IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

# Taxonomy, Benchmarks and Computational Approaches for Semantic Error Correction in Natural Language Processing

**Satvika Shrivastava, Swati Gangwani, Madhvi Soni, Jitendra Singh Thakur**

Department of Computer Science and Engineering, Jabalpur Engineering College, Jabalpur 482011, India

**Abstract:** While modern Natural Language Processing (NLP) systems excel at finding basic grammar and spelling mistakes, identifying and correcting semantic errors (problem in the meaning of the text) remains an intricate challenge. Unlike grammar faults, semantic errors often use correct spelling and rules, but they break context, disrupt logical flow, or misuse words. This paper provides a structured investigation into semantic error detection and correction. We present a novel, fine-grained taxonomy that divides semantic errors into contextual mismatches and wrong word choices. We then evaluate standard datasets to highlight their lack of semantic coverage, compare traditional rule-based methods with modern deep learning models like Transformers and analyse system performance metrics across standard evaluation dimensions, identifying critical bottlenecks and outlining strategic trajectories for future NLP architectures.

**Keywords:** Semantic Error Correction, Contextual Error Detection, Distributional Semantics, Lexical Collocation, Neural, Machine Translation, Language Modelling.

## 1.INTRODUCTION

Across the Globe, English serves as the primary medium for international communication, spoken actively by hundreds of millions of non-native learners. During the language acquisition process, these learners commit semantic errors far more frequently than syntactic ones. This disparity stems from fundamental linguistic mechanics: while syntax is governed by explicit, rule-based grammatical constraints, semantic validity relies on nuanced contextual appropriateness [1]. Consequently, semantic errors present a formidable challenge, as they disrupt both the structural flow and core meaning of an entire sentence without violating basic grammar rules.

Detecting and rectifying these errors is exceptionally difficult because standard surface-level parsing algorithms fail to catch them. Resolving semantic errors demands human-level cognitive reasoning—the ability to read between the lines, interpret discourse-level logic, and evaluate wider contextual relationships. As noted by early discourse researchers, analyzing language beyond the isolated sentence level plays a central role in linguistics, based on the principle that context offers far deeper insight into how meaning is assigned to utterances. Despite this necessity, automated syntactic and semantic analysis of an erroneous sentence remains notoriously complex, leaving semantic error detection and correction significantly less explored than its syntactic counterparts [1].

The persistence of this gap means that automated semantic error processing remains largely in its research and developmental stages. While conventional Grammatical Error Correction (GEC) tools have achieved impressive proficiency, they frequently miss subtle semantic mismatches, such as collocations, homonyms, or contextual word mis-selections. Transitioning from basic syntactic checks to comprehensive semantic reasoning represents a crucial frontier in Natural Language Processing (NLP), requiring models capable of processing discourse coherence rather than relying solely on local dictionary lookups.

To address this challenge, this paper provides a structured, multi-faceted investigation into Semantic Error Detection and Correction. We present a fine-grained taxonomy that categorizes semantic errors into contextual mismatches and word-choice errors, followed by a critical review of existing benchmark evaluation corpora and their limitations. Furthermore, we trace the evolution of computational techniques—from early rule-based and $n$-gram models to unsupervised distributional representations and modern deep neural sequence architectures—before highlighting key performance bottlenecks and proposing strategic trajectories for future research.

This paper is organized as follows: Section 2 defines the semantic errors. Section 3 presents our novel taxonomy of Semantic error classification. In section 4, various available datasets have been evaluated. Section 5 and 6 highlights the various approaches in the literature to detect and correct semantic errors. Finally, concluding remarks are given in section

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

7. This paper provides a deep and wide knowledge of existing techniques for new researchers who venture to explore detection and correction of Semantic Errors which is one of the most challenging NLP tasks.

## 2. SEMANTIC ERRORS

The term 'Semantic' relates to '**meaning in language',** representing the system of rules that governs how statements convey sense. Consequently, **semantic errors** occur when a sentence remains morphologically and syntactically sound yet becomes senseless or absurd within its context. Rather than breaking explicit grammatical rules, these errors violate the semantic constraints of the English language.

Semantic errors typically stem from a writer's imprecise understanding of vocabulary, leading them to select a word that fits the grammatical structure but fails semantically. The following example will give the correct view of these errors:

***Example 1: Semantic Mismatch***

***Incorrect:*** *The contractor was not happy with the progress at the construction sight.*

***Correct:*** *The contractor was not happy with the progress at the construction* ***site****.*

In Example 1, the token *sight* is grammatically valid as a noun following an attribute. However, its semantic representation (Sight: "faculty of vision") conflicts with the required locative context (site: "physical spatial location"), rendering the utterance semantically flawed. Substituting *sight* with *site* resolves the contextual error without altering the syntax of the sentence.

## 3. TAXONOLY OF SEMANTIC ERRORS

Semantic error typically stems from a limited proficiency in the target language. Soni et al [18] have classified Semantic errors into two broader categories: Contextual Errors and Word Choice Error. These error types can further be classified into subtypes. In this section, we present a more fine-grained taxonomy of semantic errors.

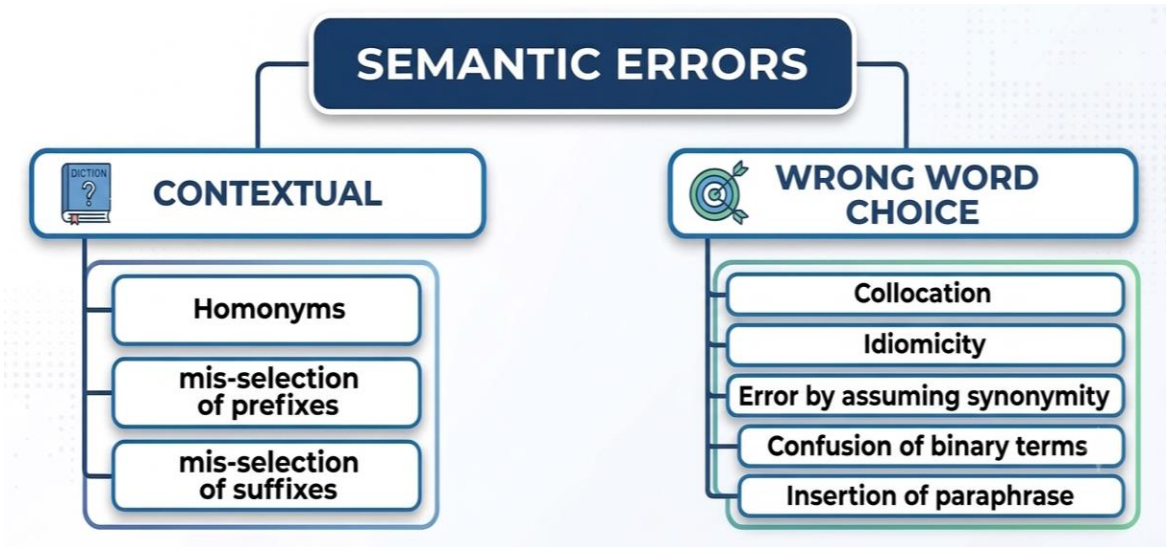


Figure 1: Classification of Semantic Errors

**1. Contextual Errors:** A contextual error occurs when a typed word is a valid real word in the language—allowing it to bypass spellcheckers—yet fails to fit the surrounding sentence context.

(a) **Homonyms:** Words that share spelling or sound but differ in meaning.

- *Example:* "The tea was **two** hot to drink." (*too*)
- *Example:* "Her mother prepared a delicious **desert**." (*dessert*)

(b) **Mis-selection of Prefixes:** Incorrect attachment of prefixes to words.

- *Example:* "This question is **disclear** for me." (*unclear*)
- *Example:* "I am **nonhappy** in my studies." (*unhappy*)

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

(c) **Mis-selection of Suffixes:** Incorrect attachment of suffixes to words.

- *Example:* "I am **interesting** in reading books." (*interested*)
- *Example:* "I am an **ambitionable** person in my life." (*ambitious*)

**2. Wrong Word Choice Errors:** A wrong word choice error occurs when an inappropriate or rare word is used in a given context, often due to limited vocabulary.

(a) **Collocation Error:** Violating natural word co-occurrence combinations.

- *Example:* "I am going to the tennis **field**." (*court*)
- *Example:* "The teacher gives us **expensive** advice." (*valuable*)

(b) **Error by Assuming Synonymity:** Treating non-interchangeable synonyms as equivalent.

- *Example:* "The teacher asked us to meet him when he is **empty**." (*free*)
- *Example:* "I will **communicate** with you through email." (*contact*)

(c) **Confusion of Binary Terms (Relational Opposites):** Misusing paired complementary terms.

- *Example:* "I have to return the books I **lent** from the library." (*borrowed*)
- *Example:* "Every day I **come** to school on foot." (*go*)

(d) **Idiomaticity:** Unnatural phrasing in set or routine expressions.

- *Example:* "I always **get up** at 6 o'clock." (*wake up*)
- *Example:* "She **heard** English music today." (*listened to*)

(e) **Insertion of Paraphrase:** Replacing a single lexical term with an awkward descriptive phrase.

- *Example:* "The women who are **carrying babies** should stay at home." (*pregnant*)
- *Example:* "Tomorrow, I have a party of **my day I was born**." (*birthday*)

## 4. DATASETS

In this section, we examine the widely used datasets used to evaluate Automated Essay Scoring and Grammatical Error Correction (GEC) systems. To date, no dedicated benchmark dataset exists exclusively for semantic errors; available evaluation corpora instead capture a broad spectrum of grammatical, lexical, and mechanical errors. Table 1 presents an analysis of primary datasets traditionally used to evaluate both general grammatical errors and specific semantic subcategories, such as collocation errors.

Table 1: Summary of available datasets

| Dataset / Corpus | Size & Scope | Target Population | Key Features & Annotation |
|---|---|---|---|
| CoNLL-2014 Shared Task [20] | 1,312 English sentences | English language learners | Annotated by 2 expert annotators; evaluated using MaxMatch ($M^2$) scorer with F0.5. |
| Lang-8 Corpus [3] | 100,051 entries (29,012 users) | Online language learners | Crowdsourced corrections; includes automatic tense/aspect annotations. |
| NUCLE Corpus [2] | 1,400 essays | Undergraduate students (primarily Asian) at NUS | Academic essay writing corpus from National University of Singapore. |

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

| | | | |
|---|---|---|---|
| CLC FCE Dataset [19] | 1,244 exam scripts (2000–2001) | Cambridge FCE exam candidates | Includes original text, marks, error annotations, and candidate demographic data. |
| Project Gutenberg [6] | Full collection of PG digital books | Standard English literary texts | Synthetically created by substituting 5% of tokens with edit-distance-1 real-word paronyms to evaluate bigram/trigram probabilistic filters |

The primary bottleneck in advancing semantic error detection and correction (EDC) is the lack of dedicated, domain-specific training and evaluation datasets. Existing corpora overwhelmingly target general Grammatical Error Correction (GEC), which contains only a tiny fraction of semantic errors. Consequently, there is an urgent need to build specialized benchmark datasets focused exclusively on semantic errors. Curating such corpora would provide the required ground-truth data to effectively train and benchmark supervised learning models.

## 5. SEMANTIC ERROR DETECTION AND CORRECTION APPROACHES

In recent decades, various approaches have been proposed globally for detecting and correcting semantic errors in text. These methods generally fall into three main categories: knowledge-based, unsupervised, and supervised deep neural network-based approaches.

**1. Knowledge-Based Semantic Detection and Correction**

Knowledge-based methods rely on pre-existing linguistic knowledge—such as language corpora or lexical databases—to evaluate whether word sequences and combinations are semantically valid. These systems use contextual probability to identify and replace erroneous words with the most suitable alternatives.

- **Dictionary of Literal Paronyms:** Gelbukh et al. [7] introduced an early approach targeting malapropisms (real words inserted incorrectly due to phonetic or structural similarity). The method uses a dictionary of literal paronyms—words within a close edit distance of the suspicious word that share identical grammatical properties—to generate correction candidates.

- **N-Gram Probabilistic Models:** These models evaluate combined bigram or trigram scores generated by neighboring words to replace real-word semantic errors with the highest-probability candidate from an edit-distance set $c(W)$.

  (a) **Fossati et al. [8]:** Proposed a mixed-trigram language model for real-word spell checking, demonstrating high hit rates for both detection and correction, though with higher false positive rates.

  (b) **Samanta et al. [6]:** Utilized Levenshtein distance ($edit\ distance = 1$) to generate confusion sets, then ranked candidates using bigram/trigram frequencies from the BYU corpus and Project Gutenberg texts.

  (c) **Multilingual Applications:** Similar n-gram strategies have been applied to non-English languages, including Bangla (Hasan et al. [9]) and Hindi (Kanwar et al. [7]).

  (d) **Hybrid Approaches:** Azmi et al. [11] combined n-gram models with machine learning to eliminate predefined confusion sets, applying preprocessing and feature extraction for detection and n-grams for correction.

**2. Unsupervised Methods**

Unsupervised methods discover word co-occurrence patterns autonomously from large, unlabeled text corpora without requiring annotated datasets of incorrect/correct sentence pairs.

- **Content Word Combination Analysis:** Kochmar [17] focused on adjective–noun (AN) and verb–object (VO) pairs, which represent a significant portion of learner content-word errors. Grounded in distributional semantics—that words with similar meanings appear in similar contexts [12][13][14]—the algorithm compares original word combinations to alternatives, scoring collocational strength alongside a machine learning classifier trained on semantic features.

- **Bidirectional LSTM Tagging:** Makarenkov et al. [15] developed a language-independent lexical substitution model using a Bi-LSTM tagger. Dense word embeddings are processed through left-to-right (prefix) and right-to-left (suffix) LSTM networks to perform classification from combined contextual representations.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

- *Limitation:* The model focuses strictly on proposing contextually better alternatives rather than explicit error detection, making direct comparison with end-to-end Grammar Error Correction (GEC) systems difficult.

**3. Supervised Deep Neural Networks & Neural Machine Translation**

Modern deep learning architectures frame error correction similarly to Neural Machine Translation (NMT), mapping an erroneous source sequence to a corrected target sequence using encoder-decoder networks, CNNs, LSTMs, and Transformers.

- **Multilayer Convolutional Encoder-Decoder:** Chollampatt et al. [4] utilized a multilayer CNN encoder-decoder to automatically correct orthographic, grammatical, and collocation errors. While effective across general text errors, performance metrics specific to semantic errors were not evaluated independently.

- **Sequence Tagging Transformers (GECToR):** Omelianchuk et al. [16] introduced GECToR, which uses a Transformer encoder pre-trained on synthetic data and fine-tuned in two stages (first on errorful corpora, then on mixed parallel corpora).

    - *Limitation:* Like many comprehensive GEC models, overall performance is evaluated on full error sets, leaving exact performance on standalone semantic errors largely unmeasured.

## 6. ANALYSIS AND COMPARATIVE STUDY

This section summarizes the primary strengths and operational constraints of the computational methodologies discussed above. Table 2 summarizes these approaches across key operational parameters.

1. **Knowledge-Based Methods:** Systems employing knowledge-based approaches report strong $F_1$ scores ranging from 80% to 90%. However, their main limitation is a strict reliance on precompiled knowledge bases, leaving them incapable of handling uncommon sentence structures or candidates beyond predefined boundaries. For instance, in the $n$-gram approach proposed in [6], the confusion set $c(W)$ generated for the target word *'his'* is restricted to $\{his, him, this, is, has\}$. Consequently, the algorithm can only evaluate replacements within this fixed set and cannot generalize beyond it. Furthermore, the evaluation sentences used to measure precision and recall in these approaches were synthetically generated by injecting errors into clean text via single-character edit operations (insertion, deletion, substitution).

2. **Unsupervised Machine Learning Methods:** These models identify structural sentence patterns and learn to infer contextually appropriate candidate words without labeled training data. Their main drawback stems directly from this reliance on unannotated corpora, which frequently leads to imprecise or contextually inappropriate word selections. For example, the Bidirectional LSTM tagger in [15] proposes lexical alternatives for a given position but lacks autonomous error localization. It cannot identify an erroneous or target word on its own and requires the target position to be specified explicitly beforehand.

3. **Deep Neural Network Approaches:** Sequence-to-sequence and tagging architectures have gained prominent traction due to their ability to transform ungrammatical sequences into fluent text while processing diverse, open-domain inputs. Rigorously trained on massive parallel corpora, these models continuously advance the state of the art in general Grammatical Error Correction (GEC) as dataset scale increases. However, their explicit performance on semantic errors remains unquantified due to the lack of dedicated semantic error datasets and specialized benchmark test-sets. Because no deep neural architecture has been trained or evaluated exclusively for semantic error detection and correction, the precise capacity of neural sequence models to resolve semantic infractions remains an open research question.

Table 2: Summary of Existing Approaches

| Approach & Year | Data | Focus | Refer. Corpora | Detection / Correction | Results |
|---|---|---|---|---|---|
| **2011 [2]** | NUS Corpus of Learner English (NUCLE) | Error correction based on L1-induced paraphrases | L1 English parallel corpus | Correction of collocation errors | MRR: 17.21 |
| **2013 [6]** | Project Gutenberg | Employs local bigrams and trigrams around target word | BYU corpus | Both | P: 71–79%, R: 81–88%, A: 85–93% |

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 9, Issue 12, December 2022

DOI: 10.17148/IARJSET.2022.91226

| | | | | | |
|---|---|---|---|---|---|
| **2016 [17]** | Cambridge Learner Corpus CLC-FCE | Compositional distributional semantics (AN & VO pairs) | BNC & ukWaC | Detection Focus | A: 65% (approx.) |
| **2018 [4]** | Concatenation of reference corpora | Multilayer convolutional encoder-decoder architecture | Lang-8 & NUCLE | Grammatical, spelling & collocation errors | P: 52.80, R: 12.80, F0.5: 32.49 |
| **2018 [9]** | Self-prepared | N-gram, Markov model for Bangla language | Self-developed corpora | Both | A: 96% |
| **2019 [15]** | ESL learner test-set | Bidirectional LSTM tagger for lexical substitution | --- | Substitution (No detection) | MRR: 0.25 |
| **2019 [11]** | Self-prepared | N-gram language model with machine learning | KSU, ANCKACST, JM corpus | Both | P: 83.5%, R: 99.2% (detection), A: 98% (correction) |

P: Precision, R: Recall, A: Accuracy, F0.5: F-Score, MRR: Mean Reciprocal Rank

## 7. CONCLUSION

Semantic error detection and correction remains one of the most challenging tasks in Natural Language Processing. Existing methodologies range from early knowledge-based rules to neural sequence architectures. Developing dedicated semantic error datasets and specialized evaluation models represents the vital next step for future research in this domain

## REFERENCES

[1]. Obeidat, H. (1986). *An Investigation of Syntactic and Semantic Errors by Arab Learners* (Doctoral dissertation). University of Illinois at Urbana-Champaign.

[2]. D. Dahlmeier, H. T. Ng, and S. M. Wu, "Building a large annotated corpus of learner English: The NUS corpus of learner English," in Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 22–31.

[3]. T. Mizumoto, Y. Hayashibe, M. Komachi, M. Nagata, and Y. Matsumoto, "The effect of learner corpus size in grammatical error correction of ESL writings," in Proceedings of COLING 2012: Posters. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 863–872.

[4]. S. Chollampatt and H. T. Ng, "A multilayer convolutional encoder-decoder neural network for grammatical error correction," CoRR, 2018.

[5]. T. Tajiri, M. Komachi, and Y. Matsumoto, "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: Association for Computational Linguistics, Jul. 2012, pp. 198–202.

[6]. Samanta, P., & Chaudhuri, B. B. (2013) A simple real word error detection and correction using local word bigram and trigram.

[7]. Gelbukh A., Bolshakov I.A. (2004) On Correction of Semantic Errors in Natural Language Texts with a Dictionary of Literal Paronyms. In: Favela J., Menasalvas E., Chávez E. (eds) Advances in Web Intelligence. AWIC 2004. Lecture Notes in Computer Science, vol 3034. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24681-7_13

[8]. Fossati D., Di Eugenio B. (2007) A Mixed Trigrams Approach for Context Sensitive Spell Checking. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2007. Lecture Notes in Computer Science, vol 4394. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70939-8_55

[9]. Hasan, K. M. & Moaj, Hozaifa & Dutta, Sanjoy. (2015). Detection of semantic errors from simple Bangla sentences. 2014 17th International Conference on Computer and Information Technology, ICCIT 2014. 296-299. 10.1109/ICCITechn.2014.7073156.

[10]. Kanwar, S., Sachan, M., & Singh, G. N-GRAMS SOLUTION FOR ERROR DETECTION AND CORRECTION IN HINDI LANGUAGE. International Journal of Advanced Research in Computer Science, 8(7), 2017.

DOI: 10.17148/IARJSET.2022.91226

[11]. A. M. Azmi, M. N. Almutery and H. A. Aboalsamh, "Real-Word Errors in Arabic Texts: A Better Algorithm for Detection and Correction," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1308-1320, Aug. 2019, doi: 10.1109/TASLP.2019.2918404.

[12]. Schfttze, Hinrich & Pedersen, Jan. (2002). Information Retrieval Based on Word Senses.

[13]. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), 211.

[14]. Pantel, Patrick. (2005). Inducing Ontological Co-occurrence Vectors. 10.3115/1219840.1219856.

[15]. Makarenkov, Victor & Rokach, Lior & Shapira, Bracha. (2019). Choosing the Right Word: Using Bidirectional LSTM Tagger for Writing Support Systems.

[16]. GECToR -- Grammatical Error Correction: Tag, Not Rewrite Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, Oleksandr Skurzhanskyi . arXiv:2005.12592 **[cs.CL]**

[17]. Ekaterina Kochmar. Error detection in content word combinations, Technical reports published by the University of Cambridge Computer Laboratory

[18]. Soni, Madhvi, and Jitendra Singh Thakur. "A systematic review of automated grammar checking in English language." *arXiv preprint arXiv:1804.00540* (2018).

[19]. Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock. "A new dataset and method for automatically grading ESOL texts." *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011.

[20]. The CoNLL-2014 Shared Task on Grammatical Error Correction overview paper is: Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task* (pp. 1-14).

 