Natural Language Engineering **22** (2): 163–186. © Cambridge University Press 2015
doi:10.1017/S1351324915000303

Non-native text analysis: A survey

SEAN MASSUNG and CHENGXIANG ZHAI
Department of Computer Science, College of Engineering, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA
e-mail: massung1@illinois.edu, czhai@illinois.edu

(*Received 25 November 2014; revised 16 July 2015; accepted 20 July 2015; first published online 7 September 2015*)

Abstract

Non-native speakers of English far outnumber native speakers; English is the main language of books, newspapers, airports, air-traffic control, international business, academic conferences, science, technology, diplomacy, sports, international competitions, pop music, and advertising (British Council 2014). Online education in the form of massive online open courses is also primarily in English—even teaching English. This creates enormous amounts of text written by non-native speakers, which in turn generates a need for grammar correction and analysis. Even aside from massive online open courses, the number of English learners in Asia alone is in the tens of millions. In this paper, we provide a survey of the two main areas of existing work on non-native text analysis, prefaced by an overview of common datasets used by researchers, comparing their attributes and potential uses. Then, an introduction to native language identification follows: determining the native language of an author based on text in the second language. This section is subdivided into various techniques and a shared task on this classification problem. Next, we discuss non-native grammatical error correction—finding and modifying text to fix errors or to make it sound more fluent. Again, we discuss different methods before investigating a relevant shared task. Lastly, we end with conclusions and potential future directions. While this survey primarily focuses on detecting and correcting non-native *English* text, many approaches are general and can be used across any language pairing.

1 Introduction

By the year 2020, the British Council (2014) estimates that there will be two billion English language learners. Some learn in the classroom; some learn online. Some may even learn through their phone or in an online class. Regardless of the medium, computational tools to enhance this educational experience will be valuable. Automatic scoring of essays—not only for grammar, but also fluency—would contribute greatly to second-language learners' understanding. User personalization for online services (including search engines and social networks) would benefit from improved user profiling. More relevant books or news articles could be recommended if the user's background and competency of English were known.

Due to these many motivating examples, research in *non-native text analysis* has prospered. This field encompasses any textual task that deals with words written in a language other than the writer's native tongue. We call the native language L1 and the second, learned language L2. Throughout this survey, we will usually assume

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

that L2 is English, though most (but not all) techniques discussed in this survey are general and could function with any pair of L1 and L2.

Existing work in non-native text analysis is somewhat scattered, appearing in workshops and shared tasks more frequently than standard proceedings. In this survey, we provide a systematic overview of these scattered works to make it easier for researchers to digest the state of the art in this emerging area.

1.1 Scope

The current work on non-native text analysis generally falls into two main categories:

(1) **Native Language Identification** (NLI): Classifying L1 based on text written in L2. Techniques can be categorized into **feature-based** (using a classifier) or **likelihood-based** (using a probabilistic model).
(2) **Non-native Grammatical Error Correction** (GEC): Detecting and correcting grammatical errors in L2 text. Techniques can be categorized into **targeted** (correcting specific errors) or **general** (correcting all errors).

Other subareas such as fluency scoring, non-native essay scoring, and text simplification for non-native speakers also exist, but we focus this survey on the two much larger areas explained above.

An excellent overview of GEC with a focus on non-native learners is available (Leacock *et al.* 2014). This short book is a concise collection on the topic and consists of many recent advances since 2010. If the reader wishes to delve into more detail in this subtopic, we suggest referencing their work, whereas our paper is a broader outline and thus not able to go into as much depth in one particular area.

Spelling correction described in Kukich (1992) is a potentially relevant task, though we choose to focus on the grammar correction aspect instead. This is not to say that spelling features do not play a role in (e.g.) NLI; rather, the method of spelling correction is not relevant to non-native text analysis. Correcting speech is also an important issue: besides the obvious speech recognition challenges in computer science and electrical engineering, speakers tend to voice their words much differently than when writing. Therefore, we do not include any investigation into these works in this paper.

Shared tasks provide a common goal and dataset to a wide array of researchers. This enables quick and accurate comparison of different methods, while simultaneously increasing interest and producing exposure for the problem in hand. Thus, we detail two relevant shared tasks in this survey that deal with non-native writers of English. Author profiling, authorship attribution, and plagiarism detection at PAN 2014 (Rangel *et al.* 2014) are also related, but we do not focus on those here, instead considering them related works since there is no specific component of non-native analysis.

Shared tasks have even started to evolve for L2s besides English. For example, the Techniques for Educational Applications Workshop had a shared task on GEC for Chinese as a foreign language (Yu, Lee and Chang 2014).

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

1.2 Organization

We now review the structure of this paper. In each section, a figure briefly summarizes the papers discussed and the methods used. If applicable, a comparison of performance is also included. In this survey, we have a particular preference for system papers as demonstrated by our sections on shared tasks. By system papers, we mean research that immediately puts theory into practice, giving real-world results which allow for better interpretation of effectiveness by the reader.

First, in Section 2, we introduce common datasets used in works described throughout the rest of this survey. They should give some perspective on the task in hand and create a mental baseline for the reader. We feel that it is important to make note of these datasets here since any future work will most likely be making use of these corpora.

In Section 3.1, we start with an introduction to NLI, usually formulated as a classification problem. Observed data is text written in L2 (where L2 means English: the second or foreign language), and a classifier predicts the writers' L1 (the native language). Although this may be considered a standard classification problem, the features chosen to target non-native writer discrimination distinguish different approaches. These range from simple unigram word counting to parse tree reranking descriptions. In future sections, we will see how classifiers themselves may be adapted or configured to take advantage of this specific domain.

Section 3.2 discusses the NLI Shared Task 2013 from the Building Educational Applications workshop. This shared task uses a corpus detailed in Section 2.

Section 4.1 discusses correcting grammatical errors, such as articles and prepositions. These techniques are usually framed as classification problems; a classifier determines if there is an error or not, and if so, predicts what the best solution would be given a fixed set of candidates. Some methods go beyond a simple classification setup and rely on more NLP-heavy techniques such as language models and grammatical parse trees to find the best corrected sentence.

Then, Section 4.2 covers the CoNLL 2013 and 2014 shared task on GEC. This task also uses a corpus described in Section 2, and systems find and correct errors in text written by non-native English speakers.

Finally, we end with a discussion and conclusions in Section 5, reflecting on what issues may still be unaddressed.

2 Non-native corpora

All the work on non-native text analysis uses some kind of non-native text corpus. In this section, we give an overview of the major datasets used in existing work, which we will survey in the next few sections. As shown in Table 1, the sizes of these corpora vary from around 1,000 to 12,000 documents with an average length of 128 to over 800 words. They cover a wide L2 range, geographically focused in Europe and Asia. Abbreviations used in the table are ARA (Arabic), BUL (Bulgarian), CAT (Catalan), CHN (Chinese), CZE (Czech), ENG (English), FIL (Filipino), FIN (Finnish), FRA (French), GER (German), GRC (Greek), HKG (Hong Kong Cantonese), IND (Indian languages), ITA (Italian), JPN (Japanese), KOR (Korean),

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Table 1. *Comparison of non-native English datasets: corpus statistics, and whether they can be used for NLI or GEC. Web links to each dataset are included in the text. Note that average document length may vary depending on tokenization method*

Dataset	Docs	Len_{avg}	NLI	GEC	L1 Languages
CEEAUS	1,008	128	✓		CHN, ENG, JPN (3)
CLEC	(1M	sent.)		✓	CHN (1)
FCE	1,244	200	✓	✓	CAT, CHN, FRA, GER, GRC, ITA, JPN, KOR, NL, POL, PRT, RUS, SPA, SWE, THA, TUR (16)
ICLE	3,640	700	✓		BUL, CZE, FIN, FRA, GER, ITA, NL, POL, RUS, SPA, SWE (11)
ICLEv2	6,085	617	✓		BUL, CHN, CZE, FIN FRA, GER, ITA, JPN, NL, NOR, POL, RUS, SPA SWE, TSW, TUR (16)
ICNALE	5,600	129	✓	✓	CHN, ENG, FIL, HKG, IND, JPN, KOR, PAK, SIN, THA, TWN (11)
NUCLE	1,397	863		✓	unknown
ETS	12,100	348	✓	✓	ARA, CHN, FRA, GER, IND, ITA, JPN, KOR, SPA, TEL, TUR (11)

NL (Dutch), NOR (Norwegian), PAK (Urdu), POL (Polish), PRT (Portuguese), RUS (Russian), SIN (Singapore languages), SPA (Spanish), SWE (Swedish), TEL (Telugu), THA (Thai), TSW (Tswana), TUR (Turkish), and TWN (Taiwanese). We now give a detailed description of each dataset.

ICLE (Granger 2003), the International Corpus of Learner English, is an early popular dataset used to investigate NLI. Its popularity is mostly due to the fact that no other large NLI corpora existed at the time the ICLE was released. It is a collection of essays on 'extremely varied' topics written by undergraduate students studying English. Additional metadata for each essay is part of the ICLE—such as gender and age—though these features are rarely used by researchers. Commonly, a five-language subset of Russian, Czech, Bulgarian, French, and Spanish languages is used, popularized by Koppel, Schler and Zigdon (2005). These subsets caught on because the L1 distribution is quite unbalanced; the subsets have a more uniform distribution of native languages. Thus, it is important to note if a subset of the ICLE is used so accurate comparisons may be made. The corpus is publicly available, but requires a license.[1]

ICLEv2 (Granger *et al.* 2009) is an expanded version of the ICLE. There are no major differences in the text files themselves (aside from the fact there are more),

[1] http://www.uclouvain.be/en-cecl-icle.html

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

since the ICLE is a subset. Version two additionally comes with a *concordancer*, a graphical interface that allows easy search and simple exploratory analysis over the dataset. However, when running experiments detailed in future sections of this paper, researchers tend to use the raw text data from the corpus instead of any built-in software. The corpus is publicly available, but requires a license.[2] The maintainers of ICLEv2 are currently working towards releasing a third version.

CEEAUS (Ishikawa 2009) is the Corpus of English Essays written by Asian University Students. Unlike the ICLE, the essays are restricted to two topics: *It is important for college students to have a part-time job* and *Smoking should be completely banned at all restaurants in the country*. A common vocabulary for each prompt emerges due to this fact, perhaps allowing a more meaningful word usage analysis to occur. Further restrictions attempt at creating the most uniform testing conditions as possible: spell checking is required, but dictionary use is prohibited; the writers have twenty to forty minutes to complete their task. Aside from its small size, the main downside to this corpus is that 77 per cent of the essays are from native Japanese speakers, establishing a rather high classification baseline for NLI. A Website exists for the corpus, but it is no longer available to download[3] since it is a subset of ICNALE.

CLEC (Gui and Yang 2003) is the Chinese Learner English Corpus. It consists of one million error-tagged sentences and is mainly used for GEC. While there is no differentiation in L1, the dataset is segmented into levels of English fluency. So while NLI classification is not possible, an analysis on the skill level is feasible. Although free, the CLEC is only available to members of the university that created it.[4]

FCE, created by Yannakoudakis, Briscoe and Medlock (2011), is the First Certificate of English dataset—a corpus of essays written by students studying English that took the Cambridge Assessment's English as a second or other language exam. The dataset was compiled as a benchmark for automated essay scoring systems, and contains about eighty types of correction annotations in each essay. Like the ICLE, additional metadata is also included, such as grade, age, and native language. There are ten different prompts, each with a varying number of responses. This corpus may be downloaded freely online.[5]

ICNALE (Ishikawa 2013), the International Corpus Network of Asian Learners of English, is an expanded version of CEEAUS. Like the ICLEv2, the previous version is a subset of this one. Aside from adding more Asian L1 languages, the language distribution of the dataset is much more balanced. Professional proof readers corrected the part-time job essays, and their revised versions are also included. This provides an opportunity to see how the non-native text is translated into sounding more native—with this information, it is possible to do much more than NLI. The last segment of the ICNALE is a collection of 1,900 recorded

[2] http://www.uclouvain.be/en-277586.html
[3] http://language.sakura.ne.jp/s/kaken_ceeaus.html
[4] http://langbank.engl.polyu.edu.hk/corpus/clec.html
[5] http://ilexir.co.uk/applications/clc-fce-dataset/

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

speeches and their transcripts. After registering online, the corpus is available[6] to freely download.

NUCLE (Dahlmeier, Ng and Wu 2013), the National University of Singapore Corpus of Learner English is also a collection of essays on a wide variety of subjects. The essays were written by students studying English at the Center for Language Communication at NUS and annotated and corrected by English instructors. Overall, nearly 47,000 errors were detected by the instructors with good rater agreement. Findings on the corpus revealed that overall errors in non-native text are actually quite rare (at least in this corpus), on the order of around four per one hundred words. Since this corpus was not intended for NLI, there is no information about the L1 language of each essay. This is the dataset used by the CoNLL 2013 shared task described in Section 4.2, and is available for download online.[7]

ETS (Blanchard *et al.* 2013) is a collection of essays from the Test of English as a Foreign Language, a component of the Educational Testing Service. It is a standardized test measuring English proficiency for students seeking to enroll in college. There are 1,100 essays per L1, sampled from eight prompts. The purpose of the corpus was to further enable work in NLI, and was used by the BEA8 shared task described in Section 3.2. ETS is publicly available, but requires a license.[8]

From our comparison, we have two datasets that are subsets of another (CEEAUS and ICLE). There is likely no reason to prefer CEEAUS over ICNALE or ICLE over ICLEv2, since the newer versions are simply expanded. This leaves six main datasets: CLEC, ETS, FEC, ICLEv2, ICNALE, and NUCLE. Of these six, CLEC, ETS, FCE, ICNALE, and NUCLE, could be used for error analysis since they contain annotations. All except CLEC and NUCLE can be used for NLI since the L1 information is part of the datasets.

As for content, all are essays written by university students, but only ICNALE and ETS have the choice of essay topics strictly controlled—though the former has two prompts and the latter has eight. ICNALE L1s are clearly focused in Asia; NUCLE writers are most likely from Singapore, while FCE, ICLEv2, and ETS L1s are more evenly distributed across the globe. Of course, CLEC consists of all native Chinese authors.

Considering these differences, we need to think about the following questions when interpreting results. What is the baseline accuracy? How difficult is it to distinguish L1s? Does the number of documents (and document length) have any bearing on an impression of the results?

Although not a standardized corpus, the website Lang-8[9] is also a popular location to retrieve non-native English. This is a site where users write in a second language and get corrections from native speakers. Corpora derived from this site have been used for NLI (e.g. Brooke and Hirst (2011)) and GEC (e.g. Tajiri, Komachi and Matsumoto (2012)).

[6] http://language.sakura.ne.jp/icnale/download.html
[7] http://www.comp.nus.edu.sg/ nlp/corpora.html
[8] https://catalog.ldc.upenn.edu/LDC2014T06
[9] http://www.lang-8.com

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

In the next sections, we systematically review the two areas of non-native text analysis.

3 Native language identification

NLI is usually the first step in any second language error correction or author profiling system. Identifying the native language of an anonymous text was first popularized by Koppel *et al.* (2005). Brooke and Hirst (2012) does an extensive survey of NLI feature efficacy, and develops a robust model that works well when used across corpora. NLI tasks are most commonly evaluated solely on a small learner corpus usually consisting of student essays (as shown in Section 2). It was previously thought that lexical features would be biased or overfit towards essay topics, but a cross-corpus evaluation showed that this was not the case (Brooke and Hirst 2012).

For a more in-depth discussion of authorship attribution, we recommend the reader consult two well-known surveys (Stamatatos 2009) (Koppel, Schler and Argamon 2009). Many techniques common to authorship attribution and author profiling are also relevant to NLI.

As mentioned in Section 2, the ICLE was an early popular dataset to evaluate NLI tasks, especially using a subset of five European languages partitioned by Koppel *et al.* (2005). A table has been created (Table 2) to portray the summary described in Brooke and Hirst (2012). In addition to the results displayed in the table, Wong and Dras also attempted to perform dimensionality reduction with LDA (Blei, Ng and Jordan 2003) as feature generation; however, this was not a successful method.

Techniques for NLI can be categorized into two methods: feature-based and likelihood-based. Feature-based methods rely on informative features derived from the text and are fed to standard machine learning algorithms (usually SVM or MaxEnt). Likelihood-based methods learn a probabilistic model (usually a grammar or language model) for each L1 and assign a label based on the maximum likelihood model. We now move onto specific techniques applied in NLI.

3.1 Techniques in native language identification

The first feature-based method (Tsur and Rappoport 2007) found that incredibly simple top 200 frequent bigram character features fed to SVM led to 66 per cent accuracy on the five native languages. They claimed that word choice of non-native speakers is influenced by the phonology of their native language (as evidenced by the effectiveness of the character features). This approach is compared to a unigram words baseline which achieved only 47 per cent accuracy. They finally hypothesized that using a spoken-language corpus would achieve even stronger results favoring character bigrams. Their reasoning was that much less conscious effort is put into speaking than writing. For analyzing transcripts of spoken words, the ICNALE corpus may be applicable, described in Section 3.

NLI has also been approached through contrastive analysis (Wong and Dras 2009): the idea that errors in text are influenced by the native language of the

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Table 2. *Summary of native language identification results listed in (Brooke and Hirst 2012) for the ICLE corpus; accuracies added to chart.* * *indicates feature-based methods and* + *indicates likelihood-based methods*

Paper	Method	Accuracy
(Tsur and Rappoport 2007)	character *n*-grams*	66 per cent
(Wong and Dras 2009)	syntactic errors*	74 per cent
(Wong and Dras 2011)	syntactic rules*	80 per cent
(Wong *et al.* 2012)	adaptor grammars*+	76 per cent
(Swanson and Charniak 2012)	tree substitution grammars*+	78 per cent
(Stolerman *et al.* 2013)	language family*	80 per cent

author. They investigated three error types as features: subject-verb disagreement, noun-number disagreement, and determiner misuse. These error types are then used as 'stylistic markers' for NLI features with an SVM classifier. To find these errors in text, they used an open source grammar checker[10], as opposed to professionally edited text. Interestingly, ANOVA showed that the features had a measurable effect, but after combining their contrastive features with existing methods, they were not able to significantly increase the classification accuracy from Koppel *et al.* (2005).

The previous authors (Wong and Dras 2011) follow their work on contrastive analysis, attempting to amend its shortcomings. Instead of error types, they use two different features obtained from grammatical parse trees: horizontal slices (production rules) and parse rerankings. They claim these are the first pure syntactic features used in NLI. For the production rules, they immediately applied information gain dimensionality reduction. The reranking features are those contained in the Charniak parser[11] and Stanford Parser[12] trained on the Wall Street Journal. Unlike the previous two attempts, the authors found MaxEnt to outperform SVM as the classifier. Additionally, five-fold cross validation was performed (as opposed to ten-fold), which means the accuracies cannot be precisely compared with previous work. In any event, they report a final accuracy of 80 per cent, which was the highest reported as of 2012.

Wong, Dras and Johnson (2012) explore the last author's—Mark Johnson's—adaptor grammars (Johnson, Griffiths and Goldwater 2006) to generate features. Simply, adaptor grammars are a non-parametric extension to PCFGs (probabilistic context free grammars). They can learn arbitrary-length word sequences (collocations); for example, *gradient descent* and *cost function* were learned as phrases in a machine learning topic. These adaptor grammars are used in two ways: in the first, collocations are used as features in a MaxEnt classifier. In the second, the grammar is trained on each class (representing native language). At test time, the most probable grammar to have generated the text is selected. For both tasks, the authors use five-fold cross validation on seven native languages. In the feature-based

[10] http://queequeg.sourceforge.net/index-e.html
[11] http://cs.brown.edu/ ec/
[12] http://nlp.stanford.edu/software/lex-parser.shtml

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

classification, they achieved 76 per cent; in the language model-based classification, they achieved only 50 per cent, a performance similar to the unigram word baseline from Tsur and Rappoport (2007).

Swanson and Charniak (2012) made use of tree substitution grammars (TSGs) (Blunsom and Cohn 2010). TSGs are a tree-rewriting formalism that defines operations on partial (parse) tree objects. For example, subtrees may be added or removed from a base tree. Benefits of using this method are priors which prefer smaller production rules and the ability to capture long-range dependencies. Various induction methods are compared to generate features, and five-fold cross validation on seven native languages is performed. All TSG features outperformed the CFG baseline (at 73 per cent). The highest TSG induction method was Bayesian induction at 78 per cent.

Massung, Zhai and Hockenmaier (2013) also make use of grammatical parse tree features, but mainly focus on their structural aspects as opposed to the syntactic category labels. In one classification task, they found these features to work well in determining the nationality of student essay writers from the CEEAUS dataset. These structural parse tree features may be applicable in other tree-based objects such as adaptor grammars, TSGs, and dependency parses, but this has yet to be explored.

Lastly, Stolerman, Caliskan and Greenstadt (2013) use language family as a feature for NLI. Specifically, they took ICLEv2 documents classified with low confidence using unigram features and reclassified them into one of the Germanic, Slavic, or Romance language families. From there, the document was classified as an L1 using a subset of training data only from the language family it was classified as. This method is a type of hierarchical classification problem, and achieved about a 6 per cent increase over the baseline in NLI accuracy when using information gain feature selection.

In summary, Table 2 lists the comparable accuracies from experiments run on the ICLE/ICLEv2 subset of five European languages. In general though, accuracies between 70 per cent to 80 per cent are standard for a wide variety of techniques and corpora.

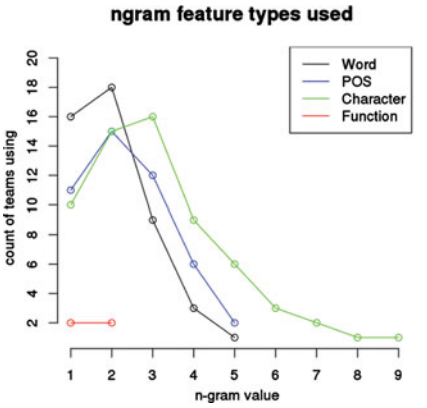
3.2 Shared task in native language identification

The eighth Building Educational Applications workshop held a shared task dedicated to NLI (Tetreault, Blanchard, and Cahill 2013). Its goal was to increase visibility of the problem and standardize results on a larger dataset: ETS. In total, twenty-nine teams participated in the shared task.

There were three separate subtasks, differentiated by what data was allowed for training:

- **Closed Training**: Use only the ETS corpus for training. All twenty-nine teams attempted this task.
- **Open Training I**: Use any data *except* the ETS corpus for training; the evaluation is still done on the ETS test set. Only three teams attempted this task.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press



Classifier	Count
SVM	13
Ensemble	4
MaxEnt	3
DFA	1
String kernel	1
PPM	1
k-NN	1

Fig. 1. (Colour online) Number of teams using particular *n*-gram features (left) and number of teams using a particular machine learning algorithm (right). DFA and PPM refer to discriminant function analysis and prediction by partial matching.

- **Open Training II**: Use any data *including* the ETS corpus. Only four teams attempted this task.

Overall results for the closed training were relatively close; the top thirteen teams had between 80–83.6 per cent accuracy. The open training I subtask teams performed much lower (probably as expected) at around 33 per cent, 38 per cent, and 57 per cent accuracy. Finally, in the open training II subtask, none of the four teams were able to beat the highest closed score. The most useful results for evaluating models are probably those of the closed task, since all training data is uniform, though it is interesting to see approaches employed to increase the amount of training data. For example, in the open subtasks, other NLI datasets were used such as ICLE, FCE, and ICNALE (all described in Section 2). Additionally, some teams crawled Lang-8, also described in Section 2.

The most common features used were *n*-grams of lexical tokens such as words or part-of-speech tags; Figure 1 compares the *n*-grams used by all teams. Aside from these, some less common features were grammatical: dependency parses, TSGs (Blunsom and Cohn 2010), parse tree rules, and adaptor grammars (Johnson *et al.* 2006) were used. Spelling error features were also captured by three of the teams. These features are all described in work in Section 3.1. Below, we outline a few of the more unique feature representations.

Skipgrams (Guthrie *et al.* 2006) were an *n*-gram variant feature used by several teams. Instead of considering an *n*-gram to consist of *n* adjacent words in the text, a *k*-skip *n*-gram allows up to *k* words total to be skipped in a sequence of tokens. Given the sentence '*They all studied statistics Monday evening*', normal 3-grams would be {*They all studied, all studied statistics, studied statistics Monday, statistics Monday evening*}. A few possible 2-skip 3-grams are {*They all studied, They studied statistics, They statistics Monday*}. This attempts to capture important phrases without regard to interspersed function words. Using skipgrams was shown to reduce perplexity on a testing set, as well as offer a viable alternative to increase the corpus size (which is usually not feasible). Since most NLI datasets we have examined are relatively

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

small, using skipgrams could help improve language modeling. While interesting, none of the teams in the top third used skipgrams in their systems.

For one feature type, the system by LIMSI used a form of machine translation called 'back translations' (Lavergne *et al.* 2013). A back translation attempts to capture a writer's lexical preference for a word sense. For example, LIMSI found that the English word sense for *awkward* is more likely to be written as *clumsy* by native Spanish speakers. These word preferences were used as features in addition to other features: word *n*-grams, spelling mistakes, and grammatical mistakes. Adding the back translation features slightly increased the task's overall classification accuracy, but the team still performed poorly overall.

The CMU-Haifa team used a ratio of passive to active verbs in their system (Tsvetkov *et al.* 2013). They operated under the common assumption that English uses passive voice more frequently than other languages, and the amount of passive use may vary based on the writer's L1. Recall that passive voice is a literary technique that shifts attention away from the one performing an action. '*The passive voice is often used*' is a sentence in the passive voice; '*English speakers often use the passive voice*' is a sentence in the active voice. The passive voice is characterized by the verb *to be* and the past participle of the verb. In the previous sentences, we have *is used* and *use* conjugated differently. Using the passive ratio feature exclusively yielded a 12 per cent accuracy on a 9 per cent baseline. In combination with four other main features, the accuracy was not significantly improved. However, further investigation could be warranted to examine the usefulness of such stylistic composition features.

SVM was by far the most prevalent classifier employed by the teams, as displayed in Figure 1. Below, we outline some of the more uncommon classifiers.

Discriminant function analysis from statistics was used for increased interpretability over other methods (Kyle *et al.* 2013). Using this model, they were able to find correlations between different L1s. For instance, Japanese and Korean L1s were highly correlated in their underuse of the words {*all, any, but, different, or, person, this, your*}. Unsurprisingly, languages of similar origin were also correlated—the Romance languages were often misclassified as one another due to increased correlation of *n*-gram tokens. Unfortunately, this team performed poorly overall.

Originally designed to operate on DNA sequences, the string kernel model (Popescu and Ionescu 2013) is given a stream of characters. Specifically, a kernel based on local rank distance is used for NLI. The simplest string kernel function $f(w_1, w_2)$ counts the number of substrings of a particular length that w_1 and w_2 share. It is also quite intuitive to introduce a normalized version that is not biased by long words. Local rank distance is then an extension that counts the sets of *similar* *n*-grams between w_1, w_2. In their NLI context, each word was actually a character *n*-gram. This method has the advantage that no syntactic information is needed; no parsing or even sentence or word segmentation is required. In training, they found that $n \in [5, 8]$ gave the best results, allowing them to take third place overall in the closed task.

Similar to the string kernels previously, Bobicev (2013) uses a method which requires very little text processing: prediction by partial matching, a statistical compression method. The output of prediction by partial matching is a language

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Table 3. *Increasing state-of-the-art NLI accuracy results for the ETS dataset*

Paper	Method	Accuracy
(Jarvis *et al.* 2013)	SVM on lexemes, lemmas, POS	83.6 per cent
(Bykh and Meurers 2014)	ensemble many lexical/syntactic	84.8 per cent
(Ionescu *et al.* 2014)	character string kernels	85.3 per cent

model which can be used to find the highest likelihood L1 given some unknown text. Both words and characters were used as tokens, though character features performed much worse than words, obtaining a precision of 37 per cent (compared to 70 per cent from the word tokens) on the 9,900-document training set. This performance placed the prediction by partial matching features in the bottom third of contestants.

In conclusion, most methods used by teams in the NLI shared tasks were very standard (e.g. *n*-grams of words with SVM). In fact, the winning team (Jarvis, Bestgen and Pepper 2013) with 83.6 per cent accuracy used SVM with unigram to trigrams of log-entropy weighted lexemes, lemmas, and POS tags. A small number of teams investigated some unique text representations and classification techniques. Given options between feature-based and likelihood-based strategies, most teams focused on features. While not always successful, these new methods can be further explored and analyzed in future work. Table 3 summarizes the results discussed on the ETS dataset. After the contest ended, others have surpassed the highest-scoring Jarvis *et al.* team.

First, Bykh and Meurers (2014) considered an ensemble of various features, consisting of context-free grammar rules and forty *n*-gram representations of words, POS tags, and lemmas for $n \in [1, 10]$. They found training a separate classifier for each feature type gave definite advantages over concatenating feature vectors.

Second, Ionescu, Popescu and Cahill (2014) improved over Bykh and Meurers (2014) by using only character-based features. However, the way that the character features are used is advanced. They combined various string kernels with two learning methods, kernel ridge regression and kernel discriminant analysis. We suggest that the reader consult their paper for a more in-depth description of these tools. Their best results were when string lengths were in the range $[5, 8]$. Seven of their kernel and learning combinations outperformed the first place shared task team (Jarvis *et al.* 2013), and their best two combinations best the Bykh and Meurers score. Their highest-scoring setup was kernel discriminant analysis on a weighted combination of a presence bits kernel (dot product of occurring characters) and the intersection string kernel.

4 Non-native grammar correction

We now transition from NLI to non-native grammar correction. While the former is often tackled as a one-step task, the latter usually relies on classification as a component, but also consists of other parts that are able to capture a deeper

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Table 4. *Comparison of grammatical error correction strategies.* * *indicates targeted approaches and* + *indicates general approaches*

Paper	Method	Target
(Lee and Seneff 2006)	LM with PCFG scoring*	articles, prepositions, and word forms
(Gamon 2010)	classifier ensemble*	articles and prepositions
(West *et al.* 2011)	bilingual random walk+	word sense
(Dahlmeier and Ng 2011a)	machine translation+	collocation errors
(Dahlmeier and Ng 2011b)	structure optimization*	articles/prepositions
(Madnani, Tetreault and Chodorow (2012)	machine translation+	common errors

syntactic meaning (such as dependencies or language modeling). Correcting machine translated text is a related issue, but we do not discuss it here; instead, please see Corston-Oliver, Gamon and Brockett (2001) or Gamon, Aue, and Smets (2005).

In the Section 4.1, we discus some key papers that display a wide variety of techniques. We also briefly touch on training data issues as well as evaluation of corrected text. Section 4.2 outlines the CoNLL shared task in GEC for 2013 and 2014.

4.1 Techniques in non-native grammar correction

Some grammar correction methods are targeted towards a very specific subset of errors, often categorized by the corpus; others attempt to solve more general errors concerning word sense or collocations. Evaluation for grammar correction is much more varied and unstandardized in comparison to the configurations from NLI seen in the previous section. Which non-native corpus is used also dictates the types of errors that can be corrected. Table 4 compares the different methods discussed in this section and categorizes them as either targeted or general approaches.

Lee and Seneff (2006) train a trigram language model on a lattice of alternatives, where 'alternatives' are prepositions, articles, and auxiliaries that may or may not occur between words in the original text. For example, the sentence *I want flight Monday* can be corrected by inserting two tokens as such: *I want a flight on Monday*. Their algorithm first strips all such alternatives from the original sentence. So far, this is not much different from the article and preposition corrections. However, they additionally change each remaining word in the input sentence to be a set of related words to the base form: $want \rightarrow \{want, wants, wanted, wanting\}$. Their language model then outputs the k-best candidates. Next, these candidates are given to a PCFG and reranked. The final output is the top-ranked sentence from the PCFG. Across all experiments, they found that reranking the language model candidates significantly increased the F measure.

West, Park and Levy (2011) use bilingual random walks between L1 and L2 word senses. For example, on one side of a bipartite graph are L1 words. There are connections from a word $w \in$ L1 to a word $w' \in$ L2 if a w could be translated into w'. w could be the English word *head*, and be translated into a physical head,

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

head of an organization, or the verb *to head*. This model was used to correct non-native sounding phrases such as *entire stranger* to the more natural *complete stranger*. This bipartite graph was combined with a language model to correct non-native sentences. In these experiments, the native language was Korean. Evaluation was performed with Amazon Mechanical Turk[13] where workers chose between the corrected sentence and the original sentence. Results were not strongly positive, since sometimes the corrected errors changed the meaning of the sentence or made it ungrammatical. In future work, the authors suggest using a richer probabilistic model such as a PCFG.

Dahlmeier and Ng (2011a) use the NUCLE corpus to find and correct collocation errors via machine translation. Here, a collocation is a phrase commonly used by native speakers. The authors propose that when a writer mentally translates from L1 to L2, some unnatural phrases result due to word choice. They give an example, '*I like to look movies*' that might be written by a native Chinese speaker since *watch* and *look* are very similar in the L1. It would be possible to correct this to the more grammatical '*I like to look at movies*', but it still does not sound natural. Instead, *look* is replaced by *watch*, resulting in the more fluent collocation *watch movies*. For their experiments, they assume the unnatural collocations have already been identified; this mimics a system where a user may ask for improvement suggestions for a snippet of writing. They train a statistical machine translation model on a parallel Chinese–English corpus to correct collocation errors in the NUCLE corpus. A log-linear model was used to score the candidate phrases which allows additional spelling, homophone, and synonym features to be incorporated. They evaluated their method as a retrieval task, where they returned the top k suggestions to fix each collocation error. Two native-English speakers judged results from 500 corrections with good rater agreement. Finally, they performed an analysis of errors and found that the main reason top-ranked phrases were not correct was due to out-of-vocabulary words.

Dahlmeier and Ng (2011b) introduce an alternating structure optimization approach to GEC. In short, alternating structure optimization is able to leverage a common structure between multiple related problems; see Ando and Zhang (2005) for a more detailed description. In this case, the related problems are selection (find features from native text) and correction (fix the errors in non-native text). Targets were article and preposition errors, again using the NUCLE corpus. It was shown that alternating structure optimization significantly outperformed a simple linear classifier as well as two unnamed commercial grammar checkers. Features included part-of-speech tags, hypernyms from WordNet, named entities, and shallow parsing tags.

Gamon (2010) combined a language model and classifier into a 'meta-classifier' that detected errors in both article and preposition use. Additionally, they investigated how much more training data is needed for the individual methods to approach the accuracy of the meta-classifier. Features for the classifier were a window of six tokens to the right and left of errors, POS tags, and lexical head features. The

[13] https://www.mturk.com/mturk/welcome

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

classifier was actually split into two steps: first, determine the likely presence of a preposition or article. Then, determine *which* article or preposition should be chosen. The language model is trained on LDC's *Gigaword* corpus and log-likelihood was used (normalized by sentence length). The meta-classifier uses features generated from the two primary models such as ratio of likelihood scores from the language model and classifier decisions. As expected, the meta-classifier outperformed the two simpler models. Prepositions were harder to classify than articles and they required more training data to reach a specific level of accuracy. Future work would be including more primary models to feed features into the meta-classifier.

Madnani, Tetreault and Chodorow (2012) applied 'round-trip' machine translation to correct generic errors in L2 text. A round-trip translation used the Google Translate API[14] to translate the candidate text from L2 to eight different languages and back again to L2. Since it is not guaranteed that the translations will preserve the meaning of the sentence, they assert that using a language model to select the most fluent choice is not acceptable. Instead, they combined alignments between the source and each round trip translation to create the final answer. This method had a better likelihood of maintaining the sentence's original meaning. In order to evaluate their system, they had human graders check whether the fixes were fluent as well as retaining the original meaning.

The next two papers we discuss consider alternatives to standard acquisition of training data for GEC. Usually, a dataset with annotated areas as described from Section 2 is used, but these papers create or acquire their own non-native errors.

Instead of training directly on non-native text, Rozovskaya and Roth (2010) took native text and intentionally inserted article errors. In order to create a realistic error distribution, they generated the article errors with the same observed distribution as the non-native corpora. Advantages to training data generation in this style are avoiding expensive data annotation, circumventing small corpus sizes, and tailoring the classifier to a particular language. Most of all, their method was shown to be superior to only training a classifier on purely native data. One final note they make is that previous baselines used in the literature are not always appropriate; baselines mentioned before are simply the majority class. While this is acceptable for a selection task, this ignores the actual error rate in the data. Usually, the error rate is much lower (shown to be about 10 per cent in their experiments) than the majority class. Therefore, they argue, a more fair comparison would be the *error reduction*, and this is indeed the measure they use to report their results. They reduced the error detection in native Chinese, Czech, and Russian text by 10 per cent, 5 per cent, and 11 per cent respectively.

The next authors (Cahill *et al.* 2013) obtained their grammatical errors from Wikipedia revisions. They filtered the article revisions dataset looking for single-word changes that corresponded to correcting article and preposition errors. Using this method, they obtained over one million corrections. They compared their mined corpus with standard error correction corpora as well as artificially-inserted errors like Rozovskaya and Roth's approach. They found that the larger 'somewhat clean'

[14] http://research.google.com/university/translate/

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Wikipedia edits were much better in increasing the F_1 of the grammar corrector as opposed to other common datasets, including a smaller 'more clean' version of their Wikipedia data. Furthermore, they found that artificial error insertion methods trained on their Wikipedia revisions data increased system accuracy compared to training on smaller corpora, and even generalized across datasets.

In sharp contrast to NLI evaluation, Chodorow *et al.* (2012) describe many issues regarding evaluation for GEC, with a focus on non-native sentences. This is a valuable paper for those engaging in any GEC task. Their main issue is the 'three-way contingency' between the original sentence, the human correction, and the system output (let alone the evaluation scripts themselves). Mentioned later (Rozovskaya and Roth 2013), this report also emphasizes the significance of the relatively low error rates in non-native sentences—this phenomenon suggests using more interpretable measures such as true and false positives and negatives. Furthermore, how is the severity of an error taken into account? For example, most native English speakers often misuse *who* and *whom*. Their conclusion from these observations is to develop a robust evaluation system based on raw error type counts. This allows bias and error skewness to be perceived by the reader while simultaneously permitting the reader or other evaluator to map the raw error data into another form such as precision, accuracy, prevalence, or bias.

4.2 Shared task in non-native grammar correction

We continue our discussion on GEC with the introduction of the CoNLL-2013 shared task (Ng *et al.* 2013) and the CoNLL-2014 shared task (Ng *et al.* 2014). The training data was the NUCLE corpus, containing annotations categorized by five error types in 2013 along with the corrected text. In 2014, there were a total of twenty-eight error types, many of which were subcategories of the 2013 types. Some systems first classified potential errors by error type and then corrected errors while others made no such distinction. Additionally, the teams received preprocessed input: sentence segmentation, word tokenization, and part-of-speech tagging were all performed.

In all, there were 1,397 essays consisting of 57,151 total sentences. Article and determiner errors were most prevalent. Testing data was considerably smaller with only fifty essays each year. Teams were also allowed to use any external resources that were non-proprietary and publicly available. Table 5 shows the methods of the top teams from both years. The five error types from 2013 are listed below:

- **Article or determiner**: '*In late nineteenth century, there was a severe air crash happening at Miami international airport*.' Correction: replace *late* with *the late*.
- **Preposition**: '*Also tracking people is very dangerous if it has been controlled by bad men in a not good purpose*. Correction: replace *in* with *for*.
- **Noun number**: '*I think such powerful device shall not be made easily available*.' Correction: replace *device* with *devices*.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Table 5. *Top teams from the CoNLL-2013 and CoNLL-2014 GEC shared task*

Team	P	R	$F_{\{1,0.5\}}$
UIUC (Rozovskaya *et al.* 2013) averaged perceptron and Naïve Bayes	46.45	23.49	31.20_{F1}
NTHU (Kao *et al.* 2013) *n*-gram and dependency language model	23.80	26.35	25.01_{F1}
HIT (Xiang *et al.* 2013) MaxEnt and rules	35.65	16.56	22.61_{F1}
CUUI (Rozovskaya *et al.* 2014) error interaction via joint inference	52.4	29.89	$45.57_{F0.5}$
CAMB (Felice *et al.* 2014) pipeline and error type filtering	46.70	34.30	$43.55_{F0.5}$

- **Verb form**: '*However, it is an achievement as it is an indication that our society is progressed well and people are living in better conditions.*' Correction: replace *progressed* with *progressing*.
- **Subject-verb agreement**: '*People still prefers to bear the risk and allow their pets to have maximum freedom.*' Correction: replace *prefers* with *prefer*.

While the contest setup and evaluation does not completely mimic real-life scenarios, it does provide useful information for those interested in GEC. Since the errors are categorized into groups, it is possible to perform an error analysis on the results.

As mentioned previously, even performing the evaluation is not a trivial task; suggested corrections—while completely valid—may not actually match the 'official' corrections. To mitigate these issues, the contest organizers permitted the contestants to submit their own gold standard corrections if they disagreed with the provided answers. Of course, the corrections were reviewed. It is also worth noting that the training and test data contained errors other than those in the five categories, but these additional errors were not taken into consideration in the final scoring.

The 2013 third-place team (Xiang *et al.* 2013) from Harbin Institute of Technology, split their system into two parts: a machine learning module (article *vs* determiner, prepositions, and noun errors) and a rule-based module (verb form and subject-verb agreement). As indicated in Table 5, the classifier used was a Maximum Entropy classifier with additional feature selection via genetic algorithms and confidence tuning based on the maximum entropy score. The rules were first preprocessed with the frequent pattern mining algorithm FP-growth to find common incorrect phrases. These common phrases were then removed from the candidate set to be corrected. A list of these common phrases was provided and may prove useful for future work. In their results, they found that their models were very sensitive to the parameters and that the confidence tuning had the largest positive effect in their pipeline.

National Tsing Hua University (Kao *et al.* 2013) finished in second place in 2013 with a system that was very similar to theirs from the previous year. In the CoNLL-2013 shared task, subject-verb agreement was not a labeled error type so the National

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Tsing Hua University system ignored these errors. In the seemingly current trend, a module was created for each error type, though their modules were more similar to each other than the previous systems. Each module used a moving window of words up to length five—five here because they leverage the Google 5-gram corpus as well as their previous tool Linggle.[15] Linggle is a linguistic search engine in which the user can specify wildcards for nouns, verbs, and adjectives (Boisson *et al.* 2013). Results were returned with those wildcards filled in ordered by increasing likelihood in the dataset. These likelihoods were used to score candidates generated by their system in each module. Their candidate voting system was created using a backoff model of lower-order *n*-grams. Verb-form errors were really the only significantly different module by including pointwise mutual information.

The University of Illinois at Urbana-Champaign (Rozovskaya *et al.* 2013) had the best-performing system in 2013. They made use of previous work for article errors (with averaged perceptron) and preposition errors (with Naïve Bayes). The remaining three error classes were dealt with Naïve Bayes with priors trained on the same Google corpus that National Tsing Hua University used. Although the CoNLL corpus already contained part-of-speech tags, they reparsed the dataset with their own tools in addition to running a shallow parser for feature generation. There was no pipeline in the system, meaning that corrections of each model were pooled together to create the final output sentence. In their error analysis, they found that an incorrect verb form can cause parsers to break, which greatly hindered rule-based methods. Since the errors were so sparse, they hypothesize, less-complicated machine learning algorithms would be more robust to the large amount of noise. This analysis provided much insight and would be useful for future work.

After the contest ended, University of Illinois at Urbana-Champaign (Rozovskaya and Roth 2013) returned to the CoNLL-2013 shared tasks with a joint learning and inference approach. Based on their observations from the shared task, they reasoned that individual (read: *independent*) classifiers per word do not capture interactions between errors and word choice. They accomplish this by combining individual classifiers using integer linear programming—a model which is able to jointly learn the error occurrences. They thus increased the F_1 score on the CoNLL-2013 data to 42 per cent, significantly higher than their first place score. They do note though, that F_1 may not be an appropriate measure; it is indeed more intuitive to measure the increase in correctness of the original data since it is quite likely that the grammar correction systems actually introduce errors themselves.

Cambridge University (Felice *et al.* 2014) was the second-best team in 2014. Without additional error suggestions for evaluation, they were in first place. They used a pipeline-based system which first ran rule-based corrections with a tool designed for English learners at their school. A language model selected the best candidate sentences from the rule-based edits. Then, those candidates were sent to a statistical machine translation system trained on parallel English data, which included NUCLE and FCE (see Section 2). Then, a large 5-gram language model from Microsoft was used to select candidates for the final stage. This is the main

[15] http://linggle.com

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

novelty of their system: with these final candidates they filter the error correction types to ignore unnecessary corrections. Their classifier was able to label almost 70 per cent of error types correctly on development data. They filtered out the labeled types that had zero precision, which consisted of reordering errors, run-ons, comma splices, and acronyms. Although the filtering did increase their $F_{0.5}$ score, the difference was not statistically significant. This is an example of how evaluation metrics play an integral part in system design.

Columbia University and the University of Illinois at Urbana-Champaign (Rozovskaya *et al.* 2014) was the top team in 2014. They improved upon their system from the previous year, and also included their joint inference work (Rozovskaya and Roth 2013). The last addition was the inclusion of new classifiers for the different error types; in total they had sixteen classifiers targeting the twenty-eight error types. The classifier learning algorithm for each error type was one of Naïve Bayes, Naïve Bayes with priors, and averaged perceptron.

In conclusion, we saw a variety of techniques for correcting grammatical errors, which can be categorized into targeted *vs* general strategies. Targeted strategies focus on errors of only specific types (such as the shared task) which general correctors try to improve the overall fluency of the L2 text.

5 Conclusions

We realize that support for dealing with non-native text is becoming increasingly important; many applications benefit from automatic author profiling and GEC. We call this collection of challenges and tasks *non-native text analysis*. This survey detailed common datasets and covered the two main applications in non-native text analysis: NLI and GEC.

In Section 2, we saw the main corpora used by researchers for NLI and GEC. Compared to other text datasets, these are relatively small and only consist of essays written by students. A larger corpus would be beneficial, as would a corpus of text other than essays such as blogs, social media, or research papers. Finally, the availability of such datasets are directly proportional to the amount of impact they may have; freely available corpora are much more likely to be used than their proprietary counterparts.

Section 3 detailed NLI as a classification problem. We also saw how a shared task was able to elicit many quality works from the community in an organized fashion. We subdivided NLI approaches into two techniques: feature- and likelihood-based. Much more effort has been put into feature-based methods using standard machine learning algorithms. Instead of classifiers and probabilistic models, is it possible to use other text mining algorithms? For example, can clustering methods be optimized for use with NLI? What is the significance of outliers and dense subclusters? How do collaborative filtering systems change when additional information about learning styles of users are known?

GEC was discussed in Section 4. We saw how most approaches used a targeted strategy: first classify errors from a limited list and then attempt to fix them. In future work, we would like to see more general approaches applied that are not restricted

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

to specific error types. Aside from grammaticality, fluency can also be considered: exactly what phrases or syntactic structures make an essay sound 'native'? Are specific topics handled with different grammar? Or are some differences inherently cultural? How are ethnicity, culture, and race related in our context? Just because one's text may seem like native English, does that mean it is American English? If American, is it Northeastern or Southern? Dialect plays a very important role in fluency.

Finally, almost all work surveyed covered English as L2. We mentioned a few works and shared task in the introduction where English was not the target, and many techniques were general in terms of L1 and L2, but in order to prove the effectiveness of a technique, it is essential to showcase it in a direction other than L1 to English. Of course, a main issue with this is the popularity of non-native English datasets, but given other L2 corpora, an evaluation should be possible.

In time, the field *non-native text analysis* will evolve into *non-native text mining*. Aspects from this area discussed will be combined into unified algorithms and intelligent applications. For example, an NLI system can determine a user's L1, and use that knowledge to offer sophisticated annotations and corrections to their own text in L2 while simultaneously summarizing and condensing text from other sources. Indeed, other areas aside from these will arise to further the understanding of non-native writing or to help non-native speakers better understand their second or third language.

References

Ando, R., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* **6**: 1817–53.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. 2013. TOEF11: a corpus of non-native English. Technical Report, Educational Testing Service (ETS). (https://www.ets.org/)

Blei, D., Ng, A., and Jordan, M. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Blunsom, P., and Cohn, T. 2010. Unsupervised induction of tree substitution grammars for dependency parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, Massachusetts, USA, pp. 1204–13.

Bobicev, V. 2013. Native language identification with PPM. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 180–87.

Boisson, J., Kao, T., Wu, J., Yen, T., and Chang, J. 2013. Linggle: a web-scale linguistic search engine for words in context. In *Proceedings of Association for Computational Linguistics (Conference System Demonstrations)*, Sofia, Bulgaria, pp. 139–44.

British Council 2014. How Many People Speak English? http://www.britishcouncil.org/learning-faq-the-english-language.htm

Brockett, C., Dolan, W., and Gamon, M. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp. 249–56.

Brooke, J., and Hirst, G. 2011. Native language detection with 'Cheap' learner corpora. In *Proceedings of the 2011 Conference on Learner Corpus Research*, Louvain-la-Neuve, Belgium, pp. 37–47.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Brooke, J., and Hirst, G. 2012. Robust, lexicalized native language identification. In *Proceedings of the International Conference on Computational Linguistics*, Mumbai, India, pp. 391–408.

Bykh, S., and Meurers, D. 2014. Exploring syntactic features for native language identification: a variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, pp. 1962–73.

Cahill, A., Madnani, N., Tetreault, J., and Napolitano, D. 2013. Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, pp. 507–17.

Chodorow, M., Dickinson, M., Israel, R., and Tetreault, J. 2012. Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012*, Mumbai, India, pp. 611–28.

Corston-Oliver, S., Gamon, M., and Brockett, C. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France, pp. 148–55.

Dahlmeier, D., and Ng, H. 2011a. Correcting Semantic Collocation Errors with L1-induced Paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, pp. 107–17.

Dahlmeier, D., and Ng, H. 2011b. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon, USA, pp. 915–23

Dahlmeier, D., Ng, H., and Wu, S. 2013. Building a large annotated corpus of learner english: the NUS corpus of learner english. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 22–31.

Felice, M., Yuan, Z., Andersen, Ø., Yannakoudakis, H., and Kochmar, E. 2014. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, Baltimore, Maryland, USA, pp. 15–24.

Gamon, M. 2010. Using mostly native data to correct errors in Learners' writing: a meta-classifier approach. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, USA, pp. 163–71.

Gamon, M., Aue, A., and Smets, M. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In *Proceedings of the European Association for Machine Translation (EAMT)*, Budapest, Hungary, pp. 103–11.

Granger, S. 2003. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. In *Teachers of English to Speakers of Other Languages Quarterly*, pp. 538–46.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. 2009. *The International Learner Corpus of English*, Version 2. Presses Universitaires de Louvain.

Gui, S., and H. Yang, H. 2003. Zhongguo Xuexizhe Yingyu Yuliaohu (Chinese Learner English Corpus). In *Shanghai Waiyu Jiaoyu Chubanshe*.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, pp. 101–11.

Ionescu, R., Popescu, M., and Cahill, A. 2014. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1363–73.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Ishikawa, S. 2009. Vocabulary in interlanguage: a study on corpus of english essays written by Asian university students (CEEAUS). In *Phraseology: Corpus Linguistics and Lexicology*, pp. 87–100.

Ishikawa, S. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of english. In *Learner Corpus Studies in Asia and the World*, pp. 91–118.

Jarvis, S., Bestgen, Y., and Pepper, S. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 111–18.

Johnson, M., Griffiths, T., and Goldwater, S. 2006. Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models. In *Neural Information Processing Systems*, pp. 641–8.

Kao, T., Chang, Y., Chiu, H., Yen, T., Boisson, J., Wu, J., and Chang, J. 2013. CoNLL-2013 shared task: grammatical error correction NTHU system description. In *Proceedings of the 17th Conference on Computational Natural Language Learning: Shared Task*, Sofia, Bulgaria, pp. 20–25.

Koppel, M., Schler, J., and Argamon, S. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 9–26.

Koppel, M., Schler, J., and Zigdon, K. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, pp. 624–8.

Kukich, K. 1992. Techniques for automatically correcting words in text. In *ACM Computing Surveys*, pp. 377–439.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. R. 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* **20**(6): 33:1–33:31. ISSN 1073-0516.

Kyle, K., Crossley, S., Dai, J., and McNamara, D. 2013. Native language identification: a key N-gram category approach. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 242–50.

Lavergne, T., Illouz, G., Max, A., and Nagata, R. 2013. LIMSI's participation to the 2013 shared task on Native Language Identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 260–65.

Leacock, C., Chodorow, M., Gamon, M., Tetreault, J. 2014. Automated Grammatical Error Detection for Language Learners, 2nd ed. In G. Hirst (ed.), *Morgan and Claypool (Synthesis lectures on human language technologies)*.

Lee, J., and Seneff, S. 2006. Automatic grammar correction for second-language learners. In *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, pp. 1978–81.

Madnani, N., Tetreault, J., and Chodorow, M. 2012. Exploring grammatical error correction with not-so-crummy machine translation. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*, Montreal, Canada, pp. 44–53.

Massung, S., Zhai, C., and Hockenmaier, J. 2013. Structural parse tree features for text representation. In *Proceedings of the International Conference on Semantic Computing*, Irvine, California, USA, pp. 9–16.

Ng, H., Wu, S., Briscoe, T., Hadiwinoto, C., Sustano, R., and Bryant, C. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, Ann Arbor, Michigan, USA, pp. 1–14.

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

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Popescu, M., and Ionescu, T. 2013. The story of the characters, the DNA and the native language. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 270–78.

Rangel, F., Rosso, F., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., and Daelemans, W. 2014. Overview of the 2nd author profiling task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, Sheffield, England, UK.

Rozovskaya, A., and Roth, D. 2010. Training paradigms for correcting errors in grammar and usage. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, pp. 154–62.

Rozovskaya, A., and Roth, D. 2013. Joint learning and inference for grammatical error correction. In *Proceedings of Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp. 791–802.

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

Rozovskaya, A., Chang, K., Sammons, M., Roth, D., and Habash, N. 2014. The Illinois-Columbia system in the CoNLL-2014 shared task. In *Proceedings of the 18th Conference on Computational Natural Language Learning: Shared Task*, Ann Arbor, Michigan, USA, 34–42.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* **60**: 538–56.

Stolerman, A., Caliskan, A., and Greenstadt, R. 2013. From language to family and back: native language and language family identification from english text. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, Atlanta, Georgia, USA, pp. 32–9.

Swanson, B., and Charniak, E. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, Jeju Island, Korea, pp. 193–97.

Tajiri, T., Komachi, M., and Matsumoto, Y. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea, pp. 198–202.

Tetreault, J., Blanchard, D., and Cahill, A. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 48–57.

Tsur, O., and Rappoport, A. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, Prague, Czech Republic, pp. 9–16.

Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., and Dyer, C. 2013. Identifying the L1 of non-native writers: the CMU-Haifa System. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, USA, pp. 279–87.

West, R., Park, A., and Levy, R. 2011. Bilingual random walk models for automated grammar correction of ESL author-produced text. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, Portland, Oregon, USA, pp. 170–79.

Wong, J., and Dras, M. 2009. Contrastive analysis and native language identification. In *Australasian Language Technology Association Workshop 2009*, pp. 53–61.

Wong, J., and Dras, M. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, pp. 1600–10.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press

Wong, J., Dras, M., and Johnson, M. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, pp. 699–709.

Xiang, Y., Yuan, B., Zhang, Y., Wang, X., Zheng, W., and Wei, C. 2013. A hybrid model for grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, Sofia, Bulgaria, pp. 115–22.

Yannakoudakis, H., Briscoe, T., and Medlock, B. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 180–89.

Yu, L., Lee, L., and Chang, L. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. In *Proceedings of the 22nd International Conference on Computers in Education*, Nara, Japan, pp. 42–7.

Zheng, Z., Wu, X., and Srihari, R. 2004. Feature Selection for Text Categorization on Imbalanced Data. *SIGKDD Explor. Newsl.* **6(**1): 80–89. ISSN1931-0145.

https://doi.org/10.1017/S1351324915000303 Published online by Cambridge University Press