing us with insights about the process of language acquisition.

While automated assessment is an active area of research for English, approaches for the automatic proficiency classification of learner essays according to the European CEFR scale were recently proposed for German (Hancke and Meurers, 2013), Swedish (Östling et al., 2013) and Estonian (Vajjala and Lõo, 2013). In this paper, we focus on the proficiency classification of Estonian learner essays. We started with the feature set described in Vajjala and Lõo (2013) and added more features to the list. We also used a more fine-grained subset of the same base corpus, consisting of four CEFR proficiency levels. We show that our approach improves the overall classification accuracy for this task reaching up to 79% for a four-level classification. We compare this approach with modeling the problem as regression and show that classification performs better in terms of accuracy and the direction of error. Apart from these, to gain a better understanding of the modeling process, we also studied the issues of feature selection, most predictive features for classification between categories and for overall classification, and correlations between features. In sum, we investigate proficiency classification both from a prediction as well as an interpretational perspective.

Rest of this paper is organized as follows: we start with an overview of contemporary research in proficiency classification of free text responses in Section 2 and describe the corpus and features we used in Section 3. Section 4 describes our experimental setup and explains the classification and regression experiments we performed along with our results. Section 5 briefly discusses feature selection and correlational analysis with all the features. Section 6 concludes the paper with pointers to future work.

2 Background

Automated Assessment (AA) of learner essays can be useful either as a method of scoring them by their proficiency or for understanding the distinctive features of language at a given proficiency level. AA has been active area of research in the field of language testing for a few decades now. Several assessment exams like GRE, GMAT that have language proficiency as a component have been using automated assessment systems as one of the scoring methods along with human graders. These AA systems that are primarily developed for English use a wide range of linguistic and structural features to score student essays (e.g., Burstein, 2003;

Zhang, 2008; Williamson, 2009; Burstein and Chodorow, 2010; Yannakoudakis et al., 2011; Crossley et al., 2011).

Apart from these approaches whose primary purpose is to predict the learner proficiency, there have been studies that did a qualitative analysis of distinctive features between proficiency levels in Second Language Acqusition (SLA) literature. Kyle and Crossley (2014) used a range of lexical sophistication indices and showed that the measures explain 47.5% of the variance in holistic scores of lexical proficiency of second language English learners. Characteristics like lexical richness, syntactic complexity, error patterns of learners and other characteristics too were studied in the recent past (e.g., Tono, 2000; Lu, 2010, 2012; Vyatkina, 2012). Although this strand of research is primarily focused on English, recent research has started to focus on other languages as well (Gyllstad et al., 2014).

With the creation of learner corpora in various European languages, automatic approaches for classifying learner essays into various proficiency levels began to emerge. Approaches for morphologically-rich languages also made use of language specific morphological features, which were not explored before in the case of English. Östling et al. (2013) reported on a proficiency classification approach for Swedish based on a corpus of 1,700 learner essays spanning four levels, obtained from the high-school exams conducted at a national level in Sweden. Along with the features like word length, sentence length, POS tag densities and corpus based entropy features, they also used spelling and compound splitting error based features for this task and achieved an overall accuracy of 62% for four level classification. Hancke and Meurers (2013); Hancke (2013) described a proficiency classification approach for German based on European CEFR standards using a broad range of lexical, syntactic and morphological features, considering German language structure into account. For a five level graded corpus with about 200 texts per level, they achieved a classification accuracy of 64.5%.

Vajjala and Lõo (2013) developed a proficiency classification approach for Estonian learner corpus and achieved an accuracy of 66% for a three level (A,B,C) corpus consisting of 250 texts per level, using a collection of POS and morphological features. We extended this work by adding more features and working with a fine-grained corpus spanning four levels on the CEFR scale. Further, we modeled the problem as both classification and regression and compared their performance. We also explored predictive features between categories and inter-feature correlations.

3 Corpus and Features

3.1 Corpus

Our experiments are based on the Estonian Interlanguage Corpus (EIC)[1] released online by Talinn University. It is a corpus of texts written by learners of Estonian as a second or foreign language. Most of the texts are originally obtained from language examinations conducted by various government bodies in Estonia. These texts include essays, personal and official letters and answers to language exercises. The learners come from diverse language backgrounds, although majority of them are native speakers of Russian. The corpus currently consists of around 12,000 documents in total[2]. The grading of this corpus is an ongoing project. In our analysis, we only used a subset of the whole corpus that is currently annotated with proficiency level. The version used in this paper was crawled from the EIC website in July 2014. As the

[1] http://evkk.tlu.ee/

[2] http://evkk.tlu.ee/statistics.html

corpus annotation is still under development, the latest version that is available on the web may have more texts than the version we used[3].

The corpus we used in this paper consists of 879 texts in total, belonging to four proficiency levels A2, B1, B2, C1. Since only one document each was annotated as A1 and C2 respectively in the corpus, we removed those levels from our experiments. According to the CEFR definitions, A2 represents a basic language user, B1 and B2 represent independent user and C1 represents a proficient user.

This corpus was used for Estonian proficiency classification earlier in Vajjala and Lõo (2013). However, we could not use the same version as the assigned grades changed for the current version. In the previous version of the corpus, proficiency levels were estimated based on the meta information collected from teacher or based on the subjective opinion of the data enterer (Eslon, 2014). The current version of the corpus has proficiency levels estimated by three qualified professional graders and divided into six CEFR categories instead of three levels A,B,C as in the older version. Thus, the proficiency level of certain texts got modified in the process and the current version of proficiency annotation of the corpus is considered precise and accurate (Eslon, 2014). However, it has to be noted that we do not have access to the individual grades given by the three graders. We only have the final grade assigned to the text. We also do not have any information about the inter-annotator agreement about the grades per text.

The crawled corpus consisted of HTML documents, which were parsed using HtmlUnit[4] and plain-text of the learner writing was extracted using Xpath expressions. Figure 1 shows some basic statistics about the corpus we used, in terms of the number of texts per category, average and the range for number of words per text. It can be noticed that the corpus is not evenly distributed across all levels. Hence, while modeling as classification, we consider this aspect and report our experiments with two versions of the dataset - one, a balanced version with equal training samples for all categories, the other an unbalanced version. Further, there is a broad range for all the categories. We used a normalized version of the dataset for our experiments.

Proficiency Level	**#Docs**	**Avg. #words**	**Range**
A2-level	196	145.9	[23, 636]
B1-level	384	226.3	[39, 1267]
B2-level	207	368.5	[30, 1749]
C1-level	92	704.1	[180, 4508]

Table 1: The Estonian Interlanguage Corpus

3.2 Preprocessing

These texts were POS-tagged using TreeTragger [5], (Schmid, 1994) a probabilistic part of speech tagger which has Estonian parameter files to tag Estonian data. The tag set was derived from the Tartu Morphologically Disambiguated Corpus tag set[6] and consists of morphological information

[3] We can share our version of the corpus with anyone who wants to replicate the experiments.

[4] http://htmlunit.sourceforge.net

[5] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[6] `http://www.cl.ut.ee/korpused/morfkorpus/`

along with basic POS information. All the features reported in this paper are calculated based on TreeTagger's output.

Table 2 shows an example output from TreeTagger for the sentence *Tänapäeva meediat valitseb suur spekter erinevaid tehnilisi abivahendeid, on arvuti, internet ja selle kõrval ka muu digitaalne kommunikatsioon*, taken from a C1 level essay in the corpus. (The sentence means: "*Today's media is dominated by a large spectrum of different technical tools, computer, internet, in addition to other digital communication.*")

word	**tag**	**lemma**
Tänapäeva	S.com.sg.gen	täna_päev+0
meediat	S.com.sg.part	meedia+t
valitseb	V.main.indic.pres.ps3.sg.ps.af	valitse+b
suur	A.pos.sg.nom	suur+0
spekter	S.com.sg.nom	spekter+0
erinevaid	A.pos.pl.part	erinev+id
tehnilisi	A.pos.pl.part	tehniline+i
abivahendeid	S.com.pl.part	abi_vahend+id
,	Z.Com	,
on	V.main.indic.pres.ps3.sg.ps.af	ole+0
arvuti	S.com.sg.nom	arvuti+0
,	Z.Com	,
internet	S.com.sg.nom	internet+0
ja	J.crd	ja+0
selle	P.sg.gen	see+0
kõrval	K.post	kõrval+0
ka	D	ka+0
muu	P.sg.gen	muu+0
digitaalne	A.pos.sg.nom	<unknown>
kommunikatsioon	S.com.sg.nom	<unknown>
.	Z.Fst	.

Table 2: TreeTagger Output: An Example

As we can see from the table, the output is rich in terms of the morphological information. For example, the tag *A.pos.pl.part* indicates - *Adjective-Positive-Plural-Partitive Case*. More detailed information on what each tag means can be found on the morpho-syntactic categories description for the Tartu Morphologically Disambiguated Corpus[7]. While suffixes are indicated in the lemma column separated by a "+" symbol, compound words are separated by an underscore. For example, in the above sentence, there are two compound words - *Tänapäeva* (täna + päev ⇒ today + day = nowadays) and *abivahend* (abi + vahend ⇒ help + tool = helping tool/aid)

3.3 Features

We started with the feature set described in Vajjala and Lõo (2013) and added a few additional features, primarily lexical richness features from Lu (2012). In total, our feature set consists of

[7] `http://www.cl.ut.ee/korpused/morfliides/seletus`

78 features.

First, several surface features, such as number of words, number of sentences in a document, mean word and sentence length in a document were considered. However, as can be seen from Table 1, the number of words in documents is unequally distributed across the classes. C1 level documents have almost five times more words than A2 on an average. This may create a bias towards this feature. While this is perfectly valid in a predictive algorithm, we wanted to understand how much can the morpho-syntactic features contribute to the task without surface measures. Hence, we excluded these surface features from the feature set. We used sentence length while doing a replication and comparison with previous work though. We also briefly present about the variation in accuracy upon including the surface features, in the results section.

Vajjala and Lõo (2013) described several features considering the morphological complexity of Estonian. These consist of the average number of nouns and adjectives in a text belonging to each of the 15 cases that exist in Estonian, average number of verbs belonging to the five moods (indicative, conditional, imperative, quotative and justive), two tenses (present, past), two voices (personal, impersonal), three persons (first, second, third), polarity (positive, negative) and other morphological features. A description of various declensions and conjugations in Estonian can be found in the Estonian Morphology guide [8]. In addition to all these features, we added the number of compound words per text and number of different cases present in the text as additional features in this category.

Vajjala and Lõo (2013) had some of the lexical variation measures from Lu (2012) (lexical variation, noun, adjective, adverb and modifier variation). We additionally implemented the other lexical richness measures described in Lu (2012) that covered the aspects of lexical density and diversity. Although these formulae are actually for English, since they are only depended on the various word counts and since we are not aware of any equivalent formulae for Estonian, we used the same formulae for Estonian as well. These measures were shown to be good predictors of learner language quality in English as second language oral narratives. Apart from this, we used the proportion of various POS tags in the text following previous work.

We also explored word and POS language models initially but since we only reached baseline performance with them, we discarded that feature set for further experiments. One reason for the poor performance of word models could be the morphological richness of Estonian which results in data sparsity for building good language models. For the POS models, we faced two issues: while using only the base POS tags resulted in all categories looking alike, using the entire morphological tag resulted in data-sparsity. We did not explore lemma/stem based language models and factored language models yet. A short experiment considering only a feature set consisting of morphological suffixes did not result in an improvement in the result. So, we discarded this feature set too, for the subsequent experiments reported in this paper.

Finally, we did not implement any syntactic features in our approach since we are not aware of any state-of-the-art parsers for Estonian. We also did not implement any features based on spelling and grammar error patterns as we are not aware of any off-the-shelf tool we can use for automatic annotation of learner errors. We also did not verify the output of the Treetagger for possible errors in tagging, as we are not aware of an automatic approach to detect them.

[8] http://lpcs.math.msu.su/ pentus/etmorf.htm

4 Experiments

Our corpus is a collection of texts spanning multiple proficiency levels. The proficiency levels can be assumed to be discrete or continuous, and with varying degree of difference between succeeding levels (i.e., difference between A2 and B1 may be less than that of B1 and B2). This allows us to conceptualize the problem of proficiency classification as belonging to nominal or interval or ordinal scales. Accordingly, we investigated this dataset by considering the problem as classification and regression. We did not explore ordinal representation yet. We used WEKA (Hall et al., 2009) for training the machine learning models and for feature selection.

4.1 Evaluation Measures

We used multiple evaluation measures based on the choice of learning approaches. We evaluated classification performance in terms of its prediction accuracy. Additionally, we report the confusion matrices and F-scores per class to compare the performance with balanced and unbalanced datasets. For linear regression, we report Pearson correlation and Root Mean Square Error (RMSE) as evaluation measures. All the evaluation was performed in a 10-fold Cross Validation setting.

We are not aware of any direct measure of comparison between classification and regression approaches using the same data. Hence, we used three measures after rounding off the regression prediction to the nearest integer value:

1. Percentage of exact matches (This is the same as accuracy for classification.)

2. Percentage of instances where the prediction is within one-level of the actual value (This is closely related to the prediction error and adjacent accuracy in regression models.)

3. Percentage of errors where the prediction is higher than the actual level. This measure was considered with the assumption that in a placement testing scenario, assigning a learner actually belonging to A2 as say, B2 is more undesirable compared to assigning a B2 learner to A2 level.

4.2 Modeling as Classification

We used the Sequential Minimal Optimization (SMO) implementation in WEKA (Platt,1998) in all our classification experiments for easy comparison with previous work with these features. Since the training corpus used in this paper differs from that used in Vajjala and Lõo (2013) in terms of the grades assigned to the texts, a direct comparison of results is not possible. Hence, we started with a replication of the classification experiment with the feature set used in their paper using the new four-level corpus described in Table 1. This resulted in a classification accuracy of 73.7% (F-score 0.74) using an unbalanced dataset and 72.3% (F-score 0.72) using a balanced dataset consisting of 92 texts per category. We consider these as our baseline measures for the rest of this paper.

After establishing this baseline, we tested the model with all our features. The model with all the features received 79% accuracy for the unbalanced corpus and 76.9% for the balanced corpus, which is clearly an increase of about 5% over the previously reported feature set. Table 3 shows the confusion matrices for both unbalanced (left) and balanced (right) versions and Table 4 shows the F-scores per category for both the versions.

(a)class. as →	A2	B1	B2	C1	(b) class. as →	A2	B1	B2	C1
A2	150	46	0	0	**A2**	79	11	2	0
B1	16	340	27	1	**B1**	9	66	16	1
B2	1	58	136	12	**B2**	3	15	60	14
C1	0	3	21	68	**C1**	0	1	13	78

Table 3: Confusion matrices for Unbalanced and Balanced training datasets

Category	F-score in un-balanced dataset	F-score in balanaced dataset
A2	0.83	0.86
B1	0.82	0.72
B2	0.70	0.65
C1	0.79	0.85

Table 4: F-scores per category, for Balanced and Unbalanced training datasets

It is interesting to note that the unbalanced corpus did not create a particular bias for or against any one level. However, the balanced version resulted in a drop in accuracy by ∼2%. Though the prediction accuracy for A2 and C1 clearly increased in the balanced version, this also resulted in reducing the F-score for both B1 and B2. It is difficult to decide which version of the corpus is better for classification - balanced or unbalanced, in this case. However, since the model trained on the unbalanced version results in better accuracy without completely being skewed towards majority classes, we can perhaps consider it as the better performing model between the two. While the increase in accuracy compared to what was reported in Vajjala and Lõo (2013) (67%) can be attributed to the fine-grainedness of the dataset we use now, the performance improvement between their feature set and ours on the new dataset shows that the increased accuracy is not entirely a data artifact.

We also trained binary classifiers for all the class combinations to understand if it is easy or difficult to classify between pairs of classes. Table 5 summarizes the experiments, performed considering equal instances from both classes in each case, in terms of classification accuracy.

Classes Used	Classification Accuracy
A2 vs B1	83.2%
A2 vs B2	92.4%
A2 vs C1	98.4%
B1 vs B2	77.2%
B1 vs C1	95.1%
B2 vs C1	86.3%

Table 5: Binary Classification Accuracy

As the table shows, all the binary classifiers achieved accuracies much higher than the four-class classification accuracy for the balanced corpus (76.9%), excepting (B1 vs B2). This encourages us to explore a multi-level classification approach like the cascades used in Vajjala and Lõo (2013) in future, which could result in an improvement over the current accuracy.

Finally, we also verified if adding the surface features (word length and sentence length) contributes to improving the classification performance. There was a slight drop (∼0.2-0.5%) in accuracy for both unbalanced and balanced datasets and the drop was not statistically significant.

4.3 Modeling as Regression

Another way besides classification is to look at proficiency level prediction as regression. As mentioned earlier, these proficiency levels can be seen as scores on a numeric scale too, since proficiency is a continuous variable though the levels assigned are discrete. Further, regression also allows us to output a prediction on a scale, where it is possible to see predictions that lie between two discrete levels. Hence, we also modeled proficiency prediction as regression. We trained a Linear Regression model in WEKA, with default settings[9]. This regression model achieved a Pearson correlation of 0.85 and an RMSE of 0.49. For the feature set of Vajjala and Lõo (2013), the numbers were 0.77 and 0.58 respectively. Since a majority of previous research on this problem treated it as classification, we have no comparable results on regression. One related result is that of Hancke (2013), who reported a Pearson correlation of 0.78 and RMSE of 0.68 for a German proficiency assessment dataset consisting of 5 levels.

4.4 Comparing Classification and Regression

As mentioned in Section 4.1, we compare classification and regression in terms of exact and within one level prediction accuracy and in terms of the direction of error. Table 6 shows the comparison between the performance of classification and regression in terms of these measures. As it can be seen from the table, both classification and regression perform comparably in terms of within one level accuracy. However, classification seems to work slightly better than regression in terms of exact accuracy as well as the direction of error.

Model	**Exact Acc.**	**Within 1-Level Acc.**	**% errors where Predicted > Actual**
Classification	79%	99.43%	46.5%
Regression	76%	99.2%	50.7%

Table 6: Classification Vs Regression - Comparison

5 Feature Selection

While the above experiments will let us conclude about the efficiency of the features to predict CEFR levels accurately, understanding what features play a significant role in distinguishing between levels is interesting from a general linguistic perspective. Further, the question of how much can we predict with how few interpretable features is interesting from an application perspective. Hence, we investigated feature selection approaches and correlations between features. We used three feature selection methods implemented in WEKA and built classifiers with these extracted feature subsets. The three methods differ in their approach to feature selection. They are:

1. Information Gain - evaluates an attribute in terms of its information gain with respect to the class. Hence it considers attributes individually, irrespective of the correlations between them.

[9] M5 attribute selection, eliminateCollinearAttributes option set to TRUE

2. CfsSubsetEval (Hall, 1998) - chooses a feature subset such that the amount of redundancy between the features in the selected subset is less and together, they have a higher predictive ability with respect to the class.

3. ReliefFAttributeEval (Kira and Rendell, 1992; Kononenko, 1994) - selects individual features by repeated sampling of instances and comparing the value of the feature for the sampled instance and the nearest instances belonging to same and different classes.

Table 7 shows the classification accuracies with all the three feature selection methods, using the unbalanced dataset. Since there was no specific bias against any class and since model with unbalanced dataset gave a higher accuracy, we report all the further results only with that dataset. As the table shows, CfsSubsetEval attains almost the same classification accuracy (78.3%) as the best performing model so far (79%) with a much smaller feature set. The difference was found to be statistically insignificant (using the Corrected Paired T-Tester implementation in WEKA).

Method	# Features	Accuracy
Information Gain	10	73.5%
CfsSubsetEval	27	78.3%
ReliefFAttributeEval	10	74.5%

Table 7: Classification Accuracy with Feature Selection

From the group of 27 features of CfsSubsetEval, Table 8 shows the top-10 features from the CfsSubsetEval set, ranked by their Information Gain. Features indicated with an asterisk (*) are the ones that were not used for this task before.

Feature	Group
Corrected Type Token Ratio (CTTR*)	LexVar
Root Type Token Ratio (RTTR*)	LexVar
# 2nd person inflected verbs/# words (numPs2)	Morph
# sub-ordinating conjunctions/# words (numSubC)	Morph
# verbs in active voice/# words (numActive)	Morph
Squared Verb Variation (SVV1*)	LexVar
# distinct cases used in the document (CaseNr*)	Morph
Corrected Verb Variation (CVV1*)	LexVar
# conjunctions/# words (numConj)	Morph
# interjections/# words (numInterj)	Morph

Table 8: 10 Predictive Features (CfsSubset, ranked by Information Gain)

Five of the 10 features in this list are the ones not used in previous research for this task. However, some feature pairs in this list like (CTTR,RTTR), (SVV1,CVV1) are still variations of the same ratio and are expected to be highly correlated with each other. While this correlation between features may not affect the prediction performance as such, studying the correlations may be more useful from the perspective of understanding the process of proficiency acquisition and will also be useful in building models where less number of features explain can more variation in the data without repetition.

5.1 Correlations between Features

Table 9 lists the correlations between some pairs of features from the top-10 features listed in Table 8.

Feature 1	Feature 2	Correlation
CTTR	RTTR	0.999
CVV1	SVV1	0.976
RTTR	SVV1	0.804
CTTR	SVV1	0.804
RTTR	CVV1	0.795
CTTR	CVV1	0.795
numSubC	numConj	0.764
RTTR	CaseNr	0.719
CTTR	CaseNr	0.719
numConj	numInterj	-0.623

Table 9: The Most Correlated Features

Several features in the most predictive features have a high degree of correlation between each other, as shown in the table. It would perhaps be sufficient to use only one of them to achieve the same amount of predictability. Further analysis is needed to choose a refined feature set that can be as predictive in terms of modeling but also more interpretable in linguistic terms.

5.2 Predictive Features between Categories

As a final experiment, we investigated the most predictive features between categories. The motivation for this exploration has been to understand if there is a change in the features that are more useful, as the proficiency increased. One hypothesis could be that the morphological features are more predictive between lower proficiency levels and lexical richness features like TTR will be more predictive at higher proficiency levels. While beginning to learn Estonian, learners may have issues with its morphological complexity. But, as they become more familiar and proficient with the language, the effect of morphology may diminish and that of lexical richness may grow. Table 10 lists the top-5 most predictive features for binary classification between proficiency levels going from least to highest ranked in terms of their Information Gain.

A2 vs B1	B1 vs B2	B2 vs C1
numInterj	RTTR	numPs2
numPs2	CTTR	# imperatives/# words
numConj	SVV1	# abessive case/# words
numSubc	CVV1	CaseNr
# affirmative verbs/# words	# compound words/# words	# compound words/# words

Table 10: Most Predictive Features Between Categories

Interestingly, all the top-5 features are morphological features in the comparisons between (A2,B1) and (B2,C1). But, the comparison between the middle levels (B1,B2) is dominated by the lexical richness features. While we do not have any intuitions about the reasons for this morphology -> lexical richness -> morphology turn with the increase in proficiency, SLA research may be able to offer some perspectives on this. It would be interesting to combine computational modeling with SLA to develop an interpretable model of proficiency assessment, for which this research could a good starting point. We did not check whether this list of top-5 features varies depending on the size of the sample chosen for the analysis. To estimate the top features in this case, we used all the data available for the chosen categories.

6 Conclusions and Outlook

To conclude, we built models for automatic proficiency classification of Estonian learner texts based on the CEFR scale. We used a collection of morphological and POS tag density based features including lexical richness measures from English SLA research. The best model reported in this paper reaches a prediction accuracy of 79%. Our results show a substantial improvement over the previously reported results for Estonian and are also considerably higher than the accuracies reported on this task for other languages (German and Swedish). We can conclude from our experiments that this linguistic modeling of proficiency holds promise in the direction of developing an automatic proficiency assessment system for Estonian.

The nature of the CEFR categories allows us to model the problem as being on different scales. So, we considered nominal and interval scales and modeled the problem as both classification and regression. Comparing both of them in terms of 3 evaluation measures, we concluded that we get slightly better results by treating proficiency assessment on CEFR scale as classification instead of regression.

We experimented with feature selection strategies to understand how far can we get with how few features and found that we can reach almost the same accuracy with a small subset of 27 features. Along with this, we also did a correlational analysis between features and our experiments showed that most of the features are highly correlated with each other. We are yet to develop a solution to deal with this issue.

It has to be noted that we looked at only one dimension of proficiency in this set of experiments, ignoring aspects of syntax, discourse, learner errors, relation of the text to the question asked etc. Also, since we ignored the possibility of tagging errors by TreeTagger, we need to caution that the results need to be interpreted keeping the tool in context. However, despite these two limitations, we believe our experiments still demonstrate the value of using language specific linguistic information for performing proficiency classification of Estonian learner text.

6.1 Outlook

We are currently working on assessing what models are statistically different from each other using significance testing and by studying the fold-wise difference in a 10-fold cross-validation setup along with average performance difference between models. This may provide us a better way to compare various prediction models when the performance difference is small and also understand the effect of sampling of instances per fold on the performance of the models.

The dataset is still under development and it would be interesting to see how far can we get with the current feature set, when more annotated data becomes available in future. Although the results appear very promising at the moment, there is a lot of scope for improvement and exploration of new feature groups. Our initial experiments with n-gram models failed, but exploring factored language models and using data-driven word-frequency, spelling error and suffix-frequency features may perhaps be more useful for the task. Further, subject to the availability of required tools, we plan to explore the role of syntactic features of texts in proficiency classification.

From a modeling perspective, exploring cascade approaches like the ones in (Vajjala and Lõo, 2013) may help in improving the accuracies further. Investigating other learning algorithms and considering the dataset as ordinal - are other interesting directions that could be explored. Finally, since the results clearly establish the impact of morphological features for this task, it would be interesting to verify if this is the case for other morphologically rich languages where such learner corpora are available.

Acknowledgments

We thank the anonymous reviewers for their comments and Dr Detmar Meurers for his useful suggestions. This research is partially funded by LEAD Graduate School (GSC 1028, http://purl.org/lead), a project of the Excellence Initiative of the German federal and state governments. Sowmya Vajjala is a doctoral student of the LEAD Graduate School.

References

Burstein, J. (2003). *The e-rater Scoring Engine: Automated Essay Scoring with Natural Language Processing*, chapter 7, pages 107–115. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Burstein, J. and Chodorow, M. (2010). *Progress and New Directions in Technology for Automated Essay Evaluation*, chapter 36, pages 487–497. Oxford University Press, 2nd edition.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

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

Eslon, P. (2014). Eesti vahekeele korpus (Estonian Interlanguage Corpus). *Keel ja Kirjandus*, 6:436–451.

Gyllstad, H., Grandfeldt, J., Bernardini, P., and Källkvist, M. (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written l2 english, l3 french and l4 italian. *EUROSLA Yearbook*, 14(1):1–30.

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

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, Newzealand.

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. and Meurers, D. (2013). Exploring CEFR classification for german based on rich linguistic modeling. In *Learner Corpus Research 2013, Book of Abstracts*, Bergen, Norway.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Ninth International Workshop on Machine Learning*, pages 249–256.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*, pages 171–182.

Kyle, K. and Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, –:–.

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*.

Östling, R., Smolentzov, A., Tyrefors Hinnerich, B., and Höglin, E. (2013). Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Tono, Y. (2000). A corpus-based analysis of interlanguage development: analysing pos tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora*, pages 323–340.

Vajjala, S. and Lõo, K. (2013). Role of morpho-syntactic features in Estonian proficiency classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*.

Williamson, D. M. (2009). A framework for implementing automated scoring. In *The annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME)*.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (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*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics. Corpus available: `http://ilexir.co.uk/applications/clc-fce-dataset`.

Zhang, B. (2008). Investigating proficiency classification for the examination for the certificate of proficiency in english (ECPE). In *Spaan Fellow Working Papers in Second or Foreign Language Assessment*.