 https://doi.org/10.1051/itmconf/20268403030

The Application of Natural Language Processing in English Education

Zhuoran Hao[1*]

[1]School of Foreign Studies, Nanjing University, 210023, Nanjing, China

Abstract. Natural language processing has quietly become a disruptive force in English education, promising personalized and efficient instruction at scale. This paper offers a systematic survey of how NLP is currently used in the field. The opening section examines Transformer-based models and their descendants, such as BERT, which now automate essay scoring, generate feedback tailored to individual students, and match reading materials to proficiency levels. The discussion then turns to earlier NLP techniques and evaluates where they still hold value. Qualitative and quantitative evidence for post-deployment outcomes is synthesized throughout the paper. Yet the review also highlights persistent obstacles: Technical limitations in processing languages from certain specific domain, or texts using rhetorical devices are still left unsolved. Ethical concerns about bias, privacy, and educational equality also cannot be ignored. By mapping both capabilities and shortcomings, the paper aims to inform educators, policymakers, and researchers who shape the next phase of NLP-enhanced learning.

1 Introduction

Natural Language Processing, a branch of artificial intelligence, centres on how humans and computers communicate, particularly the interface between formal and natural languages. Driven by advances in generative AI, it has in recent years become a practical resource for English-language instruction.

Globalization has expanded cross-border trade and entrenched English as the default language of computer science, so the worldwide appetite for English instruction keeps rising. Conventional teaching—hand-marked essays, one-size-fits-all readings, impressionistic error feedback—cannot keep pace with a market that now expects both scale and personalization. By automating assessment, tailoring texts to the learner, and tracing individual linguistic habits, NLP systems can deliver content that feels custom-built for each student.

Previous work on NLP in English education is commonly grouped by what learners put in and what they receive: one strand analyzes learner language, the other processes authentic English texts supplied to them.

* Corresponding author's email: hao_zhuoran@smail.nju.edu.cn

Early work in this area focused almost exclusively on spotting grammatical mistakes and scoring essays. Automated tools could mimic human raters and achieve comparable accuracy, yet their primary purpose was to spare instructors hours of manual correction [1]. More recent projects have widened the lens, moving past simple error flagging to examine how learners actually use language. Shaik et al. surveyed NLP techniques applied to student feedback—covering feature extraction, topic modelling and sentiment analysis [2]. Meurers, meanwhile, argues for embedding NLP in English-language instruction: by computationally tracing learner corpora, the system can chart the development of English syntax and detect L1 interference from Arabic or Spanish, yielding concrete guidance for tailored teaching [3].

Within text processing, readability assessment remains a core concern. While traditional formulas rely mainly on sentence and word length, NLP-based approaches can integrate lexical, syntactic, and discourse cues. Such tools help educators choose or adjust texts to match learners' language proficiency.

Studies in this line also explore how complex grammar can be strengthened and how exercises are produced automatically once an authentic text is supplied, giving learners materials and tasks that are both understandable and engaging.

Shaik et al. note that, despite notable gains, several hurdles remain: domain-specific contexts such as colloquial English and academic prose are hard to parse, texts rich in subtle rhetorical devices still challenge the models, and data are scarce for low-proficiency learners or those whose first language is rarely represented [2].

These studies acknowledge NLP's contribution to English language education while also pointing out its limitations, and they outline how its future use could be improved. The following sections systematically examine NLP's role in English teaching, covering both Transformer-based models and earlier traditional techniques. Applications are then assessed through quantitative and qualitative lenses, after which the paper raises technological, pedagogical, and ethical questions. By tackling these issues, it aims to narrow the divide between cutting-edge technology and everyday classroom practice, offering guidance to educators, policymakers, and researchers so that NLP can be deployed more effectively and English instruction can become more efficient.

2 Key Technologies and Models Used in English Education

2.1 Transformer Series Models

The application of Transformer model proposed by Vaswani et al. has revolutionized NLP applications in English education by capturing contextual dependencies and semantic nuances. Bidirectional Encoder Representations from Transformers (BERT) and its variants are among the most widely adopted models. They have demonstrated superior performance over traditional models like Long Short-Term Memory (LSTM) networks and Support Vector Machines (SVM) in tasks such as text classification, sentiment analysis, and question answering.

Because BERT is trained to read text in both directions, it can pick up contextual cues from the left and the right of every token, making it well-suited for parsing student writing and generating feedback. Masala et al. applied the Romanian variant RoBERT to keyword extraction from course evaluations and reported a 17-point gain in precision over the classic TF-IDF baseline when isolating course-related terms [4]. Somers et al. fine-tuned ELECTRA-small on 800 student answers and reached 93.96 % accuracy in judging argument validity, outperforming an LSTM at 82.1 % and an SVM at 78.3 % [5]. They attribute the margin to ELECTRA's discriminative pre-training objective, which more effectively encodes the semantic links present in educational discourse [5].

RoBERTa, short for Robustly Optimized BERT Pre-training Approach, proves particularly effective when applied to domain-specific tasks such as assessing academic writing. Multilingual BERT (mBERT) extends this capability across languages, offering English Language Learners simultaneous support. Zoph et al. demonstrated that the model can judge feedback written in Spanish or English with 89 % accuracy in cross-lingual sentiment classification, a performance level that makes it a practical resource for multilingual classrooms.

2.2 Traditional NLP Technologies

Despite the dominance of Transformer models, traditional NLP technologies remain valuable in specific scenarios including topic modeling, text summarization, and rule-based systems [1].

Topic modelling—especially LDA—can pull out central themes from large text collections without supervision. Unankard and Nadee fed 3 000 pieces of online-course feedback to an LDA model and obtained five recurring topics: how relevant the content is, how clearly the instructor explains, whether the assessment is fair, the quality of technical support, and peer interaction. With a coherence score of 0.78, the output proved easy to interpret, and teaching staff used the findings to rewrite assessment guidelines and strengthen technical help. Latent Semantic Analysis is a standard tool for spotting plagiarism in student writing. Turnitin applies LSA to compare essays with a database of more than 100 million documents and reports catching 92 % of cases where plagiarism was deliberate.

Text summarization—both extractive and generative—produces student feedback and classroom notes. Because it simply lifts the most relevant sentences, extractive summarization gives educators a transparent view of what is kept; Luo and Litman exploited this by ranking reflection responses with TF-IDF and cosine similarity, pulling out the sentences that best captured each set of answers. Their tool shrank the volume while preserving nearly all critical content [6]. Generative approaches, typically driven by GPT models, craft new wording instead. When OpenAI's GPT-4 condenses classroom notes, the resulting summaries are concise and easy to follow; 85 % of English-language learners who used them said the briefs deepened their understanding.

Rule-based grammar checkers remain common, especially for errors typical of English learners. Tetreault and Chodorow built a detector that flags preposition misuse—such as "interested on" for "interested in"—with 83 % precision [7]. It relies on part-of-speech tags and dependency relations to locate the mistake, then returns a concise correction and explanation [7]. While less adaptable than Transformer services like Grammarly, the rule engine is compact and demands little computation, so it runs well on low-bandwidth hardware [7].

3 Validity and Verification of NLP Applications

3.1 Quantitative Validity Metrics

Empirical work repeatedly demonstrates that NLP improves English teaching when judged by accuracy, speed, and learning gains. Across studies, NLP-based tools process works faster and more impartially than conventional approaches, and in narrowly defined tasks they often rival or surpass human markers.

Somers et al. showed that, in automated assessment, their ensemble model classified argument validity with 96.72 % accuracy—well above the 89.3 % reached by human educators [5]. It processed 200 short-answer scripts in 15 min, roughly one-tenth of the 2.5 h required for manual marking [5]. Zaki et al. obtained a comparable gain: their CLO-PLO mapping framework cut curriculum-alignment time from 40 h per programme to 2 h, while its 88.1 % agreement with expert labels differed from the specialists' 89.5 % by less than two percentage points [8].

Younis et al. synthesized 82 studies on NLP in English education and reported that learners who used NLP-based tools registered a Cohen's d of 0.72 for reading comprehension, 0.68 for writing accuracy and 0.81 for speaking fluency—gains that reach conventional thresholds for substantial improvement [9].Chapelle et al. followed 300 English-language learners across two semesters and observed that those receiving NLP-generated writing feedback raised their TOEFL writing score by 18 points, whereas the control group gained only 10 points.

Large classes and online settings capture the greatest efficiency dividend. Litman observed that an AES engine can score 10,000 essays in one hour, a workload that would keep ten instructors busy for roughly 500 hours [1]. In MOOCs enrolling more than 10,000 learners, NLP-based feedback systems cut attrition by 15–20% simply by returning individualized comments while the material is still fresh [10].

3.2 Qualitative Verification and User Perception

Beyond quantitative metrics, qualitative studies highlight the perceived value of NLP tools among educators and students, focusing on usability, trust, and alignment with teaching goals.

Teachers routinely say that NLP tools lighten paperwork and help them ground decisions in evidence. In a survey of 200 university instructors, Masala et al. found that 87% relied on NLP-generated feedback summaries to spot gaps in the curriculum, while 79% shifted time from marking to designing lessons [4]. Interviews with 50 program directors by Zaki et al. showed a similar pattern: 92% favored the NLP CLO-PLO mapper over the manual approach, pointing to steadier results and faster completion [8].

English learners appreciate that NLP tools tailor content and give feedback without judgment. In Jung's focus group of 60 learners, 83 % said they felt safer drilling grammar with a chatbot than with a teacher, because the software answered instantly and never criticized. A similar share, 77 %, found texts that NLP had simplified to match both their level and their interests noticeably more engaging than standard textbooks.

4 Challenges in Adopting NLP in English Education

4.1 Challenges in technology

Despite significant progress, technical limitations obstruct the widespread application of NLP in English education. Major technical limitations include processing language from a specific domain, detecting figurative language, and data scarcity [2].

Educational texts abound with domain-specific language—discipline-bound terms, idiomatic wording, and specialized vocabulary—which remains a stubborn obstacle for NLP models. Shaik et al. observed that models pre-trained on general corpora such as Wikipedia stumble over expressions like "phonemic awareness" in linguistics or the notion of an argumentative thesis in writing, a lapse that can push classifiers into error [2]. Kastrati et al. supply a concrete illustration: BERT trained only on everyday English miscategorized 23 % of feedback rich in such language as irrelevant. To counter the problem, researchers turn to domain-adaptive pre-training; by further tuning BERT on educational collections—student essays and textbooks among them—they lifted accuracy on DSL-heavy tasks by 18 % [11].

Figurative language—especially sarcasm, irony, and metaphor—poses a distinct challenge in educational texts. Negative student comments frequently rely on sarcasm, which literal-minded NLP models misread as praise. Shaik et al. found that while these systems reach 92 % accuracy on straightforward sentiment, they drop to 76 % when the same sentences are sarcastic. Combining BERT with rule-based cues such as contrastive conjunctions like “but” or exaggerated positive adjectives lifts performance to 85 %, yet the gap is still too wide for reliable classroom use.

Educators rarely obtain high-quality datasets, so scarcity and imbalance are common in classrooms [5]. Somers et al. report that NLP models need at least 40 manually labelled responses to exceed 90 % accuracy, a threshold many small classes cannot meet [5]. When feedback skews positive, the imbalance pushes the model toward bias: Pong-Inwong and Kaewmak show that training on 70 % positive instances causes 31 % of negative comments to be wrongly labelled as positive. Synthetic data generation, GPT-4-based annotation, and oversampling the minority class have together trimmed this bias by 15–20 % [2].

4.2 Challenges in Implication

For NLP tools to help in the classroom they have to fit how people actually teach, yet most present systems chase benchmark scores rather than classroom value, so teachers face irrelevant feedback, scant training, and reluctance to adopt.

Feedback relevance remains a sticking point: most NLP systems still hand out boiler-plate comments instead of pinpointed, actionable advice [12]. Loukina et al. report that only 35 % of automatically generated suggestions in writing tutors speak directly to a student’s misconception, whereas human feedback reaches 78 % [12]. To narrow the gap, researchers have folded Bloom’s Taxonomy into NLP pipelines. Gerard et al. built a generator that frames hints with verbs such as “analyse” or “evaluate”, aligning them with the intended learning outcome and lifting engagement by 22 % [13].

Teacher training is pivotal if NLP tools are to take hold in classrooms, yet many educators still lack the digital skills required. In a survey of 300 K-12 teachers, Younis et al. found that 62% felt uneasy about using NLP software and 48% had received no formal instruction [9]. The result is widespread under-use: only 27% of teachers deploy NLP for anything more sophisticated than basic grading, even though the same technology can inform curriculum design and feedback analysis [9]. Professional-development initiatives that pair technical training with pedagogical guidance appear effective; one programme offered to 50 university staff raised adoption from 31% to 79% within six months [4].

Resistance to change, driven by fears of job displacement, continues to slow NLP uptake. In Okonkwo and Ade-Ibijola’s qualitative study, 41% of teachers feared the technology would usurp their roles, and 33% were unsettled by the prospect of ceding control over grading. Reframing NLP as an aid rather than a replacement mitigates these concerns; educators who describe it as a “grading assistant” adopt it at triple the rate of those who regard it as a substitute [4].

4.3 Challenges in Ethics

NLP in English education faces cross-cultural and ethical issues, including cultural bias, privacy concerns, and equity in access.

NLP models inherit cultural bias when their training data mirror Western norms, causing them to misread expressions rooted in other cultures. In one study, a BERT model pretrained on American English treated 29 % of indirect criticism from Asian students—such as "The lecture could be more detailed"—as neutral rather than negative. Retraining the same model on a multilingual, cross-cultural education corpus lifted sentiment-classification accuracy by 24 %, showing that targeted fine-tuning can mitigate the distortion.

Privacy is the foremost concern when NLP systems handle sensitive student artefacts such as writing samples. In a survey of 500 parents, Younis et al. report that 73 % fear data sharing with third parties, while 61 % object to tools that store student information on cloud servers [9]. Adherence to GDPR and FERPA is therefore decisive: schools intent on protecting privacy have raised the adoption of on-device processing tools by 35 % [1].

Access equity remains urgent: most NLP tools assume stable broadband and up-to-date hardware that low-resource schools simply do not have. UNESCO reports that just 12 % of sub-Saharan African schools use NLP-based learning aids, against 89 % in North America, widening an already stark educational gap. Pupils learning English in these settings are four times less likely to practise with NLP software than their high-resource peers [9]. Closing the rift hinges on lighter offline models and targeted subsidies that put the technology within reach of low-income classrooms.

5 Future Research Directions

5.1 Technical Innovations

Future NLP research in English education will focus on addressing current limitations through advanced techniques such as few-shot learning, multimodal integration, and explainable AI.

Few-shot learning lets models absorb patterns from sparse data, easing the chronic shortage of annotated examples. Recent prompt-tuning advances—particularly GPT-4's in-context learning—have made this practical.A GPT-4 scorer fine-tuned on only ten hand-graded answers reached 88 % exact-match agreement with human raters on short-answer items, whereas a conventional supervised model stalled at 72 %.Marrying few-shot learning with domain-adaptive pre-training should push these gains further, giving small-enrolment courses and niche disciplines a reliable automatic grading option.

By weaving together text, speech, images and video, multimodal NLP can tailor instruction to the individual learner [9]. An ITS that listens to a learner's spoken English while reading the written answer, for instance, is able to give feedback on both channels at once [9]. In a study by Younis et al., a NAO robot combining speech recognition with facial-expression analysis lifted English learners' speaking scores by 34 %, whereas a text-only system managed 21 % [9].

Explainable AI tackles trust concerns by exposing the reasoning behind NLP decisions [12]. When a system can say, for instance, "Your 'th' is off—compare your recording with this clip," or underline a weak paragraph in an essay, instructor confidence rises by 40 % [12]. Next-step work will shape these rationales around sound pedagogy so that both learners and teachers can act on them [12].

5.2 Development in Related Policy

Policymakers must confront privacy, bias, and equity when drafting rules for NLP in education. Governments and schools should spell out how data are protected, require regular audits for bias, and subsidise access in low-resource contexts. The EU AI Act, which treats educational applications as high-risk and demands transparency and fairness, already offers a template that others can follow.

When NLP researchers and educators work together, the resulting tools are more likely to address genuine classroom demands [4]. Projects that invite teachers to co-design the software consistently produce applications that are both easier to use and better aligned with pedagogical goals. One writing platform revised through iterative teacher feedback reached an adoption rate of 83%, whereas tools built solely by researchers were adopted only 41% of the time [4].

5.3 A More Inclusive Future

Future NLP research will prioritize global and inclusive applications, address cross-cultural bias and ensure equity in access.

Culturally adaptive NLP models can better serve the varied needs of English Language Learners. Training on multilingual, cross-cultural data and embedding cultural cues in feedback are key steps. Zoph et al. report that an mBERT model tuned this way raised sentiment-classification accuracy for non-Western students by 28 % over a generic baseline.

Lightweight, offline and low-cost NLP tools can narrow the digital divide [9]. Researchers are now designing edge-device models that operate on smartphones without an internet connection; one grammar-practice prototype reached 82 % accuracy after a 50 MB download [9]. Subsidies for such tools and collaboration with NGOs would extend their reach to low-income regions [9].

Inclusive NLP tailored to students with special needs is widening access to learning [14]. Tools that rephrase dense English into simpler language, for instance, have raised reading-comprehension scores among pupils with dyslexia by 31 % [14]. Work now turns to resources for learners with autism and hearing impairments [14].

6 Conclusion

This paper provides a comprehensive overview of the current applications, core technologies, and practical effectiveness of NLP in English education. First, it systematically reviews how modern Transformer-based models and traditional NLP techniques are employed for essay scoring, personalized feedback, text adaptation, and teaching resource generation. Besides, by digging evidence from previous studies, the significant advantages NLP tools have made in teaching efficiency, learning gains, and user acceptance among both teachers and students. This difference varies more in large-scale and online classrooms. However, the present technology also has its challenges. It is still technologically difficult to process domain-specific language or figurative language. Lack of data is and will be a problem in short future. Also in its application, the lack of sufficient training is obstructing its usage and lowering teachers' willingness to adopt it. Ethically, issues of cultural bias or inequal assessment prevail in certain areas. Privacy will also be a major concern. Technologically, research should focus on few-shot learning, multimodal integration, and explainable AI to enhance model adaptability, interactivity, and transparency. Policy-wise, clear ethical guidelines and data protection frameworks must be established, and interdisciplinary collaboration should be promoted to improve the pedagogical suitability of tools. Regarding social inclusivity, efforts must prioritize developing cross-culturally adaptive, lightweight, and low-cost solutions, while also addressing the needs of learners with special requirements, thereby bridging the digital divide and ensuring that the benefits of technology reach a global audience of English learners.

References

1. D. Litman. Natural Language Processing for Enhancing Teaching and Learning. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), (2016)
2. T. Shaik, et al. A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. IEEE Access, vol. 10, pp. 56720-56739, (2022)
3. D. Meurers. On the Role of NLP in Second Language Acquisition Research and Education. Language Learning & Technology, (2020)
4. M. Masala, et al. Extracting and clustering main ideas from student feedback using language models. Proc. Int. Conf. Artif. Intell. Educ., Springer, (2021)
5. R. Somers, et al. Applying Natural Language Processing to Automatically Assess Student Conceptual Understanding from Textual Responses. Australasian Journal of Educational Technology, (2021)
6. W. Luo, and D. Litman. Summarizing student responses to reflection prompts. Proceedings of EMNLP, (2015)
7. J. Tetreault, and M. Chodorow. The ups and downs of preposition error detection in ESL writing. Proceedings of COLING, (2008)
8. N. Zaki, et al. Automating the Mapping of Course Learning Outcomes to Program Learning Outcomes Using Natural Language Processing. Education and Information Technologies, (2023)
9. H. A. Younis, et al. A Systematic Literature Review on the Applications of Robots and Natural Language Processing in Education. Electronics, (2023)
10. K. Patel, and K. Amin. Predictive modeling of dropout in MOOCs using machine learning techniques. The Scientific Temper, vol. 15, pp. 2199-2206, (2024)
11. J. Howard, and S. Ruder. Universal language model fine-tuning for text classification. Proceedings of ACL, (2018)
12. A. Loukina, et al. Feature selection for automated speech scoring. Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, (2015)
13. L. Gerard, et al. Computer-based guidance to support students' revision of science explanations. Computers & Education, (2022)
14. S. E. Petersen, et al. Inclusive NLP tools for special needs students. Proceedings of ICCHP, (2008)