Role of Morpho-Syntactic Features in Estonian Proficiency Classification

Sowmya Vajjala
Seminar für Sprachwissenschaft
Universität Tübingen
sowmya@sfs.uni-tuebingen.de

Kaidi Lõo
Seminar für Sprachwissenschaft
Universität Tübingen
kaidi.loo@student.uni-tuebingen.de

Abstract

We developed an approach to predict the proficiency level of Estonian language learners based on the CEFR guidelines. We performed learner classification by studying morpho-syntactic variation and lexical richness in texts produced by learners of Estonian as a second language. We show that our features which exploit the rich morphology of Estonian by focusing on the nominal case and verbal mood are useful predictors for this task. We also show that re-formulating the classification problem as a multi-stage cascaded classification improves the classification accuracy. Finally, we also studied the effect of training data size on classification accuracy and found that more training data is beneficial in only some of the cases.

1 Introduction and Motivation

Every year, language learners across the world learn various languages and take tests that measure their proficiency level. The Estonian language proficiency examination[1] in particular is usually taken by the immigrant population for citizenship and/or employment needs in Estonia. Assessing learner texts to classify them into relevant proficiency levels is usually done by human evaluators and is often a time consuming process. An approach to automate this process would complement the human annotators and reduce the overall effort in evaluating learner texts for their proficiency. Investigating features that follow any sort of trend across the various proficiency levels among learners is a first step in building such automatic proficiency classification systems. This is the main motivation for our research.

Several factors might play a role in determining a learner's proficiency in a given language. Since we study the learner corpus of Estonian, a morphologically complex language with an elaborate declension and conjugation system, we hypothesized that studying the role of morpho-syntactic features would be a good starting point to perform proficiency classification. We used the Estonian Interlanguage Corpus (EIC)[2], a publicly accessible corpus of written texts produced by learners of Estonian as a second language, for this purpose. All the texts were annotated with a proficiency level that is based on the Common European Framework of Reference for Languages Council of Europe (CEFR). We constructed various proficiency classification models based on this corpus by using features motivated primarily by the morphological complexity of Estonian and found that true to our hypothesis, they turn out to be good predictors of the proficiency level.

We also studied the effect of breaking up the main classification task into sub-tasks and cascading them. We show that this approach increases the overall accuracy of proficiency classification. In addition, we studied the effect of training data size and found that it does not have a significant impact in most of the classification tasks we performed. To summarize, we studied the task of proficiency classification for Estonian by studying both the aspects feature engineering and model construction.

[1] http://www.ekk.edu.ee/

[2] http://evkk.tlu.ee/wwwdata/what_is_evk

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 63–72,
Atlanta, Georgia, June 13 2013.

The rest of this paper is organized as follows: Section 2 briefly surveys related work and explains the context of our research. Section 3 describes our corpus and the experimental setup. Section 4 describes our feature set. Section 5 describes our experiments and results. Section 6 concludes the paper with a discussion on results and directions for future work.

2 Related Work

With the availability of computer based learner corpora, research focusing on studying the criterial features that correlate with proficiency levels began to emerge. A wide body of research exists on studying the syntactic complexity of texts produced by learners across different proficiency levels, their lexical richness and the errors they make (e.g., Lu, 2012; Vyatkina, 2012; Tono, 2000) . Learner data from both longitudinal and cross sectional studies was analyzed to understand the linguistic patterns among learners of different proficiency levels, in Second Language Acquisition (SLA) research.

Automatic proficiency assessment of learner texts is another active area of related research, which plays an important role in language testing. Automated systems are now being used both for evaluation of language learners and for offering feedback on their language proficiency (e.g., Williamson, 2009; Burstein et al., 2003). Forms of text used for assessment include mathematical responses, short answers, essays and spoken responses among others (Williamson et al., 2010). Standardized tests like GRE and GMAT too use such systems to complement human scorers while evaluating student essays automatically (Burstein, 2003; Rudner et al., 2005). Zhang (2008) discusses proficiency classification for the Examination for the Certificate of Proficiency in English (ECPE) in detail, by comparing procedures based on four types of measurement models. The problem of automatic student classification i.e., making inferences about a student's skill level by using some form of data about them is an active area of research in Educational data mining (e.g., Desmarais and Baker, 2012; Baker 2010).

But, automatic approaches for classifying language learners into standardized proficiency levels (e.g., the European CEFR levels[3], Common Core Standards[4]) is a relatively new area of interest.

Supnithi et al. (2003) used a dataset consisting of audio transcripts by Japanese learners of English to build a proficiency classification model with a feature set that modeled vocabulary, grammatical accuracy and fluency. This dataset had 10 levels of proficiency. Hasan and Khaing (2008) performed proficiency classification with the same dataset using error rate and fluency features. Dickinson et al. (2012) developed a system for classifying Hebrew learners into five proficiency levels, using features that focus on the nature of errors in a corpus of scrambled sentence exercise questions.

Proficiency Classification so far has been predominantly focused on the correlation of error-rate with proficiency. Although error-rate is a strong indicator of a learner's proficiency in a language, considering other factors like lexical indices or syntactic and morphological complexity would help in providing multiple views about the same data. Providing a non-error driven model, Crossley et al. (2011) studied the impact of various lexical indices in predicting the learner proficiency level. Using a corpus of 100 writing samples by L2 learners of English classified in to three levels (beginner, intermediate, advanced), they built a classification system that analyses language proficiency using the Coh-metrix[5] lexical indices.

Most of the research about the distinguishing factors among learners of various proficiency levels has focused on English. However, issues like morphological variation, which may not be strong predictors of learner proficiency in English, could be useful in proficiency classification of other languages. Hence, in this paper, we study the texts produced by the learners of a morphologically rich and complex language, Estonian and show that morphology can be a good predictor for learner proficiency classification.

We build our classification models using the Estonian Interlanguage Corpus (EIC), which contains texts produced by learners of Estonian as a second language. We modeled our approach based on the features motivated by the morphological complexity of Estonian. To our knowledge, this is the first

[3] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

[4] http://www.corestandards.org/

[5] http://cohmetrix.memphis.edu

work that studies the role of morphology based features for proficiency classification in general and in Estonian in particular.

3 Corpus and Experimental Setup

3.1 Corpus

The Estonian Interlanguage Corpus (EIC)[6] was created by the Talinn University. It is a collection of written texts produced by learners of Estonian as a second language. Most of the learners were native speakers of Russian. The corpus consists mainly of short essays, answers to questions, translations and personal letters. The texts are annotated with error types and incorrect forms. The corpus also provides information about the learner's age, gender, education and about other languages known to the learner. Descriptive statistics about the corpus are available on their website[7]. The corpus contains around 8000 documents (two million words), most of which are texts from the Estonian language proficiency examination. The length of the texts varies in general between 50 and 1000 words (Eslon, 2007).

Information about the learner's level of competence is based on the CEFR guidelines[8] and is decided by human annotator judgement. Until late 2008, Estonian language proficiency was tested by conducting proficiency exams at three levels - the lowest level A, the medium level B and the highest level C. Later, the CEFR standards were adapted, dividing the development of language proficiency into six levels (A1, A2, B1, B2, C1, C2). A1 indicates a basic proficiency and C2 indicates a mastery.

For our current work, we use a sub-corpus consisting of 2000 texts that can be accessibly through the EIC. These texts are spread across three broad levels A, B, C instead of the more fine grained six levels and contain all kinds of texts including short answers. Although these texts also have an annotated version containing information about error-types and corrections, since our aim in this paper is to study the effect of morpho-syntactic features, we considered the raw texts produced by the learners as they were, without looking at the error annotations. Table 1 shows a summary of the entire corpus that was made available.

We prepared a test set consisting of 50 documents from each category, picked randomly. This test set was not used to train any of the classifiers we used in this paper. Further, to avoid a training bias towards any class, we used equal number of instances from all classes during all our binary and three-class training processes.

Proficiency Level	#Docs	Avg. #tokens
A-level	807	182.9
B-level	876	260.3
C-level	307	431.8

Table 1: The EIC Corpus

3.2 Pre-processing

All the texts in our corpus were POS-tagged with the TreeTagger[9] and the tagged output was then used to extract the required features. The TreeTagger (Schmid, 1994) is a probabilistic part of speech tagger, which contains parameter files to tag Estonian data. The tag set was derived from the Tartu Morphologically Disambiguated Corpus tag set[10]. As mentioned earlier, we do not use the error annotation information for these learner texts, in this paper.

4 Features

Our choice of features were primarily motivated by the nature of the morphology of Estonian.

4.1 The Estonian Language

The Estonian language has about one million native speakers. It belongs to the Finnic branch of Uralic languages and is known for it's complex morphology. It is both an agglutinative and a flectional (fusional) language. Some of the prominent features of Estonian language include:

- 14 productive nominal cases

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

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

[8] http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

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

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

- no grammatical gender (either of nouns or personal pronouns) and no articles (either definite or indefinite)
- the verbal system lacks a morphological future tense (the present tense is used instead)
- relatively free word order (relations between words are expressed by case endings)
- extensive compound word formation
- impersonal voice (specific to the Finnic languages and similar to passive voice. The verb is conjugated in "fourth person", who is never mentioned)
- Most of the inflected words in Estonian have two distinctive parts: the stem and the formative. For example, *raamatutele* (book, plural, allative) consists of the stem *raamatu* and the formative *tele*, which in turn consists of plural marker *te* and allative case marker *le* (Erelt et al., 2007, p. 203).
- Unlike most of other Finnic languages, Estonian also has flective features, i.e., the same morpheme may have different shapes in different word forms. For example, the stem *jalg* ("foot", singular, nominative) may appear as *jala* (singular, genitive) or *jalga* (singular, partitive) and plural marker may appear as *d*, *de*, *te* or *i* or merged with the stem as in *jalad* (plural, nominative), *jalgade* (plural, genitive) and *jalgu* (plural, partitive) (Erelt et al., 2007, p. 203).

As many of these characteristics are morphological in nature, we hypothesized that this morphological complexity of Estonian may play a role in the process of language learning and hence may be a useful predictor for proficiency classification. Hence, we built our feature set primarily focusing on the morphological properties of the learner texts. Apart from these features, we also included other features based on the Parts of Speech and lexical variation.

4.2 Morphological Features

In Estonian, as in other Finnic languages, nominals (nouns, adjectives, numerals and pronouns) and verbs are inflected for number and case. Estonian nominals are inflected in 14 different cases. Three of the nominal cases are grammatical cases, i.e., nominative, genitive and partitive. They fulfill mainly a syntactic purpose and have a very general grammatical meaning. All the other cases are semantic cases, and they have a more concrete meaning than grammatical cases, which often can be explained by means of adverbs or adpositions (Erelt et al., 2007, p. 241). We considered the proportion of nouns and adjectives tagged with various cases per document and included them as our declension features. The cases we considered in this paper are: nominative, genitive, partitive, illative, inessive, elative, allative, adessive, ablative, translative, terminative, essive, abessive, comitative and short singular illative, i.e., aditive case.

The verb in Estonian has finite forms that occur as predicates and auxiliary components of complex predicates and non-finite forms. Finite forms are inflected for mood, tense, voice, aspect, person and number. The verb has altogether five moods: the indicative, conditional, imperative, quotative and jussive. It has two simple tenses: the present and the past, two voices: personal and impersonal, affirmation and negation. Non-finite forms behave differently. Participles are inflected for voice and tense, present participles also for case and number, and supines for voice and case. There is one infinitive and one gerund, which can be explained as the inessive case form of the infinitve (Erelt, 2003, p. 52). In this paper, we considered the proportion of verbs belonging to various tense, mood, voice, number and person categories as our features.[11].

4.3 POS features

We included the various degrees of comparison of adjectives and the proportion of words belonging to various parts of speech among our features. This group of features also included the proportion of adpositions (=prepositions+postpositions) along with the proportion of prepositions and postpositions separately. We also included the proportion of coordinating conjunctions and subordinating conjunctions along with that of all conjunctions.

[11]Examples of various forms of declension and conjugation can be found in the Estonian morphology guide at: `http://lpcs.math.msu.su/~pentus/etmorf.htm`

4.4 Lexical Variation features

Lexical variation, also called lexical range indicates the range of vocabulary displayed in a learner's language use. We implemented the measures of lexical variation that are used in the English SLA research to measure the lexical richness of the learners of English as a second language (Lu, 2012). These included the noun variation, verb variation, adjective variation and verb variation which indicated the ratio of the words with the respective parts of speech to the total number of lexical words (instead of all words).

4.5 Text Length Feature

Since text length is one of the most commonly used measures of learner proficiency and also because of the variation in average text length across the proficiency levels (Table1), we included the number of word tokens per document as a feature.

4.6 Most Predictive Features

Apart from these individual feature groups, we also performed a feature selection, to identify the most predictive ones among all our features. We used the Correlation based Feature Subset (CFS) selection method in WEKA for this purpose. CFS chooses a feature subset considering the correlation and the degree of redundancy between the features. Table 2 consists of a list of the most predictive and non-redundant features after ranking all the selected features based on their Information Gain. This list consisted of five verb morphology based features followed by three nominal declension features.

Feature	Group
Nominative case	NounMorph
Impersonal	VerbMorph
Personal	VerbMorph
Num. words	TextLength
Present tense	VerbMorph
2nd person verbs	VerbMorph
Prepositions	POS
Allative case	NounMorph
Imperatives	VerbMorph
Translative case	NounMorph

Table 2: 10 Most Predictive, Non-redundant Features

It is interesting to note that several characteristics that are prominent in Estonian (cf. Section 4.1) figured among this list of most predictive features. *Nominative* being the top predictor can be explained due to the difference in (the number of) cases between Estonian and other languages. For example (Eslon, 2011) found in her corpus study based on the same corpus that the learners frequently use nominative case instead of genitive and partitive case. So, it is to be expected that the usage of the nominative case changes as the proficiency increases. *Impersonal* and *personal* voice are distinctive features in Estonian and other Finnic languages, as they are different from the active and passive voice that typically exist in other languages (Erelt, 2003). This may make them difficult to master for language learners, making them one of the top predictors for proficiency. Further, Estonian has more postpositions than prepositions. Hence, one could that the use of prepositions will be replaced by postpositions as the language acquisition progresses (Ehala, 1994).

5 Experiments and Results

We first studied the effect of the individual feature groups as well as their combination for a three class classification of Estonian learners into A, B and C classes. We also studied the impact of a stacking ensemble on the overall classification accuracy and found out that it did not result in a significant improvement on the test set. Hence, we further investigated the problem as a collection of multi-stage two-class cascades instead of a single stage three class classification. For all our classification experiments, we used the WEKA (Hall et al., 2009) toolkit. We report the overall classification accuracy as our evaluation metric.

5.1 Three Class-Classification

We first considered the learner classification as a single step, three class classification problem. Since 50 documents from each category were separated as a held-out test set (cf. Section 3.1), we built our three-class models with 250 texts per category as our training set to ensure that there is a balanced distribution between classes. We trained multiple classification models considering the individual feature

groups and the most predictive feature group. Table 3 shows the classification accuracy of various feature groups, reported using the Sequential Minimal Optimization (SMO) implementation in WEKA (Platt, 1998).

Features	10-Fold CV	Test set
Random baseline	33.33%	33.33%
Noun Morph.	56.64%	52%
Verb Morph	57.55%	58%
POS	52.99%	47.33%
Lex. Variation	43.36%	47.33%
Text Length	33.72%	34%
All Features	**62.45%**	59.33%
Noun+Verb Morph	**61.45%**	58%
Top10 features (Table 2)	57.34%	56.58%

Table 3: Estonian Learner Proficiency Classification with various Feature groups

Although the classification accuracies overall are not very high, it can be seen from the results that the morphological variation does play a key role in proficiency classification of Estonian. While the verbal morphology features performed best as an individual feature sub group, the addition of lexical variation and POS features to the morphological features added very little to the overall classification accuracy.

Text length turned out to be the most predictive single feature among the top features. It can be seen from Table 3 that this feature alone resulted in a classification accuracy of 34%, which is just above the random baseline (33.33%). But the fact that the C level in general contained a higher number of essays and translations compared to other categories of text like letters and short answers (than the A and B levels), thereby resulting in longer texts in general, may have resulted text length being the single most predictive feature. The Top-10 features also performed on par with the individual morphological feature subgroups.

5.1.1 Ensemble Model

Since ensemble models are known to obtain a better performance than their constituent models, we compared the performance of a stacking ensemble against its individual constituent models. We trained three classification models on the entire feature set, using the same train-test sets as explained before and trained an ensemble model with three classifiers. We used the StackingC implementation of WEKA (Seewald, 2002) to combine the models, with a linear regression model as our meta classifier. Table 4 shows the classification accuracies for the individual classifiers as well as the ensemble on a 10-fold CV of the training set and on the held out test set. The ensemble did not result in any significant improvement ($<$1%) compared to the best model amongst the three of its individual components (SMO). The ensemble's performance on the test set was poor compared to the best classification model.

Classifier	10-Fold CV	Test set
SMO	62.45%	59.33%
Logistic Regression	59.37%	52%
Decision Tree	57.29%	52.33%
Stacked Ensemble	**63.28%**	57.33%

Table 4: Proficiency Classification With an Ensemble

5.2 Classification Through Two-Class Cascades

Since combining the classifiers as a stacking ensemble did not work, we turned to reformulating our problem as a cascade of two-class classifiers. Cascade generalization is the process of sequentially using a set of small classifiers to perform an overall classification task. Gama and Brazdil (2000) showed that a cascade can outperform other ensemble methods like stacking or boosting. Kaynak and Alpaydin (2000) proposed a method to sequentially cascade classifiers and showed that this improves the accuracy without increasing the computational complexity and cost. Although the creation of our classifier cascades in this paper is not the same as any of the above mentioned research, their conclusion that cascading subsets of classifiers to build an overall classifier can possibly result in a better accuracy was the main motivation for this experiment.

The SMO implementation in WEKA also considers multi-class classification as a combination of pairwise binary classifications. But, in our subsequent experiments, we combine our two-class classifiers as a multi-stage cascade rather than a multi-expert stacking ensemble. For these experiments,

we first built the various binary classifiers that were later used to construct the cascades. We chose our combinations both by using a One vs All (OvA) as well as a One vs One (OvO) strategy. Thus, six binary classifiers were created, namely:

- (A, B) classifier
- (B, C) classifier
- (C, A) classifier
- (A and Not A) classifier
- (B and Not B) classifier
- (C and Not C) classifier

In all the cases, our training data consisted of equal number of instances per class. In the cases of the last three classifiers, the training data for NotA, NotB and NotC categories consisted of instances from both the classes that were included in the respective "Not-" classes. The data from the held-out test set was not included in any of these binary classification experiments. The training data size for each classifier has a different size depending on the classes involved. In all cases, the number of training samples per category is equal to the number of documents belonging to the category with the least number of documents. Hence, in cases involving the C-class (ABC, AC, BC, CnotC), we trained the classifiers with 250 documents per category. In all the other cases (AB, AnotA, BnotB), we trained the classifiers with 750 documents per category. Table 5 summarizes the training data size and the classification accuracies using 10-fold cross validation. All the models were trained using the SMO algorithm.

Classifer	Training data size	Accuracy
A,B	750 per cat	70.8%
B,C	250 per cat	74.59%
A,C	250 per cat	85.93%
A,NotA	750 per cat	74.20%
B,NotB	750 per cat	60.04%
C,NotC	250 per cat	79.69%

Table 5: Binary Classifications of Estonian Learners

This binary classification shows that there is a clear trend among the features across the proficiency levels. In the case of a pair-wise classification between classes, the highest classification accuracy was achieved for the binary classifier that considered the A and C classes. Although the classification accuracies of the binary classifiers (A,B) and (B,C) are considerably higher than the overall three class classification accuracy (Table 3), they are very low compared to that of the binary classifier (A,C). The confusion between the three classes is the highest when it involves the middle class, B. This confirmed the ordinal nature of proficiency classification. In the second set of binary classifiers, again, the classifier with a poor performance turned out to be (B,NotB).

To take advantage of the fact that the two-class classification is much more accurate than the three-class classification, we studied three class classification by building multi-stage classifier cascades using the above binary classifiers. Based on the output of the first stage (which is the most accurate classifier), we feed the test instance to one of the remaining classifiers to get the final prediction.

5.2.1 Cascade-1

For the first cascade, we considered the pairwise binary classifiers that used a One vs One (OvO) strategy from Table 5. We constructed a classifier cascade as follows: For each test instance,

- Classify the instance using the classifier (A,C).
- If A, re-classify the instance using the classifier (A,B).
- if C, re-classify the instance using the classifier (B,C).

5.2.2 Cascade-2

For the second cascade, we considered the second set of binary classifiers from Table 5, which use a One vs All (OvA) strategy. The cascade is constructed as follows: For each test instance,

- Classify the instance using the classifier (C,NotC).
- If C, classify the instance as C.
- Else, re-classify the instance using the classifier (A,notA).

The choice of these particular combinations of cascades was motivated by two factors:

- To understand the performance of OvO and OvA binary classifier cascades independently
- To start with the classifier that has the highest accuracy as the first stage.

Table 6 compares the performance on the test set of the cascaded classifiers against the normal 3-class classifier and a classifier ensemble. Compared to a normal three-class classifier, the cascaded approach showed more than 5% improvement in the classification accuracy using both the cascades. Compared to Cascade-1, Cascade-2 performed even better with a 66.66% classification accuracy on the test set. Since binary classification for certain pairs seemed to be possible with higher accuracy than the three-class classification, reformulating three class classification as a cascade of binary classifications may result in a better classification accuracy. This was the initial motivation for the choice of cascade classification. Our results clearly showed that it was a fruitful experiment.

Classifer	**Accuracy**
Cascade-1	**64.66%**
Cascade-2	**66.66%**
3-class,without cascade	59.33%
3-class ensemble	57.33%

Table 6: Comparison of Cascade classification

The cascades need more exploration though. Also, although the morphological features turned out to be useful predictors of proficiency classification, the classification accuracies are still not very high. Two possible explanations could be that our features are good but not sufficient or that the training data was insufficient.

It is clear from our various classification experiments that the morphological features are good predictors of proficiency levels. But, surely, there is much more to language proficiency than morphological complexity. So, exploring more features will be the natural next step to improve the overall classification accuracy. However, to gain some more insights at this level, we studied the effect of training data sizes on the various classification tasks we performed.

5.3 Effect of Training Sample Size

We took all the seven different classification models we used in the earlier experiments and studied the impact of gradually increasing the training data size on classification accuracy. For this purpose,we trained all the classifiers with the complete feature set using the SMO algorithm. The classifiers studied include the three class ABC classifier and the binary classifiers AB, BC, AC, AnotA, BnotB and CnotC. Table 7 summarizes the effect of splitting the respective training sets into various train-test splits, on the classification accuracies.

classifier	**50-50**	**60-40**	**70-30**	**80-20**
ABC	56.73%	60.05%	61.76%	62.76%
AB	71.07%	71.3%	71.2%	72.04%
BC	71.33%	72.35%	71.73%	74.86%
AC	86.31%	84.95%	84.15%	85.55%
AnotA	75.39%	75.20%	76.65%	75.82%
BnotB	59.05%	57.95%	56.91%	58.08%
CnotC	77.34%	77.56%	77.27%	76.52%

Table 7: Effect of training size on classification accuracy

As the table shows, training data size had an impact only on some of the classification tasks. For the three class classification, training set size had a clear effect. Although our corpus had a large number of texts from A and B compared to C (Table 1), since we used balanced training sets to train all models, the three-class model had relatively fewer number of documents per category (250) compared to, say, the AB classifier (750 per category). Reduction of this small training set further by 50% decreased the three class classification accuracy from 62.76% (when 80% of the data was used for training) to 56.73%. So, in this case, training data size had an effect.

However, an interesting observation is that this small training sample size (250 documents per category) did not have any impact on the classification performance of the classifier (A,C). This classifier consistently performed at a higher level compared to all the other classifiers even when the training data was only 50% (125 documents per category). Al-

though it is possible that the length of the document played a role here, there was little difference in the performance ($< 1\%$) even after removing the text length feature. This indicates a strong differentiation between the texts of the language learners of levels A and C, in terms of the features we used.

In case of the other classification tasks, only the (B,C) classifier showed some effect of the training data on its overall classification accuracy. While there might be other reasons that we did not notice yet, it is possible that the inter class overlap between (A,B) is more compared to the overlap between (B,C) at least in terms of the features we considered. Also, the fact that the B-level lies in between A and C could also have contributed to the fact that more training data has little effect on classifiers involving data from all the three classes (AnotA, BnotB, CnotC).

6 Conclusion and Discussion

In this paper, we discussed the task of classifying learner texts into standardized proficiency levels based on the texts produced by learners of Estonian as a second language. We used the publicly accessible Estonian Interlanguage Corpus (EIC) and modeled our classifiers by considering the morpho-syntactic variation as our primary feature group. We hypothesized that the morphology may play an important role in detecting the proficiency levels as Estonian is a morphologically rich and complex language.

For building our classifiers, we experimented with various methods such as three class classifiers, an ensemble model and multi-stage cascades. Our experiments showed that the multi-stage cascades improved the classification accuracy compared to the other approaches. Our experiments also showed a clear trend across the proficiency levels. There was little classification overlap between the beginner (A) and the advanced (C) level texts but a strong overlap of both these levels with the intermediate (B) level.

We can conclude from our experiments that the morphological features can indeed play an important role in the proficiency classification of Estonian. Although the classification accuracies we achieved (60-65%) have a long way to go in terms of a real-world grading application, we believe that this is a good starting point to explore the role of morphology in proficiency classification of Estonian in particular and other morphologically rich languages in general.

As a part of our future work, we intend to investigate the role of morphology in Estonian proficiency classification further. We also want to compare the proficiency levels across various genres of texts in the corpus (e.g, essays, personal and official letters, translations etc.). Another interesting dimension we want to explore further is the distribution of specific kinds of morphological phenomena (e.g., case markers) that exist in Estonian but not in the learner's native language, across the different proficiency levels. It would also be interesting to apply insights from the theories of second language acquisition research and study their utility for proficiency classification. Apart from morphology, we also intend to study the impact of other features such as lexical sophistication, error rate, syntactic complexity and discourse coherence. Finally, on the model construction side, we plan to investigate and understand the working of cascaded classifiers better in this context.

Acknowledgments

We thank Dr Pille Eslon from the Talinn University for sharing the corpus with us. We also thank Serhiy Bykh, Dr Detmar Meurers and the three anonymous reviewers for their feedback on the paper. This research is partially funded by the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA)[12]

References

R.S.J.d. Baker. 2010. Mining data for student models. In *Advances in Intelligent Tutoring Systems*, pages 323–338. Springer.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-03)*, pages 3–10, Acapulco, Mexico, August.

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

[12] http://clara.uib.no

Scott A. Crossley, Tom Salsbury, and Danielle S. McNamara. 2011. Predicting the proficiency level of language learners using lexical indices. In *Language Testing*.

M.C. Desmarais and R.S.J.d. Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. In *User Modeling and User-Adapted Interaction, 22(1-2)*.

Markus Dickinson, Sandra Kübler, and Anthony Meyer. 2012. Predicting learner levels for online exercises of Hebrew. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 95–104, Montréal, Canada, June. Association for Computational Linguistics.

Martin Ehala. 1994. Russian influence and the change in progress in the Estonian adpositional system. In *Linguistica Uralica, 3*, pages 177–193.

M. Erelt, T. Erelt, and K. Ross. 2007. *Eesti keele käsiraamat*. Eesti Keele Sihtasutus.

M. Erelt. 2003. *Estonian language*. Linguistica Uralica. Estonian Academy Publishers.

Pille Eslon. 2007. Õppijakeelekorpused ja keeleõp. In *Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine*, pages 87–120.

Pille Eslon. 2011. Millest räägivad eesti keele käändeadendused? lähivõrdlusi. In *Lähivertailuja, 21*, pages 45–64.

Joao Gama and Pavel Brazdil. 2000. Cascade generalization.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Md Maruf Hasan and Hnin Oo Khaing. 2008. Learner corpus and its application to automatic level checking using machine learning algorithms. In *Proceedings of ECTI-CON*.

C. Kaynak and E. Alpaydin. 2000. Multistage cascading of multiple classifiers: One mans noise is another man's data. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Languages Journal*.

John C. Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research.

Lawrence Rudner, Veronica Garcia, and Catherine Welch. 2005. An evaluation of intellimetric^{TM} essay scoring system using responses to gmat awa prompts. Technical report, Graduate Management Admission Council (GMAC).

Helmut Schmid. 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.

A.K. Seewald. 2002. How to make stacking better and faster while also taking care of an unknown weakness. In *In the proceedings of the Nineteenth International Conference on Machine Learning*, pages 554–561.

Thepchai Supnithi, Kiyotaka Uchimoto, Toyomi Saiga, Emi Izumi, Sornlertlamvanich Virach, and Hitoshi Isahara. 2003. Automatic proficiency level checking based on sst corpus. In *In Proceedings of RANLP*.

Yukio Tono. 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.

Nina Vyatkina. 2012. The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*. to appear.

David M. Williamson, Randy Elliot Bennett, Stephen Lazer, Jared Bernstein, Peter W. Foltz, Thomas K. Landauer, David P. Rubin, Walter D. Way, and Kevin Sweeney. 2010. Automated scoring for the assessment of common core standards. White Paper.

David M. Williamson. 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)*.

Bo Zhang. 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*.