Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks

Ildikó Pilán, Elena Volodina
Språkbanken, University of Gothenburg
Sweden
ildiko.pilan@svenska.gu.se
elena.volodina@svenska.gu.se

Torsten Zesch
Language Technology Lab
University of Duisburg-Essen
Germany
torsten.zesch@uni-due.de

Abstract

The lack of a sufficient amount of data tailored for a task is a well-recognized problem for many statistical NLP methods. In this paper, we explore whether data sparsity can be successfully tackled when classifying language proficiency levels in the domain of learner-written output texts. We aim at overcoming data sparsity by incorporating knowledge in the trained model from another domain consisting of input texts written by teaching professionals for learners. We compare different domain adaptation techniques and find that a weighted combination of the two types of data performs best, which can even rival systems based on considerably larger amounts of in-domain data. Moreover, we show that normalizing errors in learners' texts can substantially improve classification when in-domain data with annotated proficiency levels is not available.

1 Introduction

Data sparsity is a recognized problem in many machine learning based NLP approaches since the creation of data specifically collected and annotated for a certain task or language is time-consuming and costly. Previous attempts to overcome data sparsity include transferring knowledge between different types of data through the application of models from languages and tasks where sufficient data exists to the ones where data is unavailable or sparse (Daumé III and Marcu, 2006). A common case of such a transfer learning scenario is *domain adaptation*, where training and test data belong to different domains (e.g. text genres) referred to as *source domain* and *target domain* respectively.

In our experiments, we aim at exploring the plausibility of domain adaptation as a strategy for overcoming data sparsity in the context of foreign and second language (L2) learning. More specifically, we operationalize *domain* as the type of text involved in the language learning process: on the one hand, texts from coursebooks intended for L2 learners (referred to as *L2 input texts* in this paper), and on the other hand, essays created by learners (*L2 output texts*). Our goal is to predict L2 language development stages in terms of linguistic complexity in the latter category, i.e. learner-produced texts. These stages are commonly referred to as *proficiency levels* in second language acquisition and language testing. Levels range from 'absolute beginner' to 'advanced language user' with increasing linguistic complexity as learners progress with the levels. A scale of such levels, very influential both in Europe and outside, is the CEFR – Common European Framework of Reference for Languages (Council of Europe, 2001).

In previous work, NLP methods have been successfully applied to both assessing proficiency levels in L2 input texts collected from coursebooks and output texts written by learners (see section 2). However, the two text types have always been considered separately, while we argue that there is a shared linguistic content between the two that can be used for knowledge transfer. Specifically, the output of learners is a subset of the linguistic input that they are able to understand (Barrot, 2015). Thus, incorporating knowledge from coursebook texts representing L2 input may improve the classification of proficiency levels in L2 output text. Decreasing the need for a large amount of L2 output data is particularly appealing since acquiring this type of text poses a number of challenges including copyright issues, anonymization of sensitive information, and often even digitizing hand-written material (Megyesi et al., 2016; Mendes et al., 2016; Volodina et al., 2016). Since an increasing amount of people learn foreign languages worldwide either out of necessity or as a personal interest, systems targeting the needs of this user group are especially valuable. Within this context, the automatic assessment of proficiency levels in learner-produced texts would be a powerful tool for increasing both learners' autonomy and teaching professionals' efficiency.

Research Questions In particular, this paper aims at answering the following research questions: (i) Can we overcome the lack of a sufficient amount of learner output data by incorporating knowledge from L2 input texts when performing proficiency level classification? (ii) What kind of domain adaptation technique performs best in this context? (iii) Does normalizing errors in L2 learner output benefit proficiency level classification in a domain adaptation setting?

The motivation behind error normalization is that learner output typically contains errors which may influence the performance of automatic taggers and parsers and thus, classification performance. Therefore, error normalization may allow for a more precise calculation of feature values and a more successful transfer from and to a non error-prone domain. The amount and type of errors, i.e. degree of incorrectness, however, is not explicitly considered as an indicator of proficiency for L2 learner output in our experiments in order to keep comparability with coursebook texts. Unlike linguistic complexity, incorrectness is not a relevant aspect for L2 input texts as these are authored by teaching professionals and are supposed to be relatively error-free examples of language use.

Our target language of choice is Swedish, a language considerably less resource-rich than English and for which a CEFR-level classification model of L2 learners' writing is not available yet, despite the clear need for breaking down CEFR descriptors into linguistic constituents that characterize proficiency levels for each individual language (Little, 2011; North, 2007).

Main Findings We find that, in the absence of annotated learner-written data, using a classification model trained only on coursebook texts is a viable alternative if learner errors are normalized. Furthermore, if a small amount of learner output data is available and it is combined with L2 input texts, it can even outperform a model trained only on the few in-domain instances, resulting in a prediction quality matching that of in-domain state-of-the-art systems for other languages. In a domain adaptation setting, normalizing learner errors proved to yield a substantial improvement for features based on token, character and sentence counts as well as for features based on the CEFR-level distribution of tokens.

2 Text Categorization in the Language Learning Context

The automated evaluation of learner output is primarily a text classification task which aims at determining the quality of writing and assigning an appropriate label from a given set, for example a score or grade on the continuum between pass-fail (*essay scoring*) or a level indicating learning progress (*proficiency level classification*). In a L2 learning scenario, a longer piece of learner-written text is a popular means to assess learners' proficiency level. The human assessment of learner output, however, is both time-consuming and prone to subjectivity. Different linguistic dimensions need to be taken into consideration usually requiring several iterations of re-reading and different factors may influence the decision, such as negative attitude to a learner, hunger, bad mood, and boredom. Therefore, the number of initiatives to complement (or even replace) human assessment with a more objective and more efficient supervised machine learning system has been increasing the past years, with essay grading (Burstein and Chodorow, 2010) as an important application field.

2.1 Automatic Essay Scoring

Automatic essay scoring (AES) has been an active research area since 1990s, targeting mostly English (Burstein and Chodorow, 2010; Miltsakaki and Kukich, 2004; Page, 2003). Recently, with the availability of annotated learner corpora for other languages, automatic essay grading has expanded to cover also other languages, e.g. German (Zesch et al., 2015) and Swedish (Östling et al., 2013), to name just a few.

In its nature, AES has mostly relied on machine learning approaches, exploring both supervised (Yannakoudakis et al., 2011) and unsupervised methods (Chen et al., 2010) with different degrees of success.

Östling et al. (2013) have looked at Swedish upper secondary school essays, i.e. first language learner essays, and automatically assessed them in terms of a four-point scale of performance grades with an accuracy of 62%. The authors found that this result exceeded the agreement rate between two human assessors which was as low as 45.8% which might indicate that human-like performance is a rather uncertain goal. Linguistic parameters that have over time been presumed to be strong predictors of writing quality have varied from shallow ones like text and word length (Page, 2003; Östling et al., 2013) to more sophisticated features using Latent Semantic Analysis (Landauer et al., 2003), cosine similarity (Attali and Burstein, 2006), discourse structure and stylistic features (Attali and Burstein, 2006).

2.2 Proficiency Level Classification

A closely related task to AES is classifying texts into L2 proficiency levels which consists of predicting at which language learning stage a text can be produced or understood by a L2 learner, rather than assigning a grade within a pass-fail range. The CEFR, the scale of proficiency levels adopted in our experiments, contains guidelines for the standardization of language teaching and assessment across languages and countries (Council of Europe, 2001). It provides a common metalanguage to talk about objectives, assessment, (Little, 2011), and it defines language competences at six proficiency levels (A1, A2, B1, B2, C1, C2) where A1 is the beginner level. Since the publication of the CEFR guidelines in 2001, several countries have adopted the system, but its practical application has proven to be rather non-straightforward since the descriptions of the competences at each level remain vague (Little, 2011; North, 2007).

The past few years have seen an increasing interest in the CEFR-level classification of both L2 input and output texts. In the case of coursebook texts such a classification has also been referred to as *L2 readability* and it has been investigated for, among others, French (François and Fairon, 2012), Portuguese (Branco et al., 2014), Chinese (Sung et al., 2015), Swedish (Pilán et al., 2015), and English (Xia et al., 2016).

Apart from L2 input texts, CEFR-level annotated L2 learner corpora are also available for a number of languages including but not limited to English (Nicholls, 2003), Estonian (Vajjala and Lõo, 2014) and German (Hancke and Meurers, 2013). Moreover, MERLIN (Wisniewski et al., 2013) is a trilingual learner corpus comprised of written productions of L2 learners of Czech, German, and Italian also linked to CEFR levels. Despite the availability of annotated corpora for several languages, the number of projects targeting the automatic CEFR-level classification of learner essays has remained rather limited. Previously reported results for this task in terms of accuracy include 61% for German (Hancke and Meurers, 2013) and 79% for Estonian (Vajjala and Lõo, 2014).

2.3 Domain Adaptation for Tasks Related to L2 Learning

While there is a lot of previous work on domain adaptation in general, relatively few approaches exist in the field of assessing learner output texts. Previous applications of domain adaptation to learner essays focused on exploring the transfer of models between different writing tasks that prompted students to produce the essays, e.g. expressing an opinion on a topic vs. summarizing a news article (Zesch et al., 2015; Phandi et al., 2015). Zesch et al. (2015) explore which features are transferable from one essay grading task to another task based on a different prompt. They find that by excluding some highly domain-specific features, the transfer loss can be reduced significantly without noticeable differences in overall performance.

A popular domain adaptation approach is EASYADAPT (Daumé III, 2007) that augments the original feature space with source- and target-specific versions. Phandi et al. (2015) successfully applied EASYADAPT for automatic essay scoring and Xia et al. (2016) for the CEFR-level classification of L2 input texts with native language texts as source domain.

3 Datasets

For our experiments, we use L2 Swedish data including learners' output, i.e. error-prone essays written by learners, as well as L2 input data for learners, i.e. relatively error-free texts written by experts for

		CEFR Levels				
		A2	**B1**	**B2**	**C1**	**Total**
Learner Output	**Texts**	83	75	74	88	**320**
	Tokens	18,349	29,814	32,691	60,095	**140,949**
Expert Input	**Texts**	157	258	288	115	**818**
	Tokens	37,168	79,124	101,297	71,723	**289,312**

Table 1: Overview of CEFR-level annotated Swedish datasets.

L2 learners primarily intended as reading material. Both types of data are manually labeled for CEFR levels and automatically annotated across different linguistic dimensions including lemmatization, part-of-speech (POS) tagging, and dependency parsing using the Sparv (previously known as 'Korp') pipeline (Borin et al., 2012).

3.1 L2 Output Texts

Our source of output texts is SweLL (Volodina et al., 2016), a corpus consisting of L2 Swedish learner essays on a variety of topics, manually linked to CEFR levels. The essays also contain meta-information on learners' mother tongue(s), age, gender, education level, the exam setting, and, in certain cases, topic and genre. The distribution of essays per level is given in Table 1.

The corpus includes some essays at A1 and C2 levels, but these classes were too under-represented to be included in our experiments. As for A1 level, this may depend on learners' limited ability to write due to the lack of familiarity with many linguistic constructs. In fact, the CEFR contains no descriptor for writing essays and reports at A1 level (Council of Europe, 2001, 62). C2 is lacking since courses at this level are not provided, and it is in general characterized as a near-native language competence.

Since SweLL consists of learner-produced texts, it is likely that it contains some errors which, however, have not been annotated or normalized yet in the resource. The number of non-lemmatized tokens in the resource (i.e. tokens that could not be assigned baseforms during automatic annotation), which could indicate spelling errors or creative compounding at more advanced levels is higher at lower proficiency levels, but their amount always remains within a range of 5% and 8%.

3.2 L2 Input Texts

Our L2 input texts were collected from COCTAILL, a corpus of coursebooks used for teaching CEFR-based courses of L2 Swedish (Volodina et al., 2014). The coursebooks are divided into lessons (book chapters), each of which is labeled for the CEFR level it is aimed at. Each lesson contains a variety of elements including reading texts, exercises, lists, etc. Out of these only the texts intended for reading have been included in our dataset, whose CEFR level was derived from the level of the lesson they occurred in. Table 1 gives an overview of the distribution of these texts per level. For the same reasons as in section 3.1, C2 was not included in this dataset and A1 level has been omitted to keep the classes consistent between the two datasets.

4 Feature Set

We use the feature set presented in Pilán et al. (2015) designed for modeling linguistic complexity in input texts for L2 Swedish learners. These features rely on morpho-syntactic tags, information about the CEFR level of tokens, and aspects inspired by L2 Swedish curricula. Five sub-group of features can be distinguished in this set: length-based, (weakly) lexical, morphological, syntactic, and semantic features. The detailed list of features is presented in Table 2.

Count-based features rely on the number of characters and tokens (*tkn*), extra-long words being tokens longer than 13 characters. LIX (Läsbarhetsindex) is a traditional Swedish readability formula corresponding to the sum of the average number of words per sentence in the text and the percentage of

Count	Lexical	Syntactic	Morphological	
Sentence length	A1 lemma IS	Avg DepArc length	Modal V to V	Verb IS
Avg token length	A2 lemma IS	DepArc Len $>$ 5	Particle IS	V variation
Extra-long token	B1 lemma IS	Max length DepArc	3SG pronoun IS	Function W IS
Nr characters	B2 lemma IS	Right DepArc Ratio	Punctuation IS	Lex tkn to non-lex tkn
LIX	C1 lemma IS	Left DepArc Ratio	Subjunction IS	Lex tkn to Nr tkn
Bilog TTR	C2 lemma IS	Modifier variation	PR to N	Neuter N IS
Square root TTR	Difficult W IS	Pre-modifier IS	PR to PP	CJ + SJ IS
Semantic	Difficult N&V IS	Post-modifier IS	S-VB IS	Past PC to V
Avg senses per token	OOV IS	Subordinate IS	S-V to V	Present PC to V
N senses per N	No lemma IS	Relative clause IS	ADJ IS	Past V to V
	Avg. KELLY log freq	PP complement IS	ADJ variation	Present V to V
			ADV IS	Supine V to V
			ADV variation	Relative structure IS
			N IS	Nominal ratio
			N variation	N to V

Table 2: Feature set.

tokens longer than six characters (Björnsson, 1968). Rather than a simple type-token ratio (TTR), we use a bi-logarithmic and a square root equivalent following Vajjala and Meurers (2012).

Lexical features incorporate information from the KELLY list (Volodina and Kokkinakis, 2012), a frequency-based word list compiled using a corpus of web texts (thus completely independent of our datasets), which also provides a suggested CEFR level per each lemma based on frequency bands. For some feature values, *incidence scores* (IS) are computed, in other words, instead of absolute counts, normalized values per 1000 tokens are considered to reduce the influence of sentence length. Lexical complexity is modeled with a set of weakly lexicalized features, i.e. we do not use word forms or lemmas themselves as features, but the IS of their corresponding CEFR levels instead. This aspect is especially important considering the limited size of our learner essay data. *Difficult* tokens are those that belong to levels above the overall CEFR level of the text. Moreover, we consider the IS of tokens not present in KELLY (OOV IS), the IS of tokens for which the lemmatizer could not identify a corresponding lemma (No lemma IS), as well as average KELLY log frequencies.

Morphological features include not only IS but also variational scores, i.e. the ratio of a category to the ratio of lexical tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). The IS of all lexical categories as well as the IS of punctuation, particles, sub- and conjunctions (SJ, CJ) are taken into consideration. Nominal ratio (Hultman and Westman, 1977) is another readability formula proposed for Swedish that corresponds to the ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR) adverbs, and verbs. Relative structures consist of relative adverbs, determiners, pronouns and possessives. Some features are inspired by L2 teaching material (Fasth and Kannermark, 1997) and they are based on fine-grained inflectional information such as the IS of neuter gender nouns and the ratio of different verb forms to all verbs.

Syntactic features are based, among others, on the length (depth) and the direction of dependency arcs (DepArc). Within this feature group, we consider also relative clauses as well as pre- and post-modifiers, which include, for example, adjectives and prepositional phrases respectively.

Semantic features build on information from the SALDO lexicon (Borin et al., 2013). We use the average number of senses per token and the average number of noun senses per nouns.

5 Experimental Setup

For all experiments, we use SVMs as implemented in WEKA (Hall et al., 2009) and the feature set presented in detail in section 4. Results are obtained using 10-fold cross-validation. We report the F_1 score, i.e. the harmonic mean of precision and recall, as well as quadratic weighted kappa (κ^2), a distance-based scoring function taking into consideration also the degree of misclassifications.

Experimental setup	Data used	# Training inst.	# Informing inst.
MAJORITY	D_T	288	320
IN-DOMAIN	D_T	288	320
SOURCE-ONLY	D_S	818	818
EASYADAPT	D_S with augmented features	818	1138
+FEATURE	D_T with D_S prediction as feature	288	1138
COMBINED	D_S + 60% of D_T	1010	1010
WEIGHTED	D_S ($w = 1$) + 60% of D_T ($w = 10$)	1010	1010
WEIGHTED-INSTSEL	Correctly classified D_S ($w = 1$) + 60% of D_T ($w = 10$)	505	1138

Table 3: Domain adaptation experimental setups.

5.1 Domain Adaptation

In a domain adaptation scenario, data from a source domain (D_S) is used to predict labels in a different, target domain (D_T). To overcome data sparsity, especially relevant for our learner essay data, we experiment with improving CEFR level classification by transferring information from our D_S consisting of L2 coursebook texts to D_T consisting of Swedish L2 learners' essays.

As baselines, we employ both assigning the most frequent label in the dataset (MAJORITY) and an IN-DOMAIN setup using only the learner essays in a cross-validation setup. We compare these to different domain adaptation scenarios inspired mostly by Daumé III and Marcu (2006) and Pan and Yang (2010) which differ in the type and the amount of data used as detailed in Table 3. We report the number of instances employed at the moment of training as well as the amount of instances from which information has been incorporated in some form in the final models.

In the **SOURCE-ONLY** setup, a model trained on all available source domain instances, i.e. coursebook texts, was applied directly to the target domain instances consisting of learner essays. **EASYADAPT** (Daumé III, 2007) is a feature augmentation approach which consists of triplicating the feature space by including three versions of each feature in the augmented equivalent: a general, a source-specific and a target-specific version. In more formal terms, to each feature vector x, the mapping function $\phi^S(x) = \langle x, x, 0\rangle$ is applied in the source domain and $\phi^T(x) = \langle x, 0, x\rangle$ in the target domain, 0 being a zero vector of length $|x|$. In **+FEATURE** we first train a model trained on the L2 input texts. Then, the CEFR label predicted by this system is incorporated as an additional feature for each essay instance and a new model is trained on the essays with this extra dimension. For **COMBINED** and **WEIGHTED** the training data includes not only D_S instances, but also 60% of D_T. In the WEIGHTED setup, an increased importance is given to D_T instances during training through the assignment of a higher weight (w). Finally, to obtain **WEIGHTED-INSTSEL**, we first train a model on the available D_T data and use that to classify D_S instances. Then those D_S instances that the essay-only model correctly classified are combined with 60% D_T, the latter ones receiving a weight of 10. Compared to WEIGHTED, in this setup we discard D_S instances that might be misleading when making predictions on D_T, due to differences in the underlying distributions in the two domains. A similar approach is presented in Jiang and Zhai (2007).

5.2 Error Normalization

Besides using learners' output texts in their original form, we investigate also the effects of error normalization on the domain-adapted strategies. By correcting errors we aim at bringing learners' texts closer to the standard language present in the coursebooks. Making the texts belonging to these two different domains more similar to each other may improve the domain-adapted classification performance. Moreover, since the annotation tools used were originally designed for dealing with standard Swedish, error normalization leads to a more reliable tagging and parsing, and hence to more precise feature values in the corrected learner output texts.

Previous error-normalization approaches include, among others, finite state transducers (Antonsen, 2012) and a number of, mostly hybrid, systems created within the CoNLL Shared Task on grammatical error correction for L2 English (Ng et al., 2014).

	ORIGINAL		ERROR-NORMALIZED	
	F_1	κ^2	F_1	κ^2
MAJORITY	.120	.000	.120	.000
IN-DOMAIN	.721	.886	.720	.872
SOURCE-ONLY	.438	.713	.620	.807
EASYADAPT	.503	.681	.533	.741
+FEATURE	.709	.879	**.802**	.864
COMBINED	.733	.863	.726	.885
WEIGHTED	**.747**	**.890**	.779	**.915**
WEIGHTED-INSTSEL	.733	.873	.795	.914

Table 4: Domain adaptation results with and without error normalization.

We use LanguageTool[1] (Naber, 2003), an open-source rule-based proof-reading program available for multiple languages which detects not only spelling, but also some grammatical errors (e.g. inconsistent gender use in inflected forms). We propose a two-step algorithm consisting of first obtaining correction candidates from LanguageTool and then ranking these candidates based on a word co-occurrence measure. As a first step, we identify errors in the learner essays and a list of one or more LanguageTool correction suggestions, as well as the *context*, i.e. the surrounding tokens for the error within the same sentence. When more than one correction candidate is available, as an additional step, we make a selection based on *Lexicographers' Mutual Information* (LMI) scores (Kilgarriff et al., 2004). Here we assume a positive correlation between a correction candidate co-occurring with a context word and being the correct version of the word intended by the learner. We check LMI scores for each LanguageTool correction candidate paired with the lemma of each available noun, verb, and adjective in the context based on a pre-compiled list of LMI scores. We create this list using a Korp API (Borin et al., 2012) providing LMI scores computed based on a customizable set of corpora. We use a variety of modern Swedish corpora totaling to more than 209 million tokens for our list of LMI scores. Only scores for noun-verb and noun-adjective combinations have been included with a threshold of LMI $\geq$ 50. When available, we select the correction candidate maximizing the sum of all LMI scores for the context words. In the absence of LMI scores for the pairs of correction candidates and context words, the most frequent word form in Swedish Wikipedia texts is chosen as a fallback.

Once correction candidates are ranked, each erroneous token identified by LanguageTool is replaced in the essays by the top ranked correction candidate. The normalized texts are then annotated again and feature values are re-computed.

6 Results and Discussion

Table 4 presents the results of our domain adaptation experiments first without error normalization (*original*) and then with corrected errors (*error-normalized*). In the case of the non-normalized essays, the in-domain baseline obtained using only the small amount of learner output texts in a cross-validation setup is .721 F_1 and .886 κ^2. Compared to this, transferring a model based on coursebook texts directly (SOURCE-ONLY) results in a considerable performance drop (-.283 F_1 and -.173 κ^2). When using the essays in their original, noisy form, the best performing domain adaptation setup is the weighted combination of L2 input and output texts, which outperforms even the in-domain baseline both in terms of F_1 and κ^2.

The obtained domain-adaptation results are comparable to state-of-the-art in-domain systems for other languages, like the system for Estonian described in Vajjala and Lõo (2014) with an F_1 of .78, or the one for German (Hancke, 2013) with .71 F_1 for a feature selected model distinguishing 5 classes. It is worth

[1] www.languagetool.org

	In-domain		Source-Only (Original)		Source-Only (Error-norm.)	
Feature Group	F_1	κ^2	F_1	κ^2	F_1	κ^2
All	.721	.886	.438	.713	.620	.807
Count	.499	.740	.106	-.003	*.335*	***.708***
Lexical	**.625**	**.826**	**.318**	**.507**	**.378**	.626
Syntactic	.511	.665	.118	.066	.106	.030
Morphological	.538	.743	.297	.403	.291	.419
Semantic	.299	.198	.087	.000	.087	.000

Table 5: Performance of individual feature groups.

noting, however, that both of these systems required a considerably (about three times) larger annotated in-domain corpus. This shows that additional coursebook data can benefit the classification of language proficiency levels in learner output texts, especially if only a small amount of annotated in-domain data is available.

Error Normalization Our error-normalization method corrects in total 5,080 errors in the essays which amounts to 3.6% of all tokens in the data. In absence of error-annotated Swedish resources, we manually evaluate the method by inspecting 120 normalized items out of which we find 83 correct, corresponding to 69% accuracy. Out of the normalized tokens, about 87% are categorized as spelling errors by LanguageTool. Moreover, the choice of correction candidate is based on LMI scores in 24% of all cases.

Since our feature set does not target learner errors specifically (to be able to maintain comparability when applied to coursebook text), we do not expect error normalization to influence classification results with In-Domain. Our experiment results in Table 4 show, in fact, that correcting learner errors does not have any statistically significant effect in the In-Domain setup, but it does improve performance to a great extent for most domain-adapted cases. This latter would support the hypothesis that correcting spelling and grammatical errors increases the similarity between the target and the source domain. The gain is especially large (+.182 F_1) in the case of the SourceOnly setup, which does not rely on annotated essays. EasyAdapt, which has been successfully used in an AES task previously (Phandi et al., 2015), is outperformed by most other domain adaptation methods in our case, independently from error normalization.

In terms of F_1, +Feature using the predictions of a classifier trained on the L2 input texts performs best (.802 F_1), however, the degree of misclassifications indicated by κ^2 is smallest with Weighted (.915), as in the case of the essays without error normalization. After error correction, Weighted-InstSel achieves approximately the same quality of performance for all measures as the aforementioned two best performing models Weighted and +Feature. These all improve over the In-domain baseline by about .07 F_1 and .03 κ^2.

These results show that the knowledge transfer from L2 input texts can be substantially boosted by normalizing errors in the learner-produced texts.

Contribution of Feature Groups In the next step, we investigate the contribution of individual feature groups to the classification performance both in- and cross-domain with the Source-only setup which does not presuppose the availability of annotated in-domain data. Results for our ablation test are shown in Table 5.

The most predictive features in- and cross-domain on both the original and on the normalized essays are lexical features measuring the proportion of tokens per CEFR level in the texts. Morphological features also preserve their strong predictive power when transferred between L2 input and output texts. The informativeness of syntactic and count features is very low in the cross-domain setting with the original essays, but the latter category transfers much better after error-normalizing L2 output texts. A potential explanation could be that error normalization includes also corrections of capitalization and whitespaces which might contribute to an improved detection of sentence boundaries, a central element in most of these features. Lexical features also benefit from error correction, presumably due to a more precise estimation of the CEFR-level distribution of tokens.

Direction of Misclassifications Finally, to investigate whether the transferred coursebook model predicts learner-written texts to be of higher or lower proficiency levels compared to the available annotations, we perform regression using SMO and the SOURCEONLY setup, transforming CEFR levels into numeric values. We use the normalized essays for this purpose since the automatic annotation is presumably more precise in these texts compared to their original version. Predictions within a distance of 0.5 from the numeric value representing the actual CEFR level are considered sufficiently close for being considered correct, thus the amount of errors is computed based only on cases exceeding this margin. The regression model produces .800 correlation and 1.120 RMSE (root mean squared error). We find that 64% of the erroneous predictions consider essays to be of a lower level than they actually are. This could be a data-driven confirmation of the pedagogical observation that learners' output texts are typically of a lower linguistic complexity compared to the L2 input texts written for them within the same CEFR level.

7 Conclusions

In this work we investigated the benefits of using texts from language learning coursebooks to classify proficiency levels in learner-written texts, since the latter type of data is especially costly to collect. Moreover, our experiments provide useful insights into how some simple domain adaptation techniques compare to each other for this task. Training only on source domain data did not yield a successfully transferable model between the L2 input and output texts if errors were not normalized in the learner-produced essays. With such a normalization, however, using only coursebook texts as training data produced a result rather close to what learning only from a small amount of essays did. Joining domains was useful, especially when weighted target domain instances were added to all, or a subset of the coursebook data, and learner errors were normalized. We showed that, with these two steps, it is possible to outperform a model based only on a limited amount of in-domain data. Furthermore, our results are competitive even compared to systems for other languages that make use of a considerably larger amount of in-domain data.

In the future, it would be informative to repeat the experiments for other languages, where we expect similar results. Additional domain adaptation techniques could also be explored for this task, for example, the identification of shared priors and kernel transformations. Alternatives to the current error normalization could be investigated in order to identify a broader range of incorrect tokens more precisely. More reliable error correction methods may yield further improvement to transferring classification models between these domains.

References

Lene Antonsen. 2012. Improving feedback on L2 misspellings-an FST approach. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, number 080, pages 1–10. Linköping University Electronic Press.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Jessie Barrot. 2015. Comparing the linguistic complexity in receptive and productive modes. *GEMA Online Journal of Language Studies*, 15(2):65–81.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *LREC*, pages 474–478.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. *Computational Processing of the Portuguese Language. Springer*, pages 256–261.

Jill Burstein and Martin Chodorow. 2010. Progress and New Directions in Technology for Automated Essay Evaluation. *Oxford University Press.*

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, Tao-Hsing Chang, et al. 2010. An unsupervised automated essay scoring system. *IEEE Intelligent systems*, 25(5):61–67.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Press Syndicate of the University of Cambridge.

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Cecilia Fasth and Anita Kannermark. 1997. *Form i focus: övningsbok i svensk grammatik. Del B.* Folkuniv. Förlag, Lund.

Thomas François and Cédrick Fairon. 2012. An 'AI readability' formula for French as a foreign language. In *Proceedings of the EMNLP and CoNLL 2012.*

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.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference.*

Julia Hancke. 2013. Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. *Master's thesis, University of Tübingen.*

Tor G Hultman and Margareta Westman. 1977. *Gymnasistsvenska.* Liber.

Jing Jiang and ChengXiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague, Czech Republic. Association for Computational Linguistics.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105:116.

Thomas K Landauer, Darrell Laham, and Peter W Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

David Little. 2011. The Common European Framework of Reference for Languages: A research agenda. *Language Teaching, Vol 44.3.*

Beáta Megyesi, Jesper Näsman, and Anne Palmér. 2016. The Uppsala Corpus of Student Writings: Corpus Creation, Annotation, and Analysis. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).*

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 Corpus: a Learner Corpus for Portuguese. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).*

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(01):25–55.

Daniel Naber. 2003. A rule-based style and grammar checker. Master's thesis, Bielefeld University, Bielefeld, Germany.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant, editors. 2014. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task.* Association for Computational Linguistics, Baltimore, Maryland.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Brian North. 2007. The CEFR illustrative descriptor scales. *The Modern Language Journal 91*.

Ellis Batten Page. 2003. Project essay grade: PEG. *M.D. Shermis and J.C. Burstein, editors, Automated essay scoring: A cross-disciplinary perspective*, pages 43–54.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *To appear in International Journal of Computational Linguistics and Applications*. Available at `http://arxiv.org/abs/1603.08868`.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. *NEALT Proceedings Series Vol. 22*, pages 113–127.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173. Association for Computational Linguistics.

Elena Volodina and Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of LREC*, pages 1040–1046.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *NEALT Proceedings Series Vol. 22*, pages 128–144.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European reference levels in authentic learner data. In *ICT for Language Learning 2013, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni.*

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben 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*, pages 180–189. Association for Computational Linguistics.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-Independent Features for Automated Essay Grading. In *Proceedings of the Building Educational Applications Workshop at NAACL.*

Robert Östling, André Smolentzov, Björn Tyrefors, and Erik Höglin. 2013. Automated Essay Scoring for Swedish. In *The 8th Workshop on Innovative Use of NLP for Building Educational Applications.*