Evaluating Hunspell, SymSpell, Norvig, and N-gram Spellcheckers for Azerbaijani Text

Israfil Gasim

AI Engineer
Neurotime LLC
Baku, Azerbaijan

Publication Date: 2025/05/31

Abstract: Automatic spelling correction is critical for enhancing text quality and usability across digital platforms, particularly for morphologically rich and low-resource languages like Azerbaijani. This paper presents a comparative analysis and benchmarking of four prominent spellchecking algorithms—Hunspell, SymSpell, Norvig's probabilistic model, and N-gram statistical models—implemented specifically for Azerbaijani. A comprehensive evaluation was conducted using a manually annotated corpus comprising diverse Azerbaijani text sources, simulating common orthographic errors typical in everyday language usage. Results indicate moderate effectiveness among all tested methods, with Hunspell achieving the highest accuracy (84.5%) due to its robust dictionary-based morphological handling. Despite its speed advantage, SymSpell (81.4% accuracy) requires extensive dictionary resources, making it impractical for morphologically complex languages without significant resource investments. Norvig's method (78.3%) and the N-gram model (82.1%) also demonstrated limitations related to corpus dependency and computational efficiency, respectively. The findings highlight substantial challenges posed by Azerbaijani's agglutinative structure, underscoring the inadequacy of existing general-purpose algorithms. Consequently, the paper emphasizes the urgent need for new hybrid approaches specifically tailored to Azerbaijani and similarly structured languages, suggesting directions for future research and development in spelling correction technologies.

Keywords: *Azerbaijani Language, Spellchecking Algorithms, Hunspell, Symspell. Norvig, N-gram, Agglutinative Languages, NLP*

How to Cite: Israfil Gasim (2025) Evaluating Hunspell, SymSpell, Norvig, and N-gram Spellcheckers for Azerbaijani Text. *International Journal of Innovative Science and Research Technology,* 10(5), 2574-2577. https://doi.org/10.38124/ijisrt/25may1640

I. INTRODUCTION

Spelling error correction is a fundamental natural language processing (NLP) task that improves the clarity and usability of text in many applications. A significant fraction of user-generated text contains spelling mistakes – for example, around 15% of search engine queries are misspelled – which can severely impact downstream tasks like search result matching and question answering[1][2]. Spelling correction is therefore employed in everything from web search and machine translation to document editing. It is also crucial as a post-processing step in optical character recognition (OCR) and in typing assistants: correcting OCR output or user input can dramatically improve text quality and user experience[3].

Despite extensive research in English and other high-resource languages, many languages remain under-served by robust spellchecking tools. Azerbaijani (Azeri) is one such low-resource language[1]. Thanks to its rich agglutinative morphology, Azerbaijani can form very complex words through derivational and inflectional suffixes[4]. For instance, multiple suffixes can be concatenated to a root, yielding word forms that are difficult to list in a simple dictionary exhaustively. This morphological complexity, including vowel harmony and numerous suffixes, poses a challenge for spell checkers since a small base lexicon can generate thousands of valid word forms. Moreover, linguistic resources for Azerbaijani are limited – large annotated corpora and comprehensive dictionaries are scarce[5]. The combination of a productive morphology and a low-resource setting means that typical approaches to spellchecking (which often rely on extensive wordlists or statistical models trained on big data) may struggle.

Orthographic errors are common in Azerbaijani digital text, yet users currently have few reliable tools to detect and correct them. Anecdotally, many Azerbaijani speakers avoid typing certain native characters (“ö”, “ğ”, “ı”, “ə”, “ç”, “ş”) when using standard keyboards, substituting nearest Latin letters instead[5]. This leads to systematic misspellings that are not handled by generic spellcheckers. Undetected spelling mistakes can degrade the quality of online content and even hinder communication. There is a clear need for accurate Azerbaijani spell correction – for social media posts, digital libraries, OCR of historical texts, and everyday word processing. Addressing this need is not trivial, as any effective solution must cope with the language's complex morphology and the relative lack of training data or existing tools.

In this paper, we systematically benchmark four different spellchecking methods on Azerbaijani text: (1) a Hunspell-based spellchecker, (2) the SymSpell algorithm, (3) Norvig's probabilistic algorithm, and (4) an N-gram language model approach. These approaches were chosen for their diversity and popularity in spelling correction tasks. Hunspell represents dictionary-based methods that leverage morphological rules; it is widely used in open-source software for many languages and is specifically designed to handle rich morphology through affix dictionaries. SymSpell is a recently introduced algorithm focused on efficiency – it precomputes deletions of dictionary terms to enable swift correction of candidate lookups[6]. Norvig's algorithm (described initially by Norvig in 2007) is a classic approach that generates possible corrections within a small edit distance and selects the most likely word based on a frequency model. The N-gram model approach uses statistical language modelling to suggest corrections that have the highest probability in context (e.g., using character or word n-grams to rank candidates). By evaluating these four methods on Azerbaijani data, we aim to illuminate which techniques are most effective for a morphologically complex, low-resource language and what unique challenges Azerbaijani poses. In particular, we examine each method's ability to handle agglutinative word forms and common error patterns (such as omitted diacritics or transliterated characters), as well as their speed and resource requirements.

II. RELATED WORK

Spellchecking and correction have been studied for decades, yielding a variety of algorithms. Here, we survey prior research relevant to Azerbaijani and the four methods we evaluate. Because Azerbaijani is under-represented in NLP literature, we also draw on insights from similar languages and general spellchecking studies.

Spellchecking and correction have been studied for decades, yielding a variety of algorithms. Here, we survey prior research relevant to Azerbaijani and the four methods we evaluate. Because Azerbaijani is under-represented in NLP literature, we also draw on insights from similar languages and general spellchecking studies. Azerbaijani Spellchecking: Early efforts to build Azerbaijani spell checkers have been relatively limited. A notable recent line of work applies neural network models. Mammadov (2019) introduced one of the first neural spell correction systems for Azerbaijani[4]. The authors highlighted that languages like Azerbaijani, with complex morphologies, can produce very long words with many affixes, complicating the detection of errors. Their system leveraged recurrent neural networks to correct misspellings, demonstrating the feasibility of data-driven approaches for Azerbaijani. More recently, Ahmadzadeh and Malekzadeh (2021) proposed a sequence-to-sequence model with attention to Azerbaijani spelling correction[5]. Because manually labeled data is scarce, they generated synthetic training examples by introducing random errors into correct sentences. This neural spellchecker reportedly achieved high correction rates, with an F1 score of around 75% on exact matches and over 90% when minor one-letter errors were tolerated. These results are promising, though the reliance on synthetic data highlights the low-resource issue. Isbarov et al. (2024) further incorporated ensemble deep-learning techniques to improve robustness[7]. Their work specifically addresses the limitations of simpler statistical methods and even standard single neural models, given the noise and ambiguity in real Azerbaijani text. By using an ensemble of deep learners, they aimed to increase the reliability of corrections on agglutinative languages. The continued interest in neural approaches indicates their potential, but such systems require significant data and computational resources, which are not always available for Azerbaijani.

In summary, existing literature underscores several points: (a) Azerbaijani's spelling correction needs are real and increasingly being addressed by researchers, (b) each algorithmic approach to spellchecking has strengths – e.g., Hunspell's linguistic coverage, SymSpell's speed, Norvig's simplicity, N-gram's context awareness – and weaknesses, and (c) there has been little direct comparison of these approaches on a common Azerbaijani dataset. This paper's contribution is to fill that gap by evaluating these four methods side by side on Azerbaijani text. By doing so, we also shed light on how well techniques developed primarily for English (and other well-resourced languages) transfer to a highly inflectional Turkic language. The insights from this comparison can inform not only tool development for Azerbaijani but also broader efforts on spell checking in other low-resource and morphologically rich languages.

III. METHODOLOGY

This study evaluates and benchmarks four popular spelling correction algorithms—SymSpell, Norvig, Hunspell, and N-gram models—implemented specifically for the Azerbaijani language. The evaluation focused on their performance and accuracy in automatically detecting and correcting spelling errors across diverse text types

A comprehensive corpus consisting of 25,000 sentences in Azerbaijani was created from various sources, including news articles, social media posts, official documents, and literary texts. The corpus was manually reviewed and annotated by linguistic experts to establish a gold-standard dataset for benchmarking.

Spelling errors were intentionally introduced into the dataset, reflecting typical mistakes found in everyday Azerbaijani typing, such as:

- Diacritical mark omission ("ş" → "s", "ğ" → "g").
- Incorrect use of similar-sounding letters ("qələm" → "qelem").
- Common keyboard errors (transpositions, deletions, insertions, replacements).

The resulting annotated dataset comprised approximately 30,000 misspellings across all sentence sets.

IV. RESULTS AND DISCUSSION

Table 1 below summarizes the averaged numerical performance metrics obtained from the evaluation:

Method	Accuracy	Precision	Recall	F1-Score
Symspell	81.4	79.7	82.2	80.9
Norvig	78.3	76.5	79.0	77.7
Hunspell	84.5	83.2	85.3	84.2
N-gram	82.1	81.0	83.0	82.0

Hunspell exhibited the highest performance with an accuracy of 84.5%, demonstrating superior capability in identifying misspelled words and suggesting accurate corrections compared to the other methods. Its strength lies particularly in handling known dictionary words and standard morphological patterns. However, despite these advantages, Hunspell still displayed significant limitations when faced with the complex agglutinative structures common in Azerbaijani. Errors frequently arose from the inability to correctly parse less common morphological forms or newly coined terms that were not present in its predefined dictionary and rule set.

SymSpell showcased remarkable speed (4 ms average execution time per correction), making it a strong candidate for real-time correction applications such as instant messaging and social media interactions. However, its accuracy (81.4%) was highly dependent on dictionary completeness. Given the extensive morphological complexity of Azerbaijani, maintaining a sufficiently comprehensive dictionary for SymSpell is highly resource-intensive and practically challenging. Additionally, SymSpell was less effective at recognizing and correcting morphological variants not explicitly pre-listed in the dictionary, significantly reducing its utility in formal or complex text contexts.

Norvig’s Algorithm, with an accuracy of 78.3%, performed acceptably well on simpler misspellings or high-frequency words, but struggled noticeably with morphologically derived errors. Its probabilistic correction strategy based on corpus frequencies had limited effectiveness for complex Azerbaijani morphology, as its reliance on large, accurately annotated corpora is difficult to fulfill given the low-resource status of Azerbaijani.

The **N-gram Model** demonstrated a moderate accuracy of 82.1%. Although it leveraged contextual information effectively, thereby handling homonyms and context-sensitive errors slightly better, it showed limitations regarding computational efficiency (18 ms per correction). Its performance suffered due to Azerbaijani’s syntactic flexibility, frequent morphological variants, and limited availability of large-scale, high-quality corpora for training robust statistical models.

Key limitations identified from this study for each method included:

- Hunspell struggled significantly with out-of-vocabulary and highly agglutinative forms. Errors emerged frequently due to missing morphological rules and lexicon coverage limitations.
- SymSpell was severely limited by its dictionary dependency, highlighting the impracticality of continuously updating comprehensive word lists that include all morphological variants.
- Norvig’s algorithm faced challenges due to its limited morphological analysis capabilities and corpus-dependence. Its effectiveness rapidly diminished for low-frequency or morphologically complex words.
- N-gram models experienced challenges due to Azerbaijani’s flexible syntax and limited corpus availability, negatively affecting model accuracy.

This comparative analysis highlights the necessity for a new, linguistically informed approach specifically designed for morphologically rich, agglutinative languages such as Azerbaijani. Existing standard algorithms, while effective for simpler morphological languages, consistently fall short in addressing Azerbaijani-specific linguistic challenges. Future research should focus on hybrid models combining rule-based morphological analyzers with data-driven methods, including neural approaches and contextual embeddings. Such a blended approach might significantly enhance performance by effectively addressing the limitations observed in the current algorithms. In conclusion, while Hunspell and SymSpell provide valuable starting points, developing more robust solutions tailored explicitly to Azerbaijani’s linguistic complexity is essential for achieving high-quality spelling correction in real-world applications.

REFERENCES

[1]. Microsoft Research, "Speller100: Zero-shot spelling correction at scale for 100-plus languages," unpublished, 2021. [Online]. Available: https://www.microsoft.com/en-us/research/blog/speller100-zero-shot-spelling-correction-at-scale-for-100-plus-languages. [Accessed May 10, 2024].

[2]. Markus Näther. 2020. An In-Depth Comparison of 14 Spelling Correction Tools on a Common Benchmark. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1849–1857, Marseille, France. European Language Resources Association

[3]. Had, I. S. ., Maulana Baihaqi, W., & Putriana Nuramanah Kinding, D. (2025). Improving Tesseract OCR Accuracy Using SymSpell Algorithm on Passport Data. Sinkron : Jurnal Dan Penelitian Teknik Informatika, 9(1), 374-381. https://doi.org/10.33395/sinkron.v9i1.14395

[4]. S. Mammadov, "Neural Spelling Correction for Azerbaijani Language," 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 2019, pp. 1-5, doi: 10.1109/AICT47866.2019.8981776.

[5]. Ahmadzade, A., & Malekzadeh, S. (2021). Spell Correction for Azerbaijani Language using Deep Neural Networks. ArXiv. https://arxiv.org/abs/2102.03218

[6]. Alan Juffs and Ben Naismith. 2025. Identifying and analyzing ‘noisy’ spelling errors in a second language corpus. In Proceedings of the Tenth Workshop on Noisy and User-generated Text, pages 26–37, Albuquerque, New Mexico, USA. Association for Computational Linguistics

[7]. Isbarov, J., Huseynova, K., & Rustamov, S. (2024, April). Robust automated spelling correction with deep ensembles. In Proceedings of the 2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI), Singapore, Singapore. ACM. https://doi.org/10.1145/3665065.3665070