Adapting to Learner Errors with Minimal Supervision

Alla Rozovskaya*
Queens College,
City University of New York

Dan Roth**
University of Illinois at
Urbana-Champaign

Mark Sammons**
University of Illinois at
Urbana-Champaign

This article considers the problem of correcting errors made by English as a Second Language writers from a machine learning perspective, and addresses an important issue of developing an appropriate training paradigm for the task, one that accounts for error patterns of non-native writers using minimal supervision. Existing training approaches present a trade-off between large amounts of cheap data offered by the native-trained models and additional knowledge of learner error patterns provided by the more expensive method of training on annotated learner data.

We propose a novel training approach that draws on the strengths offered by the two standard training paradigms—of training either on native or on annotated learner data—and that outperforms both of these standard methods. Using the key observation that parameters relating to error regularities exhibited by non-native writers are relatively simple, we develop models that can incorporate knowledge about error regularities based on a small annotated sample but that are otherwise trained on native English data.

The key contribution of this article is the introduction and analysis of two methods for adapting the learned models to error patterns of non-native writers; one method that applies to generative classifiers and a second that applies to discriminative classifiers. Both methods demonstrated state-of-the-art performance in several text correction competitions. In particular,

* Department of Computer Science, Queens College, CUNY, Queens, NY 11367 USA. E-mail: arozovskaya@qc.cuny.edu. Part of this work was done when the author was at Virginia Tech and the University of Illinois.

** Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. E-mail: {danr, mssammon}@illinois.edu.

Submission received: 14 May 2016; revised version received: 14 February 2017; accepted for publication: 2 June 2017.

doi:10.1162/COLI_a_00299

the Illinois system that implements these methods ranked at the top in two recent CoNLL shared tasks on error correction.[1] *We conduct further evaluation of the proposed approaches studying the effect of using error data from speakers of the same native language, languages that are closely related linguistically, and unrelated languages.*[2]

1. Introduction

This article addresses the problem of correcting grammatical context–sensitive mistakes made by English as a Second Language (ESL) writers, a subject that has recently attracted significant attention (Izumi et al. 2003; Han, Chodorow, and Leacock 2006; De Felice and Pulman 2008; Gamon et al. 2008; Tetreault, Foster, and Chodorow 2010; Gamon 2010; Rozovskaya and Roth 2010a,b, 2011; Dahlmeier and Ng 2011, 2012). ESL error correction is an important problem because the majority of people who write in English today are not native English speakers (Gamon 2011). Representative ESL errors—a verb agreement mistake, an unnecessary article, and an incorrect choice of a preposition— are illustrated in Figure 1.[3] The **source word** (the word chosen by the ESL writer) and the **label** (the correct choice, as specified by a native English speaker) are emphasized.

A standard approach to dealing with these errors, which also proved highly successful in text correction competitions (Dale and Kilgarriff 2011; Dale, Anisimoff, and Narroway 2012; Ng et al. 2013, 2014), makes use of a *machine-learning classifier paradigm* and is based on the methodology for correcting *context-sensitive spelling mistakes* made by native speakers. With the exception that learners and native writers exhibit errors of different types, most of the grammar and usage mistakes made by non-native speakers of English also fall into the category of context-sensitive errors that result in valid English words (e.g., articles or prepositions) being confused.

Traditionally, following the work on context-sensitive spelling, classifiers have been trained on native English data for a particular mistake type (e.g., preposition). Note that because native data do not have information about learner errors, the model can only use contextual cues. Thus, when the resulting classifier is applied to non-native text, the most appropriate preposition is selected based exclusively on the surrounding context, similar to a cloze task where one needs to "guess" a word that has been replaced by blank in a sentence. An error is flagged if this most likely candidate is different from the author's choice (Eeg-Olofsson and Knuttson 2003; Izumi et al. 2003; Han, Chodorow, and Leacock 2006; De Felice and Pulman 2008; Gamon et al. 2008; Tetreault and Chodorow 2008; Tetreault, Foster, and Chodorow 2010). We call this general approach to error correction the **selection** training paradigm.

In realistic ESL situations, however, the scenario is different and there is additional information that could be used by a correction system beyond that used in the selection paradigm. The ESL learner writes a text, and typically makes mistakes on

1 The Illinois system ranked first in all metrics in the CoNLL-2013 competition and scored second and first on original and revised annotation metrics, respectively, in the 2014 competition.

2 This article unifies and significantly extends material that appeared previously in Rozovskaya and Roth (2010b, 2011), Rozovskaya, Sammons and Roth 2012, and Rozovskaya and Roth (2014).

3 The example is taken from the ICLE corpus (Granger, Dagneaux, and Meunier 2002).

The ruling class of the Soviet Union is represented in Orwell's book by pigs, who **gives/give* themselves **the/∅* education and privileges, use the power and the force of terror provided by Napoleon's dogs to control the other animals in ways that are different from those **by/of* the previous ruling group.

Figure 1
Examples of representative ESL errors.

fewer than 10% of words.[4] This baseline accuracy is higher than the performance of the state-of-the-art classifiers on word selection tasks (Han, Chodorow, and Leacock 2006; Tetreault and Chodorow 2008). This high baseline also suggests that by just selecting the author's word choice we can already do better than using the context alone. In addition to this baseline bias, the author's word choice may also be informative because non-native speakers do not make mistakes randomly. For example, it is well known that the native language of the learner makes a distinctive impact on their usage of English (Odlin 1989; Gass and Selinker 1992; Montrul and Slabakova 2002; Ionin, Zubizarreta, and Bautista 2008). Therefore, we would like to use the author's word when making a decision about the appropriate correction.

One way the system can consider the word used by the writer at evaluation time is by proposing a correction only when the confidence of the classifier is high enough. However, in this partial solution, the source word is not used in training if the classifier is trained on native data. Alternatively, one can train on corrected non-native text. In that case, error patterns provide the model with additional information that is not available when training on native data. Indeed, in certain scenarios models trained on annotated learner data perform better than those trained on larger amounts of native data (Gamon 2010; Han et al. 2010; Dahlmeier and Ng 2011).

Training on native vs. annotated learner data raises the question of the trade-off between using large amounts of cheap native data and the availability of additional information provided by the expensive supervision in the form of annotated ESL data. We propose an approach that draws on the advantages of the two training sources by combining them in one model. The approach provides models trained on native data with knowledge about error regularities, *with minimal annotation costs*: A small amount of annotated learner data is used to extract knowledge about error patterns that are then "injected" into models trained on large amounts of native data. We call this approach of combining a large quantity of native training data with a small amount of annotated ESL data **adaptation with minimal supervision** and denote models trained in this way as **adapted** to learner errors.

The key contribution of this work is an analysis of novel methods of training for error correction tasks that use minimal supervision and draw on the strengths of the two standard approaches—training on native or annotated learner data—and the advantages provided by each training source. Within this method, we describe how

4 Such error rates are typical for data sets that represent learners with a college level of English; one would expect higher error rates for beginning-level learners but lower error rates for high-proficiency writers, for example, the Helping Our Own (HOO) shared task (Dale, Anisimoff, and Narroway 2012). In addition, error rates may vary by error type. For example, in the FCE corpus (Yannakoudakis, Briscoe, and Medlock 2011) around 10% of preposition occurrences are errors, and some types of noun and verb errors occur in fewer than 5% of occurrences.

error patterns are learned, and introduce two approaches for "injecting" knowledge about error regularities into a model otherwise trained on native English data. These two approaches are designed to work with two state-of-the-art machine learning algorithms. The first approach, implemented within the discriminative learning framework, is based on generating artificial errors in training, where artificial mistakes are intended to mimic those observed in non-native data. The second approach adapts to error regularities by updating the prior distribution over the correction candidates. We show that the resulting adapted models are superior to the standard methods of training either on annotated or native data alone. This is because, in contrast to training on annotated ESL data, the adaptation approach only requires a small amount of annotation to estimate the parameters related to error regularities, while context parameters can be learned from native data. In sum, the proposed methods allow us to combine the advantages of training on native and annotated learner data. The proposed adaptation framework was implemented as part of the Illinois system that came first in several text correction competitions, including the prestigious CoNLL shared tasks (Rozovskaya et al. 2014, 2013). We further evaluate the proposed approaches and study the effect of adaptation when using error data from speakers of the same native language, languages that are closely related linguistically, and unrelated languages. This article unifies and significantly extends material that appeared previously in Rozovskaya and Roth (2010b, 2011, 2014), and Rozovskaya, Sammons and Roth (2012). The novel contribution is concentrated in Section 5, and evaluates the adaptation approach by comparing performance when error statistics are drawn from the writer's native (target) language data vs. from data generated by writers of other first-language backgrounds. We also study the effect of language relatedness on the quality of adaptation, namely, whether using data from writers whose first language is linguistically related to the author's native language is better than using data from writers whose native language is linguistically unrelated. An additional contribution is introduced in Section 6. There we explain the significance of the classification approach by reviewing and reassessing the current approaches to grammatical error correction, both from the point of view of developing systems that can focus on specific error phenomena, and from the perspective of the amount of supervision (annotated learner data) that is available and is needed for training.

This article is organized as follows. Section 2 provides the relevant background and explains the training paradigms. Section 3 presents the framework for adaptation with minimal supervision. In Section 4, we present key experimental results on three types of common ESL mistakes. Language-specific adaptation and analyses are presented in Section 5. Section 6 reviews related work. We conclude with a discussion of the adaptation approaches, the advantages of each of these algorithms, and situations in which one may be preferred over the other in Section 7.

2. Training Paradigms

We start with an overview of the machine learning framework for error correction, then present the standard word *selection* approach and show why this paradigm is not appropriate, as is, for error correction tasks. Next, we describe the correction training paradigm.

The Machine Learning Classifier Framework The machine-learning classifier methodology of training on native error-free data has been adopted for correcting ESL mistakes from *context-sensitive spelling correction*, a task that involves correcting spelling

errors that result in legitimate words, such as confusing *peace* and *piece* or *your* and *you're* (Roth 1998). Correcting these errors requires consideration of the context around the target word. Because the relevant contextual information may depend on various linguistic dimensions and is highly variable, the dominant approach to correcting these errors has been to use machine learning algorithms (Golding and Roth 1996, 1999; Banko and Brill 2001; Carlson, Rosen, and Roth 2001; Carlson and Fette 2007).

In the machine learning approach, we are given a *candidate set* or a *confusion set* of confusable words, for example, {*piece, peace*}. In training, each occurrence of a confusable word is represented as a vector of features derived from a *context window* around it. A classifier is trained on text assumed to be error-free, where each target word occurrence (e.g., *peace*) is treated as a positive training example for the corresponding word. Given a text to correct, for each confusable word, the task is to select the most likely candidate from the relevant confusion set.

The machine learning classifier approach has been and remains one of the prevalent methods in ESL error correction, as is evidenced by the competitions devoted to grammatical error correction: HOO-2011 (Dale and Kilgarriff 2011), HOO-2012 (Dale, Anisimoff, and Narroway 2012), CoNLL-2013 (Ng et al. 2013), and CoNLL-2014 shared tasks (Ng et al. 2014). Thanks to these competitions, the field has also seen a number of alternative approaches. For example, the CoNLL shared tasks made available a large annotated learner data set, that enabled the machine translation approach (Felice et al. 2014; Junczys-Dowmunt and Grundkiewicz 2014) that showed competitive performance in CoNLL-2014. In this work, our focus is on the classifier-based approach with an emphasis on techniques that allow for building robust models by leveraging large amounts of native English data *without the use of expensive annotation*.

The Selection Training Paradigm In the application of the *selection* training approach to ESL error correction, a model is tailored toward one mistake type (e.g., errors involving preposition usage) and is trained on well-formed native English text with features defined based on the surrounding context. Task-specific confusion sets are formed. For instance, in preposition error correction, it is common to include the top n most frequent English prepositions. The features are generally based on the surface form, part-of-speech information, and syntactic function of words in the immediate context around the potentially erroneous word. The classifier is then applied to non-native text to select the most appropriate candidate from the confusion set in context.

A number of earlier works in ESL error correction followed the selection training paradigm, with slight differences in the choice of features and machine-learning algorithm. The latter included maximum entropy models, perceptrons, language models, and decision trees (Eeg-Olofsson and Knuttson 2003; Izumi et al. 2003; Han, Chodorow, and Leacock 2006; Chodorow, Tetreault, and Han 2007; De Felice and Pulman 2007, 2008; Gamon et al. 2008; Tetreault and Chodorow 2008; Yi, Gao, and Dolan 2008; Bergsma, Lin, and Goebel 2009; Tetreault, Foster, and Chodorow 2010).

The Source Word The key reason that the selection approach has been popular in error correction is that it does not require annotated data. Native data (presumed to be correct) is used for training the system; it is cheap and is available in large quantities. It is clear, though, that there is a problem with this standard approach of training on native data, as its decision is based solely on the context and ignores the author's word choice.

The author's word choice or the source word is an important piece of information, as non-native speakers make mistakes in a systematic manner. To begin with, learner

performance is high, and for many error types, fewer than 10% or even 5% of word usages are actually mistakes (Han et al. 2010; Rozovskaya and Roth 2010b; Dale and Kilgarriff 2011; Yannakoudakis, Briscoe, and Medlock 2011).[5] This high learner performance is better than the performance of the state-of-the-art classifiers on word selection tasks not only on learner texts but also on well-formed data (e.g., Han, Chodorow, and Leacock 2006). Furthermore, learner errors follow specific error patterns, as discussed in more detail in the next section. These error patterns may be prominent across multiple first languages or be first-language dependent. The effect of "language transfer"—applying knowledge from the native language, when learning a foreign language—has been the subject of considerable study in the second-language acquisition literature (Odlin 1989; Gass and Selinker 1992; Montrul 2000; Montrul and Slabakova 2002; Oh and Zubizaretta 2003; Ionin, Zubizarreta, and Bautista 2008). These facts have also been confirmed empirically by studies that quantitatively examine learner corpora (Han, Chodorow, and Leacock 2006; Lee and Seneff 2008). For example, speakers of languages that do not have a determiner system (e.g., Russian) tend to make 4–5 times more article mistakes in English than speakers whose first language has articles (Rozovskaya and Roth 2010b). In addition to the error regularities due to first language influence, some confusions are much more likely to occur than others across multiple first languages. For example, regardless of the first language, ESL writers are 38 times more likely to incorrectly use "in" rather than "by" in place of the correct word "on" (Table 2, Section 3.1).

Training for Correction Tasks Depending on whether the author's word choice is used by the model, we distinguish between two training paradigms. In the *selection* paradigm, the decision of the classifier depends only on the context around the author's word, whereas in the *correction* training paradigm, both the context and the source word are used. The two training paradigms and the information available to the models in training and at prediction time in each case are illustrated in Figure 2. A straightforward way to use the source word is to train on annotated learner data. In that case, the source word can also be used as a feature. The problem is that this approach requires large amounts of annotated ESL data (Gamon 2010).

Training on native versus annotated learner data raises the question of the trade-off between the useful information provided in the form of expensive supervision and the robustness obtained from training on large amounts of native data. In this article, we address the question of what is the most appropriate way to train for correction tasks given the limitations and advantages of each data source. We propose an approach to building models that incorporates the best of both modes: training on native texts to facilitate the possibility of training from large amounts of data without the need for annotation, but using the correction mode with a modest amount of annotated ESL data so that the model can adapt to writers' errors. As a result, at evaluation time, these models can make use of the potentially erroneous information provided by the writer. We show that adaptation is beneficial on various levels: when error patterns are specific for a given first language and for groups of linguistically-related languages, as well as when error patterns are collected across multiple first language backgrounds that are available and that may include various related or unrelated languages.

5 We stress that, though seemingly low, these error rates correspond to at least one mistake in every sentence.

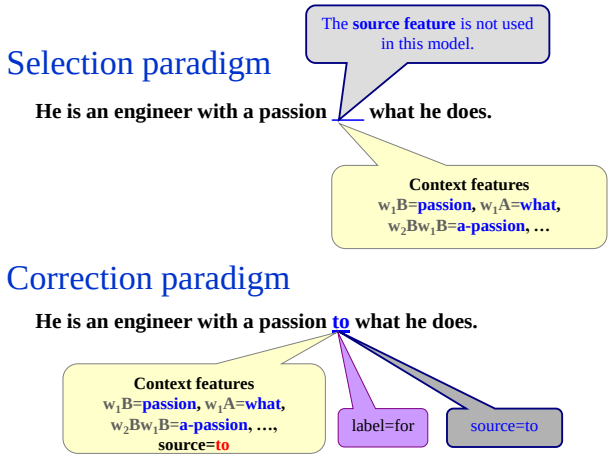

Figure 2
Two training paradigms for error correction tasks. In the *selection* training paradigm, models use only contextual information in training and when making a prediction. In the *correction* paradigm, the author's word is also used by the model.

3. Adaptation to Learner Errors with Minimal Supervision

In the previous section, we argued that the key advantage of the selection paradigm is that it is cheap, as we can use large amounts of native data. However, because learners' errors are systematic, we would like to provide the models with knowledge about typical mistakes learners make. As discussed earlier and as we show in the rest of the article, using the source word in prediction is better than just using the context. We therefore want to train on ESL data but we have a very limited amount of it. We thus need to develop ways that can use ESL data in prediction even when very little of it is available, which we do by exploiting large amounts of native data together with a small amount of ESL data. We call this approach **adaptation** to learner errors. The rest of this section describes the adaptation algorithms that we develop for this purpose.

After we describe how error patterns are learned, two adaptation methods are presented. The proposed methods are designed to work with two state-of-the-art machine learning algorithms: The *artificial errors* adaptation method is applicable for a discriminative learning algorithm (implemented here within Averaged Perceptron – henceforth AP), and the *priors adaptation* method for the Naive Bayes (NB) algorithm. These two algorithms demonstrated superior performance in a study that compared several learning frameworks for ESL error correction tasks (Rozovskaya and Roth 2011). This study also included language models and other count-based methods.

3.1 Learning Error Patterns

Error patterns are extracted from annotated learner data; these error patterns are referred to as **error statistics**. As we show here, unlike the context parameters, the error statistics are very simple, and thus we only need a small annotated sample to estimate them.

Given a specific task, we collect all source/label pairs from the annotated sample, where both the source and the label belong to the confusion set, and generate two

Table 1
Confusion matrix for article errors used in the artificial errors method. Based on the training data from the FCE corpus. The left column shows the correct article. Each row shows the author's article choices for that label and *Prob*(*source*|*label*). The numbers next to the targets show the count of the label (or source) in the data set.

Label	**Sources**		
	a (7,408)	∅ (44,916)	the (16,719)
a (7,945)	**0.885**	0.099	0.016
∅(44,096)	0.006	**0.977**	0.017
the (17,002)	0.008	0.060	**0.932**

confusion matrices, one where each cell represents *Prob*(source=s|label=l) (used in the artificial errors method, Section 3.2), and the other, inverse matrix, where each cell shows *Prob*(label=l|source=s), which is used by the priors adaptation method for NB (Section 3.3).

Tables 1, 2, and 3 show confusion matrices of the first type for article, preposition, and verb agreement mistakes, respectively. These matrices are based on the training part of the FCE data set, a corpus of annotated ESL essays (see Section 4.1). Consider, for example, the preposition matrix: We show eight rows and columns, but the entire table contains 12 rows and columns, one for each of the top 12 English prepositions;[6] each entry shows $Prob(p_i|p_j)$, the probability that the author chose preposition p_i given that the correct preposition is p_j. These probabilities are used in the *artificial errors method*. The matrix also shows the preposition count for each source and label in the data set. Given the matrix and the counts, it is possible to generate the matrix for $Prob(p_j|p_i)$, the probability that the correct preposition is p_j given that the author's preposition is p_i. This second matrix is used for adapting NB with the *priors* method.

The confusion matrices show that the distribution of alternatives for each source word is very different from that of the others. For instance, using the row corresponding to the preposition *for* in Table 2, it can be observed that instead of the preposition *for*, non-native writers are 18 times more likely to use the preposition *to* than to use the preposition *by*. This strongly suggests that errors are systematic. The matrices also give a sense of the error baseline. Specifically, as the numbers show, most words are used correctly, and the error rates are very low (the error rates can be estimated by looking at the matrix diagonals in the tables). Although the error rates vary, they are typically below 10% or 5%, and in some cases less than 1%. For example, the error rate for the preposition *of* is lower than for *on*, since 97.4% of the occurrences of the label *of* are correct, but only 85.2% of the label *on* are correct.

3.2 Adaptation for a Discriminative Model

The artificial errors method is an adaptation technique for discriminative classifiers that we implement with the AP algorithm: Learner errors are simulated in training through artificial mistakes at a rate that reflects the errors made by ESL writers, which

6 Typically, the top n most frequent English prepositions are included in the preposition confusion set. We use the set used in earlier work (e.g., Rozovskaya and Roth 2014): {in of on for to at about with from by into during}.

Table 2
Confusion matrix for preposition errors used in the artificial errors method. Based on the training data from the FCE corpus. The left column shows the correct preposition. Each row shows the author's preposition choices for that label and *Prob*(*source*|*label*). The numbers next to the targets show the count of the label (or source) in the data set. The table is based on the confusion matrix for 12 prepositions but only rows corresponding to eight most frequent prepositions are shown to make the table readable.

Label	**Sources**							
	on (1,897)	for (3,995)	of (5,945)	to (3,897)	at (2,726)	in (6,367)	with (2,011)	by (625)
on (1,829)	**0.852**	0.020	0.014	0.008	0.012	0.076	0.008	0.002
for (4,034)	0.002	**0.939**	0.019	0.018	0.001	0.007	0.001	0.001
of (5,795)	0.002	0.004	**0.974**	0.001	0.001	0.009	0.002	0.001
to (3,990)	0.004	0.015	0.010	**0.936**	0.007	0.018	0.008	-
at (2,898)	0.011	0.002	0.007	0.009	**0.889**	0.076	0.002	0.001
in (6,274)	0.040	0.003	0.011	0.003	0.013	**0.924**	0.002	0.001
with (2,046)	0.002	0.008	0.009	0.009	0.002	0.015	**0.933**	0.009
by (621)	0.008	0.010	0.011	-	0.002	0.008	0.008	**0.934**

Table 3
Confusion matrix for verb agreement errors used in the artificial errors method. Based on the training data from the FCE corpus. *S* and *INF* stand for third person singular and bare verb form, respectively. The left column shows the correct verb. Each row shows the author's verb choices for that label and *Prob*(*source*|*label*). The numbers next to the targets show the count of the label (or source) in the data set.

Label	**Sources**			
	S (627)	INF (2,205)	WAS (601)	WERE (149)
S (640)	**0.961**	0.039	-	-
INF (2,192)	0.005	**0.995**	-	-
WAS (593)	-	-	**0.997**	0.003
WERE (157)	-	-	0.064	**0.936**

is ensured by generating artificial mistakes using the confusion matrix. The idea of using artificial errors goes back to Izumi et al. (2003) and Foster and Andersen (2009). The approach discussed here refers to the *adaptation* method originally proposed in Rozovskaya and Roth (2010b) and its modified version in Rozovskaya, Sammons, and Roth (2012). In Rozovskaya and Roth (2010b), artificial errors are generated using the distribution of naturally-occurring errors. This version of the approach suffers from the low recall problem, as discussed subsequently. Rozovskaya, Sammons, and Roth (2012) describe a more general method of using artificial errors for adaptation and also solve the low recall problem. This is the method we describe here.

Generating Artificial Errors with Error Statistics In the original method, the transition probabilities *Prob*(source=s|label=l) shown in Tables 1, 2, and 3 are used to generate

artificial errors in the training data. For example, from Table 2, label *on* corresponds to source *on* in 85.2% of the cases and corresponds to source *in* in 7.6% of the cases. In other words, in 7.6% of the cases when the preposition *on* should have been used, the writer instead used *in*. Thus, for each example in the training data (which are from native text and therefore assumed to be correct), with probability 7.6% we flip *on* in the training data to *in*. Note that the label of these examples is not changed, only the surface form is replaced. The native training examples now approximate the behavior of ESL data: The prepositions will now be incorrect with the same frequency and distribution as those from the corresponding target ESL user. Note that we refer to the replacement as *source*, as we refer to ESL writer word choice: Both for training (native data) and test (ESL data), *source* denotes the surface form that the classifier sees as a feature (which could be an error), and *label* denotes the correct word. For ESL data, the *source* corresponds to the word chosen by the author, whereas for native data, the *source* corresponds to the artificially produced (and possibly erroneous) surface form.

Error Sparsity and Low Recall The *artificial errors* approach just described generates errors with the frequency of naturally occurring mistakes, thereby creating a low-recall model that tends to abstain rather than flag a possible error. This happens because of error sparsity, or the low error rates in the learner data. The values of the *source* feature thus tend to have a very skewed label distribution, that is, most of the time the source feature value corresponds to the label. The system will learn that in the majority of cases the source word corresponds to the label, and will tend to over-predict it, which will result in very few mistakes being flagged.

The Error Inflation Method The two training modes—(1) training without the source word, or the knowledge about learner errors; and (2) training with the source feature, where the classifier relies on it too much—represent two extreme choices for training for correction tasks.

In the error inflation method, in order to preserve the ability of the model to take into account learner error patterns, while also increasing the model's recall, we reduce the confidence that the system has in the source word, while preserving the knowledge the model has about likely confusions (typical errors). This is achieved by decreasing the proportion of correct examples and distributing the extra probability mass among confusion pairs in an appropriate proportion by generating additional error examples. This inflates the error rate in the training data, while keeping the probability distribution among likely corrections the same. Increasing the error rate improves the recall.

Algorithm 1 shows the pseudo-code for generating artificial errors in training data; it takes as input training examples the confusion matrix *CM* (as shown in Tables 1, 2, and 3) and the inflation constant *C*, and generates artificial source features for correct training examples. An inflation constant value of 1.0 corresponds to the *original* adaptation method that simulates learner mistakes without inflation. For each training example, a probability distribution of the artificial sources is generated inside the second *for-loop*, using the confusion matrix *CM* and the inflation constant *C*. A list of tuples *ProbRanges* is created that contains for each source the probability range represented by two numbers, *start* and $start + Prob(t)$, where $Prob(t)$ is the probability of the source t, and *start* is set to 0 at the beginning of the execution of the *for-loop*, and is increased by $Prob(t)$ on every iteration. After the *for-loop* is executed, a random number x between 0 and 1 is generated, and in the next *for-loop* we pick the target whose range includes x. This is the artificial source. Table 4 shows the proportion of artificial errors created in training using the inflation method for different inflation rates for preposition errors.

Algorithm 1 Data Generation with Inflation

Input: Training examples E with correct sources, confusion matrix CM, inflation constant C
Output: Training examples E with artificial errors
for Example e in E **do**
  Initialize $lab \leftarrow e.label$, $e.source \leftarrow e.label$
  ProbRanges $\leftarrow$ []
  Initialize $start \leftarrow 0$
  for target t in $targets$ **do**
    if t equals lab **then**
      $Prob(t) = CM[lab][t] \cdot C$
    else
      $Prob(t) = \frac{1.0 - CM[lab][lab] \cdot C}{1.0 - CM[lab][lab]} \cdot CM[lab][t]$
    end if
    $ProbRanges.append((start, start + Prob(t), t))$
    $start = start + Prob(t)$
  end for
  $x \leftarrow Random[0, 1]$
  for tuple t in $ProbRanges$ **do**
    $start = tuple[0]$
    $end = tuple[1]$
    $target = tuple[2]$
    if $x > start$ and $x < end$ **then**
      $e.source \leftarrow target$
      Break
    end if
  end for
end for
return E

Table 4
Artificial errors. Percentage of preposition examples converted into artificial mistakes in training, using the inflation method with different inflation rates.

		Inflation rate			
1.0 (Original)	0.9	0.8	0.7	0.6	0.5
7.20%	16.48%	25.76%	35.04%	44.32%	53.60%

3.3 Adaptation for the Naive Bayes Algorithm

The NB adaptation method makes use of the observation that error regularities can be viewed as a distribution of priors over the correction candidates. As shown in Equation (1), in NB, a score computed for candidate p in the context S corresponds to the joint probability of p and the feature space F. This makes candidate prior a special parameter in NB. When NB is trained on native data, candidate priors correspond to the relative frequencies of the candidates in the native corpus and do not provide any

information on the real distribution of mistakes and the dependence of the correction on the word used by the author.

$$g(S,p) = log\{prior(p) \cdot \prod_{f \in F(S,p)} P(f|p)\}$$
$$= log(prior(p)) + \sum_{f \in F(S,p)} log(P(f|p)) \quad (1)$$

In the NB adaptation method, candidate priors are changed using an error confusion matrix based on learner data that specifies how likely each confusion pair is. Importantly, *adapted* candidate priors are dependent on the author's word choice. Let s be a preposition appearing in text, and p, a correction candidate. Then the *adapted* prior of p given s is:

$$prior(p,s) = \frac{C(s,p)}{C(s)}$$

where $C(s)$ denotes the number of times s appeared in the learner data, and $C(s,p)$ denotes the number of times p was the correct preposition when s was used.

Using the information from Table 2, we can compute adapted priors, that is, *Prob*(*label*|*source*), for preposition mistakes.[7] Table 5 shows adapted priors for two author's choices—*on* and *by*. First, note how the distribution of adapted priors differs from that of the global priors: According to the global priors, the most likely candidate for both of the author's prepositions is *of*; it is four times more likely than *with*. However, the *adapted* prior of *with* when the author's choice is *by* is almost ten times higher than the adapted prior of *of* (0.028 vs. 0.003), reflecting the fact that learners are more likely to incorrectly use *with* when *by* is intended than to use *of*. The second distinction of the adapted priors is the high probability assigned to the author's choice: The adapted prior for *on* given that it is also the author's choice is 0.817, vs. the 0.07 prior based on the native data. This reflects the fact that the majority of prepositions are used correctly. Finally, higher probabilities are also assigned to those candidates that are most often observed as corrections for the author's preposition (in bold in the table). For example, the adapted prior for *in* when the writer chose *on* is 0.139, because *on* is frequently incorrectly chosen instead of *in*. Similarly, the most likely confusion for the source *by* is *with*.

To determine a mechanism to inject the adapted priors into a model, note that the NB architecture directly specifies the prior probability as one of its parameters. We thus train NB in a traditional way, on native data, and then replace the prior component in Equation (1) with the adapted prior to get the score for p of the *NB-adapted* model:

$$g(S,p) = log\{prior(p,s) \cdot \prod_{f \in F(S,p)} P(f|p)\}$$

It should be stressed that in the NB adaptation method there is no need to re-train the model to adapt to a different set of learner error patterns, as is the case with the

7 Adapted priors for article and verb agreement errors can be generated in a similar way using Tables 1 and 3.

Table 5
Examples of *adapted* candidate priors used to adapt NB (Prob(candidate label|source (author's choice))) for two of the author's choices—*on* and *by*. *Global prior* denotes the probability of the candidate in the standard model and is based on the relative frequency of the candidate in native training data. *Adapted priors* are dependent on the author's preposition. Adapted priors for the author's choice are very high. Other candidates are given higher priors if they often appear as corrections for the author's choice (shown in **bold**). Based on preposition errors in the FCE training data.

Candidate label	Global prior	Adapted priors	
		prior for when the author's choice is *on*	prior for when the author's choice is *by*
of	0.25	0.005	0.003
to	0.22	0.007	0.001
in	0.15	**0.139**	0.013
for	0.10	0.005	0.005
on	0.07	**0.817**	0.006
by	0.06	0.003	**0.928**
with	0.06	0.002	**0.028**
at	0.04	0.017	0.006

artificial errors approach. Only one model is trained, and only at decision time do we change the prior probabilities of the model. Furthermore, even though NB does not typically perform as well as a discriminative classifier when both models are trained on the same data and features (Rozovskaya and Roth 2011, 2014), the algorithm also has advantages in several training scenarios (see Section 7).

4. Key Adaptation Experiments

We now present experiments demonstrating that the adaptation approaches with minimal supervision proposed in this work outperform the standard methods of training on native data and when training on annotated learner data alone. In the following sections, we describe the data sets, and present key adaptation experiments on three error types. Section 5 presents language-specific adaptation.

4.1 Data Sets

We use native English data and an annotated corpus of learner essays. The native English data is used for training, and the ESL data is used both for training and evaluation.

Native English Data Sets Two corpora of native English are used. The first corpus, WikiNYT, is a selection of texts from English Wikipedia and the *New York Times* section of the Gigaword corpus (Parker et al. 2009). Unless otherwise stated, the training sizes for native data are fixed, as follows: 1.8M examples for agreement errors, 3M examples for articles, and 5M for prepositions (the training sizes are chosen based on the size of the confusion sets for each error type). Note that the entire WikiNYT corpus contains more training examples; these sizes are those used in the key experiments.

Table 6
Statistics on the native training data sets used in the key experiments. For the WikiNYT corpus, training sizes indicate the number of relevant training examples (articles/prepositions/verbs) in millions. (The entire WikiNYT corpus contains more examples; the table shows the sizes used in the key experiments.) Training sizes for the Web1T corpus are approximate and are based on the corpus size (10^{12} words).

Data set	Training sizes		
	Articles	**Prepositions**	**Verb agreement**
WikiNYT	3M	5M	1.8M
Web1T	40,000M	20,000M	25,000M

The second corpus is Google Web1T 5-gram (Brants and Franz 2006) (henceforth Web1T), which is a collection of *n*-gram counts of lengths one to five over a corpus of 10^{12} words. Table 6 shows the sizes of the two native data sets. Because Web1T does not come with complete sentences, we approximate the number of training examples for each error type for Web1T based on the sizes of Web1T and WikiNYT and the frequencies of the respective target words in WikiNYT. To compare with the WikiNYT, Web1T contains on the order of 10,000 more training examples for each error type.

Using the two corpora allows us to evaluate the proposed adaptation methods when applying two state-of-the-art machine learning algorithms and to demonstrate how to take advantage of the benefits provided by each data source and each machine learning framework. On WikiNYT, we train discriminatively using the AP algorithm and rich syntactic features shown to be useful for article and verb agreement errors (Lee and Seneff 2008; Rozovskaya and Roth 2014) and for preposition errors (Tetreault, Foster, and Chodorow 2010). Because of the special format of the Web1T corpus, it is difficult to generate rich feature annotations for this data, and to make use of a discriminative classifier on this corpus, as one would have to limit the surrounding context to two words on each side of the mistake. Because we wish to make use of the context features that extend beyond the two-word window, it is only possible to use count-based methods (e.g., NB or language models). We thus train the NB algorithm.

Learner Data Several annotated learner data sets were made available recently, including the data set used in the HOO competition, the CoNLL data set (Dahlmeier, Ng, and Wu 2013), and the FCE data set (Yannakoudakis, Briscoe, and Medlock 2011), a subset of the Cambridge Learner corpus. Because we want to explore the effects of first language backgrounds, we use the FCE corpus in this work. This corpus contains data from learners of multiple language backgrounds, including information on the first language of the writer. The work described in Section 5 makes use of this information. We discuss other corpora and approaches in Section 6.

The FCE corpus contains 1,244 essays (500,000 words) produced by learners of 16 first language backgrounds. The data set is fully corrected and error tagged. For the key adaptation experiments, we split the data into training and test (according to the split used in the HOO-2012 shared task [Dale, Anisimoff, and Narroway 2012]), selecting 1,000 files of the FCE corpus as training data. We use the remaining 244 documents from the FCE data for testing. Table 7 shows the number of training and test examples, as well as the number of incorrectly used instances and the error rates (the percentage of mistakes with respect to the total number of examples). The error rates are below

Table 7
Statistics on articles, prepositions, and verbs in the ESL data set. *Error rate* denotes the percentage of examples (articles, prepositions, or verbs in the data) that are mistakes.

Data set	Total examples			Errors (error rate)		
	Arts.	Preps.	Verb agr.	Arts.	Preps.	Verb agr.
Train	60,743	32,699	35,322	2,892 (4.76%)	2,936 (8.98%)	490 (1.39%)
Test	15,398	7,927	8,695	787 (5.11%)	750 (9.46%)	137 (1.58%)

10% but vary by error type (from slightly more than 1% for verb agreement, to 5% for articles, to almost 9% for prepositions).

In the key experiments, we generate combined statistics across writers of all 16 first language backgrounds. For first-language and other fine-grained adaptations, see Section 5.

4.2 Experimental Set-up

Depending on what training data source is used, we refer to the models as follows:

1. **Native-trained models**: trained on native English data in the selection paradigm
2. **ESL-trained models**: trained on annotated learner data in the correction training paradigm
3. **Adapted models**: trained on native English data and adapted using annotated learner data

Features AP models trained on WikiNYT use rich features tailored to each error type. These features were used in the components of the Illinois system in several shared tasks (Rozovskaya et al. 2011, 2013) and are presented in Appendix Tables A.1, A.2, A.3, and A.4 for convenience. The Web1T models are trained on word *n*-gram features.

Parameter Tuning Ten percent of the training data is used to optimize the inflation rate for the AP-adapted models (0.8 for articles and prepositions, and 0.85 for verb agreement).

4.3 Note on Evaluation Metrics

Various metrics have been proposed and used in error correction. These can be broken down roughly into metrics that compute the accuracy of the system and those that use the F-measure. The HOO competitions adopted F1, which can take into account both precision and recall of the systems; the M2 scorer used in CoNLL is also F-based but can take into account phrase-based edits (Dahlmeier and Ng 2012). Overall, accuracy and F-measure are equivalent, with the exception of tuning; F-measure is flexible in that it allows for calibrating precision and recall. In CoNLL-2014, F0.5 was used (precision was weighed twice as high as recall).

Three papers have been published recently that addressed the appropriateness of commonly used metrics for error correction; two of these also proposed new metrics that are different from the standard F1 adopted in the shared tasks but are variations of accuracy or F-measure. Felice and Briscoe (2015) proposed an I-measure that is accuracy-based and Napoles et al. (2015) proposed a variation of the BLEU metric used in Machine Translation (called GLEU). Further, Napoles et al. (2015) and Grundkiewicz, Junczys-Dowmunt, and Gillian (2015) also compare the outputs of the systems in the CoNLL shared task against human judgments and show that the need for new metrics is motivated by a lack of correlation between F1 and human judgments. However, the newly proposed GLEU metric does not fare much better in terms of correlation with human judgments than the F-measure and the I-measure.

Developing a new, more appropriate metric is an involved issue that is beyond the scope of this work. We are not dealing with phrase-based edits here, so we follow the evaluation metrics based on precision, recall, and F1 that were used in the shared tasks on grammar correction.

4.4 Key Adaptation Results

The key question addressed in this section is the following:

- Does adding a little bit of learner data to large amounts of native training data result in better models compared with training on native data alone?

We compare native-trained models and adapted models implemented within the two algorithms—AP and NB, trained on two native corpora—and evaluate on three types of mistakes. Adapted models use a small amount of learner data (about 6K, 3K, and 3.5K article, preposition, and verb agreement examples, respectively) extracted from an annotated FCE sample of 40K words, which constitutes 10% of the FCE training data. Additionally, we show performance for models adapted using all of the FCE training data.

Table 8 presents the results. First, we compare native-trained models and adapted models that use 10% of learner training data. For AP adapted models, we show two variants in each case: one with inflation rate 1.0 (this is the baseline adapted model that uses the error rates that correspond to naturally occurring mistakes) and one that uses an inflation rate that was optimized on the development sample. Adaptation consistently improves the performance for both algorithms and all three types of mistakes: The improvements range between 1.5 F1 points for articles to 2 F1 points for agreement and preposition mistakes. It should be stressed that we are using a small annotated sample. In the following description, we also show that training an entire model on this learner sample results in poor performance as context parameters cannot be estimated robustly.

Interestingly, when we adapt using all of the FCE training data (also shown in the table), we only observe improvements for verb agreement errors and a small boost for preposition error correction with AP models. Clearly, article error patterns can be reliably estimated using just a small ESL sample. This is because these errors are not as sparse as verb agreement errors (see Table 7) and have a smaller confusion set than preposition errors. In contrast, verb agreement errors are more sparse, thus adding more data for adaptation is beneficial.

Finally, note that NB models perform better than AP models on article and preposition error correction, even though prior work showed that AP outperforms NB

Table 8
Key adaptation experiments. Models are trained on *WikiNYT* data or *Web1T*. Two types of adapted models are shown: those that use 100 files (10%) of the *FCE* training data and those that use all of the learner training data for error statistics. Inflation rate is optimized on the development set. All differences are statistically significant (McNemar's test, $p < 0.0001$).

Model	Training size		Infl. rate	Performance		
	Native	ESL		P	R	F1
			Articles			
Native-trained (AP, WikiNYT)	3M	-	-	26.72	36.80	30.96
Adapted (AP, WikiNYT)	3M	6K	0.80	28.73	41.60	**33.99**
Adapted (AP, WikiNYT)	3M	6K	1.0	43.91	15.66	23.08
Adapted (AP, WikiNYT)	3M	60K	0.85	30.84	36.23	33.32
Adapted (AP, WikiNYT)	3M	60K	1.0	43.20	14.51	21.73
Native-trained (NB, Web1T)	40,000M	-	-	28.13	41.83	33.64
Adapted (NB, Web1T)	40,000M	6K	-	31.74	39.43	**35.17**
Adapted (NB, Web1T)	40,000M	60K	-	31.87	38.97	35.06
			Prepositions			
Native-trained (AP, WikiNYT)	5M	-	-	26.07	22.80	24.32
Adapted (AP, WikiNYT)	5M	3K	0.80	31.27	22.27	26.01
Adapted (AP, WikiNYT)	5M	3K	1.00	56.12	7.33	12.27
Adapted (AP, WikiNYT)	5M	32K	0.80	31.73	22.93	**26.62**
Adapted (AP, WikiNYT)	5M	32K	1.00	45.32	8.40	14.17
Native-trained (NB, Web1T)	20,000M	-	-	28.47	27.87	28.16
Adapted (NB, Web1T)	20,000M	3K	-	31.42	30.00	**30.69**
Adapted (NB, Web1T)	20,000M	32K	-	31.33	29.20	30.23
			Verb agreement			
Native-trained (AP, WikiNYT)	1.8M	-	-	36.43	34.31	35.34
Adapted (AP, WikiNYT)	1.8M	3.5K	0.85	33.00	72.99	**45.45**
Adapted (AP, WikiNYT)	1.8M	3.5K	1.00	48.28	30.66	37.50
Adapted (AP, WikiNYT)	1.8M	35K	0.90	36.72	68.61	**47.84**
Adapted (AP, WikiNYT)	1.8M	35K	1.00	54.88	32.85	41.10
Native-trained (NB, Web1T)	25,000M	-	-	46.67	30.66	37.00
Adapted (NB, Web1T)	25,000M	3.5K	-	43.36	35.77	**39.20**
Adapted (NB, Web1T)	25,000M	35K	-	42.06	38.69	**40.30**

(Rozovskaya and Roth 2011, 2014), when both are trained under the same conditions. This is because NB and AP results are not directly comparable here, since even though NB is trained using a less rich set of features (word *n*-grams only), it uses much more data.

4.5 A Graphical View of ESL Adaptation

We now summarize the adaptation idea graphically using the preposition error correction task and the AP learning framework as an example. As shown in Figure 3, we can view adaptation by considering a graph with an *x*-axis representing the amount of ESL data and the *y*-axis representing the amount of native data used in training. Each box corresponds to a model, and the numbers in boxes show F1 results. The boxes on the *x*-axis indicate models that are trained exclusively on ESL data, and the boxes on the *y*-axis represent native-trained models. The boxes in the middle are *adapted* models that

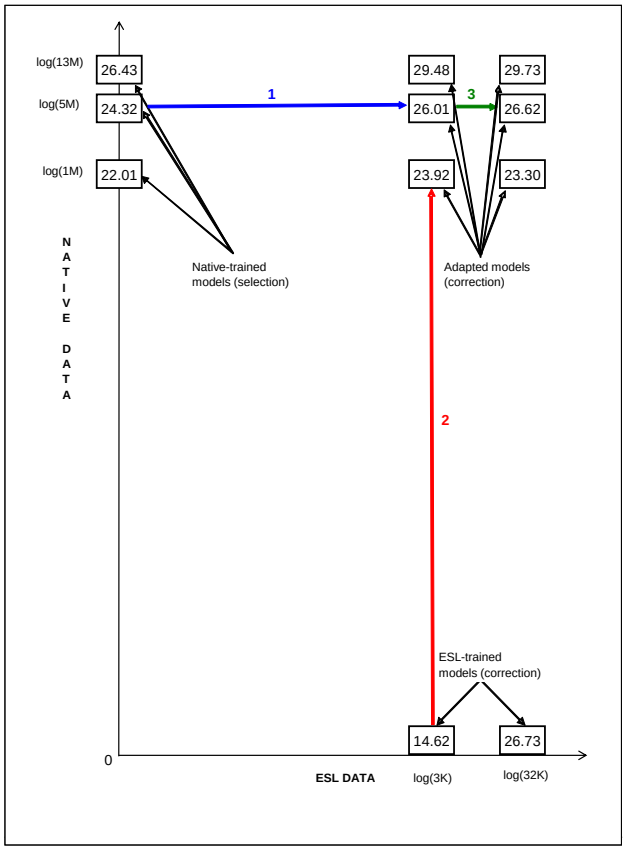

Figure 3
Adaptation: a two-dimensional summary, using *the preposition error correction task and the AP learning framework* as an example. Increasing the amount of training data is important, but one type of data is much more costly than the other. Adaptation methods are shown to gain from a small increase in the costly dimension. The *x*-axis shows the ESL data continuum as a percentage of the annotated training examples. The *y*-axis shows the native training data. Note how the native data size is in fact several orders of magnitude larger than the ESL data. The numbers in boxes show F1 results.

are trained on much native data and adapted using different amounts of ESL data. The *x*-axis is expensive, and the *y*-axis is cheap. The data sizes are shown for each axis in *log* scale, because the sizes of the native data are several orders of magnitude larger than the sizes of the ESL training data: The entire ESL data set contains 32,000 preposition examples, and the 3,000 also shown on the graph corresponds to 10% of the data (the amount used for adapting the models in Table 8). The native data size varies from 1 million to 13 million prepositions. To have a fair comparison, in all cases, we train AP classifiers using the same set of rich features. The same results are also presented in tabular form (Table 9).

First, consider models trained on learner data: A model trained on 3,000 non-native examples obtains an F1 score of 14.62. We can improve to 26.73 if we use 10 times more ESL training data (32,000 examples). Even though annotating so much data is a big effort, this is still not much data to train a robust model. On the other hand, when training on native data, increasing the training size is easy. A native-trained model that

Table 9
Adaptation using the preposition error correction task as an example. Models are trained on different sizes of native (*WikiNYT*) and learner data suing the AP algorithm. Each cell shows the F1 score. The same results are presented in Figure 3.

Amount of native training data	Amount of learner training data		
	0	3K	32K
0	-	14.62	26.73
1M	22.01	23.92	23.30
5M	24.32	26.01	26.62
13M	26.43	29.48	29.73

uses 1 million examples obtains an F1 score of 22.01, and we can improve further, to 24.32 and 26.43, if we use 5 million and 13 million training examples, respectively. We continue to improve as we add more data (28.16) when training on Web1T using the NB algorithm (Table 8). This demonstrates the importance of using native data for learning parameters relating to context features.

The results in the graph can be summarized as follows. The main lesson is that the two types of training data are complementary, although *increasing the amount of training data is important, but one type of data is a lot more costly than the other*. The adaptation methods are shown to gain very much from a small increase in the costly dimension: The key result (marked as line "1" on the graph) is that a model trained on 5 million native prepositions and *adapted* with 3,000 ESL prepositions outperforms a native-trained model that uses the same number of training examples from native data (26.01 versus 24.32 F1 points). Similarly, there is a 3-point improvement due to adaptation when the models use 13 million prepositions from the native data. These results demonstrate that even a little bit of ESL data can make a big difference.

The comparison marked as line "2" demonstrates the importance of using native data: an ESL-trained model that uses 3,000 prepositions performs very poorly, because there is not enough evidence to learn all parameters reliably. We can improve considerably to 23.92 and 29.48 by adding 1 million and 13 million native examples, respectively.

The result marked as line "3" shows that the model adapted with just 3,000 prepositions is very close to the performance of the model adapted with 10 times more data. This shows that by using a small amount of annotated ESL data, it is possible to estimate the mistake parameters reliably.

Finally, we continue to improve as more native data are used (we improve by 2 F1 points (from 24.32 to 26.43) for native-trained models when using 13 million prepositions instead of 5 million). Importantly, adding learner data always helps, regardless of how much native data is used, because *learner data contributes knowledge on the types of mistakes that are not available in the native data*: Adding a little bit of learner data for adapted models, we improve by more than 3 F1 points (from 26.01 to 29.48) compared with native-trained models that use the same amount of native data.

4.6 Prepositions: Top Three Evaluation

One thing that can be observed is that the performance numbers are quite low, especially for preposition mistakes. One reason that is related to this is the *learner* baseline, or the performance of non-native writers, which is quite high to begin with. Typically,

Table 10
Prepositions: Top three evaluation. Models are trained on *WikiNYT* data or *Web1T*. Adaptation uses 100 files (10%) of the *FCE* training data.

Model	Training size		Infl. rate	Performance		
	Native	ESL		P	R	F1
	Top 1					
Native-trained (AP, WikiNYT)	5M	-	-	26.07	22.80	24.32
Adapted (AP, WikiNYT)	5M	3K	0.8	31.27	22.27	**26.01**
Native-trained (NB, Web1T)	20,000M	-	-	28.47	27.87	28.16
Adapted (NB, Web1T)	20,000M	3K	-	31.42	30.00	**30.69**
	Top 3					
Native-trained (AP, WikiNYT)	5M	-	-	31.84	41.60	36.07
Adapted (AP, WikiNYT)	5M	3K	0.5	72.70	38.00	**49.91**
Native-trained (NB, Web1T)	20,000M	-	-	44.13	43.07	43.59
Adapted (NB, Web1T)	20,000M	3K	-	59.87	38.00	**46.49**

fewer than 10% of instances are erroneous (one would expect that a higher baseline [i.e., a lower error rate] would imply a more difficult learning task). However, although the preposition baseline is lower than for the other mistakes, the performance is also the lowest, which seems counterintuitive. We attribute this to the large confusion set (12 prepositions), and to the fact that, typically, preposition usage is highly ambiguous (multiple prepositions can be licensed in a given context [Tetreault and Chodorow 2008]), whereas the FCE annotation does not allow for multiple acceptable answers. De Felice (2008) investigated preposition performance when going beyond the top match and found that performance would improve when looking at the top *n* candidate corrections.

Given that preposition usage is highly variable, we conduct further evaluation using the top three choices provided by the classifiers (Table 10). In the top three evaluation, a classifier's prediction is counted as "correct" if the gold label is among any of the top three candidates selected by the classifier. F1 performance increases for all the models when top three choices are considered. However, adapted models fare much better: Precision increases substantially from 31.27 to **72.70** for AP-adapted and from 31.42 to **59.87** for NB-adapted models, whereas it only improves slightly from 26.07 to 31.84, and from 28.47 to 44.13 for their respective native-trained counterparts. These improvements suggest that it is not just the first candidate that is better but the overall ranking is better in adapted models. The top three evaluation thus provides additional evidence that *knowledge of error patterns available in the adapted models is extremely important*.

Whether presenting the top three candidates to the user is reasonable depends on the evaluation of how good the top three candidates are. Thus, we also wish to assess top three candidates directly because the top three F1 measure values presented are computed for a gold standard that allows only a single correct answer, when in fact there may be multiple possible correct answers. More specifically, given that preposition usage is variable, we would like to know how often the top three candidates include *multiple* valid options. To this end, for each adapted model, we selected a sample of

Table 11
Prepositions: Top three evaluation. Column 1 shows the number of correct predictions among the top three candidates (between 0 and 3). Each cell shows the percentage of cases with this number of correct predictions among the top three.

Number of correct predictions among the top 3 candidates	Percentage of preposition examples	
	(AP-adapted, %)	(NB-adapted, %)
0	6.3	14.3
1	79.4	71.4
2	14.3	12.5
3	0	1.8

200 preposition instances where the classifier's top choice was different from the gold label. These were manually annotated by one of the authors, a native English speaker, who was given a preposition in a context (where by "context" we consider the entire sentence where the preposition appears) and was asked to specify *how many* of the top three choices are acceptable in a given context. Results of the evaluation are shown in Table 11. Column 1 shows the number of correct predictions among the top three candidates (between 0 and 3). Each cell shows the percentage of cases with this number of correct predictions among the top three. We found that for the majority of the cases, at least one of the top three candidates is valid; row 1 shows that only 6.3% and 14.3% of the examples for AP- and NB-adapted models, respectively, did not have a single correct candidate among the top three choices. More importantly, 14.3% of examples for each of the classifiers have two or three valid preposition choices among the top three candidates (last two rows in the table). We believe that these numbers confirm earlier findings with respect to the highly variable preposition usage and thus justify presenting multiple options to the user, at least for preposition mistakes. Finally, it is interesting to note that AP-adapted has a higher percentage of correct prepositions in the top three evaluation than NB-adapted, even though NB-adapted does better in the top 1 evaluation. We believe that this indicates that AP-adaptation is better than NB-adaptation at ranking the appropriate candidates. This result is also consistent with those in Table 10 that show that the precision (and F-score) of AP-adapted models in the top three evaluation improves from 31.27 to 72.70, whereas for NB-adapted models the improvement is more modest (from 31.42 to 59.87). Lastly, as the results show, whereas precision numbers increase markedly, the recall does not. This happens because the top three evaluation is performed on models that were optimized with respect to top one evaluation. It is also possible to optimize models based on top three evaluation, in which case both precision and recall would increase, resulting (most likely) in improved F1 scores. However, the evaluation and additional annotation that we perform already demonstrate how multiple preposition usage underestimates top one evaluation.

5. Adaptation by Language

In the previous experiments, models were adapted based on error patterns of learners from a variety of first-language backgrounds. In this section, we investigate the question of using error patterns from the writers of the *same* first language vs. *other* (linguistically related or unrelated) language backgrounds. In this article, we consider

Table 12
Number of examples and errors for the 11 languages used in the experiments on language-specific adaptation.

First language	Error type: Articles			Prepositions			Verb agreement		
	Total	Errors	Correct (%)	Total	Errors	Correct (%)	Total	Errors	Correct (%)
Chinese	5,054	273	94.60	2,337	224	90.42	2,615	54	97.93
French	9,296	328	96.47	4,419	390	91.17	4,807	92	98.09
Greek	4,962	212	95.73	2,296	244	89.37	2,584	34	98.68
Italian	4,848	178	96.33	2,377	181	92.39	2,358	17	99.28
Japanese	5,073	331	93.48	2,387	177	92.58	2,464	43	98.25
Korean	6,006	374	93.77	2,708	215	92.06	2,932	29	99.01
Polish	5,335	397	92.56	2,507	174	93.06	2,802	45	98.39
Russian	5,703	440	92.28	2,793	229	91.80	2,769	22	99.21
Spanish	12,317	339	97.25	5,866	588	89.98	6,444	80	98.76
Thai	4,371	277	93.66	2,107	170	91.93	2,316	58	97.50
Turkish	5,275	355	93.27	2,459	213	91.34	2,802	24	99.14
Total	68,240	3,504	94.97	32,256	2,805	91.30	34,893	498	98.57

two groups of closely related languages: Russian and Polish (Slavic), and Spanish and Italian (Romance).[8] As discussed in Section 2, research in second-language acquisition as well as empirical studies in natural language processing indicate that the effect of the native language on learner error patterns is strong. In fact, because writers of related languages use similar structures transferred from their native language to English, it is even possible to recover language family structures by analyzing writing patterns of non-native speakers (Nagata and Whittaker 2013; Berzak and Katz 2015).

Here, we study to what extent having language-specific patterns affects the quality of adaptation. We use three phenomena—articles, prepositions, and verb agreement—as these occur in the FCE data from 11 first languages (shown in Table 12), excluding languages that only have a handful of examples, and study these phenomena within the NB adaptation framework, training on the native data from the Web1T corpus and adapting using learner data from different first-language groups. Specifically, we denote by *target*, *other*, *related*, and *unrelated*, error patterns obtained from the learners of the same first-language background, all other first languages (excluding the target), languages related to the target, and those unrelated to the target, respectively. The group *other* includes both related and unrelated languages to the target.

Because experiments in this section are performed individually by first-language group, we do not separate the data into training and testing partitions. Instead, results are reported in 10-fold cross-validation, where each time we evaluate on 10% of target-language data, and 90% of the target data is used to estimate prior parameters of the target-based adaptation model. Similarly, in all other settings—related, unrelated, and other—we select the same amount of data to estimate priors from other, related, or unrelated languages.

8 Section 5.2 details the definitions of related languages adapted for our experiments.

Table 13
Adapting NB with data from the target language background and other language backgrounds. Numbers in the parentheses show the relative improvement of using adapted models vs. the native-trained models. All improvements of adapted models vs. native-trained models are statistically significant (McNemar's test, $p < 0.001$).

Error	Native-trained			Adapted (other)			Adapted (target)		
	P	R	F1	P	R	F1 (rel. improv.)	P	R	F1 (rel. improv.)
Article	28.06	41.49	33.47	31.17	40.65	35.28 (5.4%)	32.21	41.33	**36.20** (8.2%)
Prep.	18.41	24.97	21.19	23.25	23.82	23.63 (11.5%)	23.80	24.95	**24.25** (14.4%)
Verb agr.	28.52	46.51	35.32	32.73	43.78	37.44 (6.0%)	33.58	45.96	**38.80** (9.9%)

The following key questions are addressed.

1. **Target priors vs. other priors.** Suppose we have annotated data from a set of writers of various linguistic backgrounds, and we need to adapt to writers from a different linguistic background that is not part of our corpus. Is there a benefit in collecting error patterns on this new set of learners for adaptation? In other words, can we do better using error patterns from writers of the target language than when adapting using data from other speakers?
2. **Target priors vs. related language priors vs. unrelated language priors.** Here, we investigate in more detail the effect of language relatedness on the quality of adaptation. For a given target language, among all languages that are different from the target, we distinguish between those that are linguistically related to the target and unrelated languages. We study whether there is a difference between adapting using target priors, priors based on languages that are closely related linguistically to the target, and priors from languages unrelated to the target.

5.1 Target Priors vs. Other Priors

Table 13 compares three types of models: Native-trained, adapted with other priors, and adapted with target priors. In all adapted models, priors are estimated on the same amount of data: from the same language (*target*); or from a set of ten other languages (*other*). We note that adapting using both target and other language data is effective compared with training on native data only. However, for all error types, there is an advantage to using target language priors over priors estimated from data written by speakers of other first language backgrounds: The relative improvement of using adapted models vs. native-trained models ranges between 5.4% to 11.5% for other-language adaptation, but ranges between 8.2% to 14.4% for target-language adaptation.

We now consider whether all languages benefit in the same way when target-language adaptation is used in place of other-language adaptation. Table 14 shows that, with a few exceptions, both target-language and other-language adaptations help for all language groups and error phenomena. With respect to the difference in the effect of target-language adaptation and other-language adaptation, whereas many language groups and phenomena benefit from target language adaptation vs. other-language

Table 14
Target-language adaptation. Adapting NB with data from the *target* language background and *other* language backgrounds. Results are presented by first-language background where languages are ordered by the amount of target error data available. **Bold** is used to indicate results where target-language adaptation is better than other-language adaptation.

Language (number of errors)	F1			
	Native-trained	Adapted (other langs.)	Adapted (target lang.)	Adapted (other+target)
		Articles		
Russian (440)	37.72	39.59	**40.14**	39.47
Polish (397)	40.93	43.33	**43.43**	43.22
Korean (374)	38.96	40.97	**41.37**	**41.34**
Turkish (355)	39.12	40.94	40.67	40.71
Spanish (339)	23.58	24.52	**25.31**	**24.82**
Japanese (331)	39.82	42.82	42.80	42.74
French(328)	26.86	27.18	**27.70**	**27.56**
Thai (277)	34.92	35.80	35.79	35.79
Chinese (273)	30.34	32.51	32.27	32.13
Greek (212)	29.89	31.84	**33.53**	**32.51**
Italian (178)	28.56	30.42	**31.13**	**30.86**
		Prepositions		
Spanish (588)	23.88	26.44	**28.07**	**27.65**
French (390)	19.85	20.95	**23.35**	**21.08**
Greek (244)	27.63	30.27	**32.03**	**31.43**
Russian (229)	17.16	20.81	**22.92**	20.85
Chinese (224)	23.11	25.36	24.77	24.88
Korean (215)	19.17	22.83	20.70	**23.17**
Turkish (213)	18.41	23.33	21.20	22.94
Italian (181)	22.97	22.50	**26.26**	**23.64**
Japanese (177)	16.13	17.99	16.66	**18.32**
Polish (174)	24.25	25.86	24.42	24.52
Thai (170)	18.12	18.87	17.23	**19.64**
		Verb agreement		
French (92)	38.99	38.79	**42.33**	**39.97**
Spanish (80)	36.59	39.74	**40.56**	39.46
Thai (58)	27.84	27.79	**28.08**	**29.99**
Chinese (54)	43.39	44.71	**47.78**	43.71
Polish (45)	49.00	52.51	**54.71**	**53.18**
Japanese (43)	34.23	33.35	**38.48**	31.92
Greek (34)	28.06	31.54	30.35	30.14
Korean (29)	33.51	36.57	32.63	**37.12**
Turkish (24)	36.90	41.00	39.29	40.93
Russian (22)	27.03	31.54	30.65	29.70
Italian (17)	21.34	25.87	24.41	**26.30**

adaptation (these results are in bold in the table), some language groups do better with adaptation based on other languages than on their own target data.

It is interesting to observe that preposition and verb agreement errors behave similarly, in that languages that do not benefit from target priors as much as from priors extracted from other languages are those that have less error evidence for prior estimation: Specifically, for verb agreement and preposition errors there is a strong correlation between the amount of target error evidence and the ability to improve over priors estimated from other languages. We conjecture that this happens because when more error data are available, differences related to language-specific mistake patterns

become more pronounced, and thus there is more benefit when using target-language data for adaptation vs. other-language data. It should be noted, however, that data size for prior estimation depends on error type, or, specifically, the error rate and the size of the confusion set (small for agreement, large for prepositions). Thus, for agreement mistakes, even 40 errors provide a good estimate of error patterns. For preposition mistakes, it is a good idea to have over 200 errors to obtain reasonable estimates. To collect preposition data, this would require about 2K words of annotated learner texts (we assume an average error rate of 10% for preposition mistakes and that prepositions typically account for 10% of words). For verb agreement errors, error rates are often less than 2%, therefore the amount of annotation required might be even higher, despite a smaller confusion set.

Another interesting observation is that, unlike for preposition and verb agreement mistakes, for articles there is no correlation between the size of error data used in adaptation and the benefit obtained from using target-language adaptation. In fact, observe that we have both more data to estimate article error patterns (Table 12) and a smaller confusion set compared with prepositions, which suggests that estimated language-specific parameters are quite robust. We conjecture that the reason for a smaller benefit from target priors for article mistakes for some of the language groups is that patterns of article errors are more similar across languages than patterns of other error types. Table 15 illustrates article usage patterns in the languages for which we use learner English data. For instance, even though Russian and Korean are not closely related linguistically, because neither of these languages has articles, the error patterns can be expected to be similar.

Finally, the last column shows performance of models that are adapted both with target and other language data, where 20% of the data for adaptation come from the target language. These results help answer the question "if we already have other-language data, would it be helpful to also acquire target-language data?" To this end, we compare adaptation using "other" language data only with adaptation that uses other+target data. For article errors, there is no clear benefit to combining target- and other-language data. This is consistent with the adaptation results on articles that do not show an advantage to using target- priors over other-language priors. However, some languages do exhibit a consistent small improvement when target-language data is added to other-language data for adaptation. Specifically, these are the same languages that also exhibit an improvement when target-language data is used *instead of* other-language data. In general, languages that benefit more from target-language priors are also the same languages that benefit when target priors are added to other priors (see the last two columns). For Greek, for example, adding target-language priors to other-language priors is particularly beneficial.

Similarly, for prepositions, typically adding target-language prior data helps in those cases when target-language adaptation is beneficial compared with

Table 15
Article usage across different first languages.

Language	Def. article	Indef. article
French/Greek/Italian/Spanish	+	+
Japanese/Korean/Polish/Russian/Thai	−	−
Turkish	−	+

other-language adaptation. For some languages, however (e.g., Thai and Japanese), other+target adaptation outperforms both the target-language and the other-language adaptations. We conjecture that this happens when target data are scarce and mixing that with other-language data introduces a good balance of common and language-specific error patterns. Finally, for verb agreement, it is not clear when adding target-language data to other data is beneficial.

Similarity of Error Patterns and Its Effect on Adaptation This observation about articles suggests that we should expect target-language adaptation to be particularly useful compared with other-language adaptation, when the error patterns based on the other languages are dissimilar to the target language error patterns. To determine whether this is indeed the case, for two phenomena (articles and verb agreement) we selected two languages that behave differently when it comes to target-language adaptation. For article mistakes, we chose Russian and Greek: Greek is a language that benefits substantially from target priors compared with other-language priors, which is not the case for Russian. Similarly, for verb agreement mistakes, we chose Japanese and Thai.

Table 16 illustrates for each error type and language two error distributions: (1) the distribution of various confusions based on the target data and (2) the distribution of confusions based on data from other languages. For example, the confusion of type *∅-the* (the definite article is incorrectly omitted) accounts for 27.5% of all article errors in the data by Greek writers, and accounts for 40.24% of all mistakes in the Russian data. Based on the evidence from other 10 language backgrounds, this confusion accounts for about 37% of all article mistakes. The relative frequency of this confusion in the Russian data is thus closer to the "averaged" frequency than for the Greek data. Overall, for the majority of article errors in the Russian data (over 66%) the distribution is very similar to "averaged" error statistics, which is not the case for Greek. More generally, we hypothesize that the reason Greek benefits most from moving from "other" priors to "target" priors is because the distribution of article errors for Greek writers is very different from the "averaged" distribution. In contrast, for Russian, both of the distributions are quite close.

If we assume that we have a measure for the distributional distance between "target" and "other" error patterns, then we wish to determine whether there is indeed a correlation between the distributional distance and the benefit from moving from "other" priors to "target" priors. To this end, for a given language background and error type, we compute weighted L^2 distance between the "target" error distribution and "other" error distribution, where the coefficients (weights) correspond to the frequency of the occurrence of the specific confusion type. More formally, let t denote the error distribution of the "target," and o denote the error distribution of the "other" (these are the distributions shown in Table 16). Then t_i and o_i refer to the frequencies of the specific confusion in "target" and "other," respectively (as shown in Table 16). For instance, the frequency of the confusion type ∅-the in the Greek target data is 27.50%. We compute weighted L^2_w distance as follows:

$$L^2_w(t, o) = \sqrt{\sum_{i=1}^{n} (t_i - o_i)^2 \cdot t_i}$$

Table 17 demonstrates that, indeed, larger L^2_w distances correspond to bigger advantages when using "target" priors in place of "other" priors.

Table 16
Comparison of distributions of article and verb agreement confusions. For each language, confusions based on the target data and data from other 10 languages are shown.

Distributions of article confusion sets

Confusions	Greek		Russian	
	Target (%)	Other (%)	Target (%)	Other (%)
∅-the	27.50	37.35	40.24	37.09
∅-a	30.50	25.11	26.43	24.51
the-∅	30.50	19.94	11.43	22.11
a-∅	7.50	8.07	9.05	7.97
a-the	1.00	5.21	7.14	4.51
the-a	3.00	4.32	5.71	3.81
	100.00	100.00	100.00	100.00

Distributions of verb agreement confusion sets

Confusions	Japanese		Thai	
	Target (%)	Other (%)	Target (%)	Other (%)
INF-S	32.57	51.42	53.45	49.32
S-INF	39.53	29.45	25.86	30.91
WAS-WERE	18.60	14.51	8.62	15.68
WERE-WAS	9.30	4.40	10.34	4.09
INF-WAS	-	0.22	1.73	-
	100.00	100.00	100.00	100.00

Table 17
Distance between error distributions vs. benefit from using target priors compared to priors averaged on other language data. *Relative advantage* reflects the relative improvement that we get between priors estimated on target data vs. other data.

Articles

Language	Weighted L^2 dist.	Target vs. other priors: rel. advantage
Greek	83.55	5.31%
Russian	43.37	1.38%

Verb agreement

Language	Weighted L^2 dist.	Target vs. other priors: rel. advantage
Japanese	126.98	15.38%
Thai	49.09	1.04%

5.2 Adaptation by Target, Related, and Unrelated Languages

We now address the second question, that of comparing the contribution of priors estimated from the target language, from related languages, and from languages that are not related to the target. Two pairs of closely related languages among the 11 are

identified: Russian/Polish and Spanish/Italian. For each of the four languages, we compare four models:

1. *Native-trained* (unadapted)
2. *Adapted with target priors*
3. *Adapted with data from related languages*: For Russian, we used Polish data; for Polish, we used Russian; for Italian, we used Spanish; and for Spanish, we used Italian.
4. *Adapted with data from unrelated languages*

Table 18 shows the performance for the four languages using these four models. First, we note that all adapted models perform better than the unadapted ones, even when adaptation only uses data from unrelated languages. Second, there is very little difference between the models adapted with target data (2) and those adapted using data from related languages (3). Only for verb agreement mistakes, there is about 1 point difference in F1 performance, whereas for the other errors results are quite close. However, adaptation using target data substantially outperforms adaptation using unrelated language data, by 1.1 to 2.7 F1 points for different errors. Similarly, there is a big gap when using data from related languages vs. unrelated ones: 1.5 F1 points or more of a difference. On average, adaptation that uses unrelated language error data gives a relative improvement with respect to unadapted models between 3.49% and 5.45%, and for adaptation that is based on related language data and target language data improvements range between 7.85% to 13.53% and between 7.17% to 16.12%, respectively.

To conclude, the experiments that separate related and unrelated language priors indicate that error data from related languages helps much more than error data from unrelated languages. Thus, when we only have data from unrelated languages available, it makes sense to invest a little bit and add data from the target language.

Table 18
Adaptation using data from target language, related languages, and unrelated languages. All models are trained using NB on the Web1T corpus. In all settings, the total number of examples used to estimate priors is the same. Averaged results for Russian, Polish, Spanish, and Italian. Best results for each phenomenon are emphasized in **bold**. In the parentheses, relative improvement of each adapted model with respect to the corresponding native-trained model is shown. All improvements of adapted models vs. native-trained models and improvements of the best adapted models vs. models adapted using unrelated language data are statistically significant (McNemar's test, $p < 0.001$).

Error type	(1) Native-trained	Performance (F1) (2-4) Adapted models		
		(2) Target	(3) Related	(4) Unrel.
Article	30.70	32.90 (7.17%)	**33.11** (7.85%)	31.77 (3.49%)
Prep.	22.40	**26.01** (16.12%)	25.43 (13.53%)	23.62 (5.45%)
Verb agr.	34.67	**38.76** (11.80%)	37.56 (8.34%)	36.07 (4.04%)

6. Related Work

Two Approaches to Grammar Correction There are two dominant approaches to grammatical error correction: Classification and statistical machine translation (MT). Not only are they orthogonal technically, they have two key differences in terms of *the amount of annotated data required* and in terms of *the types of mistakes* they address. The two approaches are complementary, and each of them is essential to getting good results (Rozovskaya and Roth 2016). This article deals with the classifier-based approach.

In terms of differences in targeting different types of mistakes, results in the literature show that the approach studied in this work can be used to improve state-of-the-art MT models, as some error phenomena, such as article and verb agreement errors, are better handled using classifiers (Rozovskaya and Roth 2016).

Regarding the amount of supervision, MT systems require significant annotation: All of the recently published MT systems are trained using Lang-8 corpus, which contains between 11M and 48M words of annotated learner data, depending on the corpus version and the pre-processing that was performed (Susanto, Phandi, and Ng 2014; Chollampatt, Taghipour, and Ng 2016; Hoang, Chollampatt, and Ng 2016; Junczys-Dowmunt and Grundkiewicz 2016; Mizumoto and Matsumoto 2016; Yuan and Briscoe 2016).

Although most of the recent publications report results on the CoNLL-2014 test set, those systems are evaluated on global behavior (i.e., the whole corpus) and do not focus or evaluate performance on specific error phenomena, as we do in this work. Further, because we care about specific first-language backgrounds, CoNLL is not an appropriate data set to use: The FCE corpus is the only one that contains data from learners of multiple language backgrounds and information on the first language of the writer, used in Section 5. Therefore, in this work we make use of the FCE data set.

However, we also show that the models described here are competitive with state-of-the-art on the CoNLL-2014 test set. Table 19 shows the performance of the classifier system from our most recent work (Rozovskaya and Roth 2016) that is based on the classifiers described here, and places it in the context of other recently published systems. The systems are divided into three categories: classifier-based, MT, and combined. For each system, the size of the annotated learner data used to train the system is also shown. Depending on the type of training data available, a specific approach should be preferred. In particular, classification systems can be trained from significantly less annotated data and can focus on specific errors, whereas MT-based systems do not have this capability. Best result in each category is in bold.

Observe that the performance of the MT systems depends heavily on the size of the supervision. Specifically, when an MT system is trained only on the CoNLL corpus, the performance is quite poor (28.25). Thus, all other MT systems use an additional source of supervision—either the Cambridge Learner corpus (CLC), which is a larger version of FCE, or the publicly available Lang-8 corpus. The Lang-8 corpus used in the literature has several versions, depending on when it was collected and how much noise has been removed from it. The smallest version used by Hoang, Chollampatt, and Ng (2016) and Chollampatt, Taghipour, and Ng (2016) contains about 11M words; the one in Mizumoto and Matsumoto (2016) is about twice as large; the version in Junczys-Dowmunt and Grundkiewicz (2016) contains about 30M words; the MT component in Rozovskaya and Roth (2016) has 48M words. Because of these differences in size and quality, the comparisons are not fair. For instance, Junczys-Dowmunt and Grundkiewicz (2016) report an F0.5 score of 52.21 using a version closer to 50M words, which was used in

Table 19
State-of-the-art systems on CoNLL-2014 data set. The systems are divided into three categories: classifiers, MT, and combined. For each system, we also show the size of the annotated learner data used to train the system. Depending on the type of training data available, a specific approach should be preferred. In particular, classification systems can be trained from significantly less annotated data and can focus on specific errors, whereas MT-based systems do not have this capability. Best result in each category is in **bold**.

System type	System name	Annotated learner data			F0.5
		CoNLL 1.2M	Lang-8 11-48M	CLC 29M	
Classif.	Susanto, Phandi, and Ng (2014)	✓			35.44
	Rozovskaya and Roth (2016)	✓			**43.11**
MT	Mizumoto and Matsumoto (2016)	✓	✓		40.00
	Yuan and Briscoe (2016)	✓		✓	39.90
	Chollampatt, Taghipour, and Ng (2016)	✓	✓		41.75
	Hoang, Chollampatt, and Ng (2016)	✓	✓		41.19
	Rozovskaya and Roth (2016)	✓			28.25
	Rozovskaya and Roth (2016)	✓	✓		39.48
	Junczys-Dowmunt and Grundkiewicz (2016)	✓	✓		**49.49**
Combined	Susanto, Phandi, and Ng (2014)	✓	✓		39.39
	Rozovskaya and Roth (2016)	✓	✓		**47.40**

Rozovskaya and Roth (2016), however, because of the differences in pre-processing, these data sets are also not the same.

Note that the system presented in this paper is better than most of the recently published works, including those in the MT category, even though the classifiers only use the CoNLL training corpus (1.2M words) as learner data, whereas all other systems, all of which use the MT approach, train on much more annotated learner data. We wish to emphasize that the present work focuses on approaches that use a small amount of annotation. This should be particularly useful when building error correction systems when very little or no annotated data is available, which is the case today for languages other than English.

Finally, it should be pointed out that, although the absolute F0.5 score for the combined system of Rozovskaya and Roth (2016) is slightly lower than the MT approach of Junczys-Dowmunt and Grundkiewicz (2016), the combined system uses a much weaker MT component (which scores 39.48 F0.5, when used by itself). We expect that a combined system that used a better MT component would perform significantly better than Junczys-Dowmunt and Grundkiewicz (2016) and, in addition, would handle better some types of mistakes that MT systems do not do well on.

Adaptation Using Artificial Errors Several other researchers study the effect of the adaptation framework that uses artificial errors. The two most closely related studies are the works by Cahill et al. (2013) and Felice and Yuan (2014). Cahill et al. researched the effects of different training paradigms with different data sets on the preposition error correction task. They show improvements when using artificial errors, although their selection models were not optimized with respect to F-score (typically, models trained on well-edited text use a threshold, as we do in this work, because

otherwise these models tend to have extremely low precision, which negatively affects the F-score). Because the original models were not optimized, it is not clear whether the improvements were due to adding artificial errors.

Felice and Yuan (2014) follow the adaptation framework originally proposed in Rozovskaya and Roth (2010b) but they introduce artificial errors by taking into account additional, contextual information, such as the POS tags of the head nouns for article errors, the surface form of the verbs for verb agreement mistakes, and semantic classes of the nouns for preposition mistakes. The adaptation method is evaluated in the context of training statistical machine translation systems for grammar correction, so that work is not directly comparable to ours. Furthermore, no improvements are reported of the baseline of training on a corpus containing natural errors.

Large-scale Error-annotated Corpora Two large-scale error-annotated corpora became available recently—Lang-8 (Mizumoto et al. 2012) and corpus Wikirev (Cahill et al. 2013)—that several researchers have used in developing grammar correction systems. Similar to results reported in this work, Cahill et al. (2013) show that using these corpora with artificially generated errors or naturally occurring errors typically outperforms models trained in the selection paradigm.[9] We wish to emphasize that the focus of the current work is on methods that use *minimal supervision*, when error-annotated data are not available, such as low-frequency errors and languages other than English.

7. Discussion and Conclusion

This article addressed the question of developing an appropriate training paradigm for error correction tasks that can consider the regularities in learner errors using *minimal supervision*. Although several large-scale error-annotated corpora recently became available in English, such corpora require significant effort to create and we are therefore unlikely to reach a state where such resources exist for all, or even many, languages of interest. Thus, the techniques that we describe are going to be very useful once this line of work extends to other languages that do not have many resources. Further, even when one has annotation of learner data in a specific genre, it could be different from the genre of the data that one needs to correct. In this case, one can envision applying the adaptation framework and training models on native data that is of a similar genre to the target learner data. More generally, the techniques that we develop should be useful once one moves out of the familiar, ideal setting where we have annotation of learner data that is both in the target language and of a similar type of writing.

The proposed training paradigm is made possible using parameters related to error patterns, which are very simple and thus do not require a lot of annotated data for estimation. This is in contrast to features encoding contextual information, which can be estimated using native English data. As a result, the proposed adaptation approach allows one to combine large amounts of native data (which is cheap and can be used to learn context features) with a small amount of annotated ESL data for estimating error patterns.

We described how error patterns can be learned from a small annotated sample and proposed two methods of "injecting" knowledge about error regularities into models

9 One issue here is that selection models are not optimized in this work and the results are reported without decision thresholds.

that are otherwise trained on native English data, within the framework of two state-of-the-art machine learning algorithms: a discriminative classifier (Averaged Perceptron) and Naive Bayes. NB and AP were chosen as two algorithms that exhibited top performance in error correction tasks, with AP being the best model (Rozovskaya and Roth 2011). It is important to emphasize, though, that although discriminative classifiers tend to outperform NB, sometimes it is preferable not to train discriminatively. For instance, we used the Web1T corpus that contains *pre-computed n*-gram counts that can be used to train a NB model, but not an AP model. Another advantage of the NB adaptation is that it does not require generating new training data and re-training in order to adapt to a different error distribution.

We showed that adapted models outperform native-trained models that use the same amount of native data for training. Crucially, the effect of adaptation remains strong as the size of the native training data is increased. This is because learner data contains knowledge about error patterns that is not available in native data. The adapted models also outperform ESL-trained models that use the same amount of data. This is because context parameters require much more data for estimation than error parameters, so it is a good idea to use native data for learning the parameters of the context features.

We first applied the adaptation approach across multiple first-language backgrounds and learned error regularities that are not specific to the first language background of the writer. Next, we used the idea that error patterns are often first-language-dependent and showed that performance can be further improved with target-language adaptation, and even in cases when error statistics are obtained from related languages or languages that exhibit similar behavior in error patterns.

In this work, we studied the adaptation framework, as applied to data produced by learners of intermediate and advanced language proficiency. In our previous work, we showed that adaptation is effective using multiple learner corpora that contain various levels of proficiency (intermediate [FCE], advanced [NUCLE, Illinois], and highly proficient non-native writers [HOO]). Our expectation is that adaptation would be effective also at beginning language levels, although it is more likely that such writing might require whole-sentence correction, which is out of the scope of this work.

Although this work focuses on a technical contribution to the text correction area and shows how to develop models that are aware of learner error patterns using minimal supervision, there is another important but orthogonal direction, which is how to eventually build a good system that is able to improve the quality of the text. This question is also related to evaluation metrics, which we touched upon in Section 4.3. The F-measure by itself does not indicate what the quality of the text is as a result of running the system; instead, this is done by looking at *accuracy* and comparing it with the learner baseline. On the other hand, accuracy measure is only a single point and in order to get improvement it pushes the system toward low recall.

Thus one important direction for future work is a focus on designing a system that can improve the quality of the text compared with the learner baseline. This evaluation should also take into account the question of annotation: Because we are starting with such a high baseline, having annotation judgments for as many valid choices as possible is key. In fact, Bryant and Ng (2015) show that evaluation performance increases with the number of annotators used to correct the same text, as multiple annotators can better account for the fact that there are typically multiple ways to correct the same text (as we discussed in Section 4.6 in the context of correcting preposition mistakes). These findings indicate that the results obtained using just one annotator (as is done in the FCE corpus) likely underestimates the system performance for all errors.

Appendix A. Features

Table A.1
Features used in the article error correction system trained discriminatively. *wB* and *wA* denote the word immediately before and after the target, respectively; and *pB* and *pA* denote the POS tag before and after the target. *headW* denotes the NP head. *NC* stands for noun compound and is active if the second to last word in the NP is tagged as a noun. *Verb* features are active if the NP is the direct object of a verb. *Preposition* features are active if the NP is immediately preceded by a preposition. *Adj* feature is active if the first word (or the second word preceded by an adverb) in the NP is an adjective. *NpWords* and *npTags* denote all words (POS tags) in the NP. Adapted models also use the author's article as a feature.

Feature type	Feature group	Feature list
Word		wB, w_2B, w_3B, wA, w_2A, w_3A, $wBwA$, w_2BwB, wAw_2A, w_3Bw_2BwB, w_2BwBwA, $wBwAw_2A$, wAw_2Aw_3A, $w_4Bw_3Bw_2BwB$, w_3Bw_2BwBwA, $w_2BwBwAw_2A$, $wBwAw_2Aw_3A$, $wAw_2Aw_3w_4A$
POS		pB, p_2B, p_3B, pA, p_2A, p_3A, $pBpA$, p_2BpB, pAp_2A, $pBwB$, $pAwA$, p_2Bw_2B, p_2Aw_2A, p_2BpBpA, $pBpAp_2A$, pAp_2Ap_3A
	NP_1	*headW*, *npWords*, *NC*, *adj*&*headW*, *adjTag*&*headW*, *adj*&*NC*, *adjTag*&*NC*, *npTags*&*headW*, *npTags*&*NC*
	NP_2	*headW*&*headPOS*, *headNumber*
Chunk	wordsAfterNP	*headW*&*wordAfterNP*, *npWords*&*wordAfterNP*, *headW*&*2wordsAfterNP*, *npWords*&*2wordsAfterNP*, *headW*&*3wordsAfterNP*, *npWords*&*3wordsAfterNP*
	wordBeforeNP	$wB\&f_i \; \forall i \in NP_1$
	Verb	*verb*, $verb\&f_i \; \forall i \in NP_1$
	Preposition	$prep\&f_i \; \forall i \in NP_1$

Table A.2
Features used in the preposition error correction system trained discriminatively. *wB* and *wA* denote the word immediately before and after the target, respectively; the other features are defined similarly. *compHead* denotes the head of the preposition complement. *wB*&*compHead*, w_2BwB&*compHead* are feature conjunctions of *compHead* with *wB* and w_2BwB, respectively. Adapted models also use the author's preposition as a feature.

Feature Type	Feature list
Word n-gram	wB, w_2B, w_3B, wA, w_2A, w_3A, $wBwA$, w_2BwB, wAw_2A, w_3Bw_2BwB, w_2BwBwA, $wBwAw_2A$, wAw_2Aw_3A, $w_4Bw_3Bw_2BwB$, w_3w_2BwBwA, $w_2BwBwAw_2A$, $wBwAw_2Aw_3A$, $wAw_2Aw_3w_4A$
Preposition complement	*compHead*, *wB*&*compHead*, w_2BwB&*compHead*

Table A.3
Features used in the verb agreement error correction system trained discriminatively. *wB* and *wA* denote the word immediately before and after the target, respectively; and *pB* and *pA* denote the POS tag before and after the target. Syntactic features are described in Table A.4. Adapted models also use the author's verb as a feature.

Feature type	Feature list
Word	wB, w_2B, w_3B, wA, w_2A, w_3A, $wBwA$, w_2BwB, wAw_2A, w_3Bw_2BwB, w_2BwBwA, $wBwAw_2A$, wAw_2Aw_3A, $w_4Bw_3Bw_2BwB$, w_3Bw_2BwBwA, $w_2BwBwAw_2A$, $wBwAw_2Aw_3A$, $wAw_2Aw_3w_4A$
POS	pB, p_2B, p_3B, pA, p_2A, p_3A, $pBpA$, p_2BpB, pAp_2A, $pBwB$, $pAwA$, p_2Bw_2B, p_2Aw_2A, p_2BpBpA, $pBpAp_2A$, pAp_2Ap_3A
Syntax	*subjHead*, *subjPOS*, *subjDet*, *subjDistance*, *subjNumber*, *subjPerson*, *subjHead*&*subjDistance*, *subjPOS*&*subjDistance*, *subjNumber*&*subjPerson*

Table A.4
Verb agreement features that use syntactic knowledge.

	Feature name	Description
(1)	subjHead, subjPOS	the surface form and the POS tag of the subject head
(2)	subjDet	determiner of the subject NP
(3)	subjDistance	distance between the verb and the subject head
(4)	subjNumber	*Sing* – singular pronouns and nouns; *Pl* – plural pronouns and nouns
(5)	subjPerson	*3rdSing* – "she", "he", "it", singular nouns; *Not3rdSing* – "we", "you", "they", plural nouns; *1stSing* – "I"
(6)	conjunctions	(1) and (3); (4) and (5)

Acknowledgments

The authors thank the anonymous reviewers for the insightful comments on the article. This material is partly based on research sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under agreement numbers FA8750-13-2-0008 and HR0011-15-2-0025, and by the Army Research Laboratory under agreement W911NF-09-2-0053. This research is also partly supported by a grant from the U.S. Department of Education and by the DARPA Machine Reading Program under Air Force Research Laboratory prime contract no. FA8750-09-C-018. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the view of the agencies.

References

Banko, M. and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL-01)*, pages 26–33, Toulouse.

Bergsma, S., D. Lin, and R. Goebel. 2009. Web-scale *n*-gram models for lexical disambiguation. In *Proceedings of 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pages 1507–1512, Pasadena, CA.

Berzak, Y., R. Reichart, and B. Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL-15)*, pages 94–102, Beijing.

Brants, T. and A. Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA.

Bryant, C. and H. T. Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-15)*, pages 697–707, Beijing.

Cahill, A., N. Madnani, J. Tetreault, and D. Napolitano. 2013. Robust systems for preposition error correction using wikipedia revisions. In *Proceedings of the 13th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-13)*, pages 507–517, Atlanta, GA.

Carlson, A. and I. Fette. 2007. Memory-based context-sensitive spelling correction at Web scale. In *Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA-07)*, pages 166–171, Washington, DC.

Carlson, A. J., J. Rosen, and D. Roth. 2001. Scaling up context sensitive text correction. In *Proceedings of the 13th Conference on Innovative Applications of Artificial Intelligence (IAAI-01)*, pages 45–50, Seattle, WA.

Chodorow, M., J. Tetreault, and N.-R. Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 25–30, Prague.

Chollampatt, S., K. Taghipour, and H.-T. Ng. 2016. Neural network translation models for grammatical error correction. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 2768–2774, New York, NY.

Dahlmeier, D. and H. T. Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-11)*, pages 915–923, Portland OR.

Dahlmeier, D. and H. T. Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-12)*, pages 568–578, Jeju Island.

Dahlmeier, D., H. T. Ng, and S. M. Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, GA.

Dale, R., I. Anisimoff, and G. Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62, Montreal.

Dale, R. and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG-11)*, pages 242–249, Nancy.

De Felice, R. 2008. *Automatic Error Detection in Non-native English*. Ph.D. thesis, University of Oxford.

De Felice, R. and S. Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, pages 45–50, Prague.

De Felice, R. and S. Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 169–176, Manchester.

Eeg-Olofsson, J. and O. Knuttson. 2003. Automatic grammar checking for second language learners—the use of prepositions. In *Proceedings of the 14th Nordic Conference of Computational Linguistics (Nodalida-03)*, Reykjavic.

Felice, M. and T. Briscoe. 2015. Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-15)*, pages 578–587, Denver, CO.

Felice, M. and Z. Yuan. 2014. Generating artificial errors for grammatical error

correction. In *Proceedings of the Student Research Workshop at the European Chapter of the Association for Computational Linguistics*, pages 116–126, Gothenburg.
Felice, M., Z. Yuan, Ø. Andersen, H. Yannakoudakis, and E. Kochmar. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL-14): Shared Task*, pages 15–24, Baltimore, MD.
Foster, J. and Ø. Andersen. 2009. Generrate: Generating errors for use in grammatical error detection. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 82–90, Boulder, CO.
Gamon, M. 2010. Using mostly native data to correct errors in learners' writing. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-10)*, pages 163–171, Los Angeles, CA.
Gamon, M. 2011. High-order sequence modeling for language learner error detection. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 180–189, Portland, OR.
Gamon, M., J. Gao, C. Brockett, A. Klementiev, W. Dolan, D. Belenko, and L. Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 449–456, Hyderabad.
Gass, S. and L. Selinker. 1992. *Language Transfer in Language Learning*. John Benjamins.
Golding, A. R. and D. Roth. 1996. Applying Winnow to context-sensitive spelling correction. In *Proceedings of the 13th International Conference on Machine Learning (ICML-96)*, pages 182–190, Bari.
Golding, A. R. and D. Roth. 1999. A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130.
Granger, S., E. Dagneaux, and F. Meunier. 2002. *International Corpus of Learner English*. Presses universitaires de Louvain.
Grundkiewicz, R., M. Junczys-Dowmunt, and E. Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP-15)*, pages 461–470, Lisbon.
Han, N., M. Chodorow, and C. Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Journal of Natural Language Engineering*, 12(2):115–129.
Han, N., J. Tetreault, S. Lee, and J. Ha. 2010. Using an error-annotated learner corpus to develop an ESL/EFL error correction system. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-10)*, pages 763–770, Valetta.
Hoang, D.-T., S. Chollampatt, and H.-T. Ng. 2016. Exploiting *n*-best hypotheses to improve an SMT approach to grammatical error correction. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 2803–2809, New York, NY.
Ionin, T., M. L. Zubizarreta, and S. Bautista. 2008. Sources of linguistic knowledge in the second language acquisition of English articles. *Lingua*, 118:554–576.
Izumi, E., K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 145–148, Sapporo.
Junczys-Dowmunt, M. and R. Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL-14): Shared Task*, pages 25–33, Baltimore, MD.
Junczys-Dowmunt, M. and R. Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP-16)*, pages 1546–1556, Austin, TX.
Lee, J. and S. Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 174–182, Columbus, OH.
Mizumoto, T., Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING-12)*, pages 863–872, Bombay.
Mizumoto, T. and Y. Matsumoto. 2016. Discriminative reranking for grammatical

error correction with statistical machine translation. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-16)*, pages 1133–1138, San Diego, CA.

Montrul, S. 2000. Transitivity alternation in L2 acquisition: Toward a modular view of transfer. *Studies in Second Language Acquisition*, 22:229–273.

Montrul, S. and R. Slabakova. 2002. Acquiring morphosyntactic and semantic properties of preterite and imperfect tenses in L2 Spanish. In A.-T. Perez-Laroux and J. Liceras, editors, *The Acquisition of Spanish Morphosyntax*. Springer, Dordrecht, pages 115–151.

Nagata, R. and E. Whittaker. 2013. Reconstructing an Indo-European family tree from nonnative English texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-13)*, pages 1137–1147, Sofia.

Napoles, C., K. Sakaguchi, M. Post, and J. Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-15)*, pages 588–593, Beijing.

Ng, H. T., S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL-14): Shared Task*, pages 1–14, Baltimore, MD.

Ng, H. T., S. M. Wu, Y. Wu, Ch. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL-13): Shared Task*, pages 1–12, Sofia.

Odlin, T. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press.

Oh, E. and M. L. Zubizaretta. 2003. Does morphology affect transfer? The acquisition of English double objects by Korean native speakers. In *Proceedings of the 28th Annual Boston University Conference on Language Development*, pages 402–413, Boston, MA.

Parker, R., D. Graff, J. Kong, K. Chen, and K. Maeda. 2009. *English Gigaword Fourth Edition LDC2009T13*. Linguistic Data Consortium, Philadelphia, PA.

Roth, D. 1998. Learning to resolve natural language ambiguities: A unified approach. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 806–813, Madison, WI.

Rozovskaya, A., K.-W. Chang, M. Sammons, and D. Roth. 2013. The University of Illinois system in the CoNLL-2013 shared task. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL-13): Shared Task*, pages 13–19, Sofia.

Rozovskaya, A., K.-W. Chang, M. Sammons, D. Roth, and N. Habash. 2014. The University of Illinois and Columbia system in the CoNLL-2014 shared task. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-14): Shared Task*, pages 34–42, Baltimore, MD.

Rozovskaya, A. and D. Roth. 2010a. Generating confusion sets for context-sensitive error correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 961–970, Boston, MA.

Rozovskaya, A. and D. Roth. 2010b. Training paradigms for correcting errors in grammar and usage. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-10)*, pages 154–162, Los Angeles, CA.

Rozovskaya, A. and D. Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-11)*, pages 924–933, Portland, OR.

Rozovskaya, A. and D. Roth. 2014. Building a state-of-the-art grammatical error correction system. *Transactions of ACL*, 2:419–434.

Rozovskaya, A. and D. Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-16)*, pages 2205–2215, Berlin.

Rozovskaya, A., M. Sammons, J. Gioja, and D. Roth. 2011. University of Illinois system in HOO text correction shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG-11)*, pages 263–266, Nancy.

Rozovskaya, A., M. Sammons, and D. Roth. 2012. The UI system in the HOO 2012 shared task on error correction. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building*

Educational Applications, pages 272–280, Montreal.

Susanto, R. H., P. Phandi, and H. T. Ng. 2014. System combination for grammatical error correction. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP-14)*, pages 951–962, Doha.

Tetreault, J. and M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, pages 865–872, Manchester.

Tetreault, J., J. Foster, and M. Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-10)*, pages 353–358, Uppsala.

Yannakoudakis, H., T. Briscoe, and B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-11)*, pages 180–189, Portland, OR.

Yi, X., J. Gao, and W. Dolan. 2008. A Web-based English proofing system for ESL users. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 619–624, Hyderabad.

Yuan, Z. and T. Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-16)*, pages 380–386, San Diego, CA.