University of Tehran Prees

Journal of Language Research

Online ISSN: 2676-3362

https://jolr.ut.ac.ir

Investigation of the Correspondence between the Translations Produced by ChatGPT, DeepSeek, Google Translate, Microsoft Translator and Reverso, and the Approved Orthography of the Academy of Persian Language and Literature

Gholamreza Medadian [1]

1. Corresponding Author, Assistant Professor of Translation Studies, English Language Department, Hazrat-e Masoumeh University, Qom, Iran. Email: gh.medadian@hmu.ac.ir

Article Info

Article Type:
Research Article

Article History:

Received:
13, August, 2025

In Revised Form:
27, August, 2025

Accepted:
7, September, 2025

Published Online:
20, September, 2025

Keywords:
orthography errors, punctuation errors, approved orthography, ChatGPT, Google Translate, Microsoft Translator, Reverso

Abstract

Given that adherence to the orthographic principles and conventions of the target language constitutes a significant dimension of translation quality, this study investigated the compliance of English-to-Persian journalistic translations generated by five online machine translation (MT) systems with guidelines of the approved orthography of the Academy of Persian Language and Literature (APLL) and standard usage rules for punctuation marks. To this end, several texts were extracted from English newspapers and were translated into Persian by five systems (ChatGPT, DeepSeek, Google Translate, Microsoft Translator, and Reverso). Through meticulous, word-by-word reading and analysis of the translations in the light of the guidelines of the Approved orthography and usage rules for punctuation marks, a total of 16 distinct categories of orthographic and punctuation errors were identified in the machine-generated translations. These were later classified into three main types: (1) errors in the use of non-breaking or half-space (Nim-Faslé) within word elements and elements of compound words, (2) errors in the orthography of secondary Persian characters, and (3) errors in the correct spacing of punctuation marks. Google, Microsoft, and Reverso produced instances of all error types. DeepSeek and ChatGPT exhibited 4 and 12 types of these errors in their outputs, respectively. Furthermore, the results indicated that none of the systems demonstrated complete consistency in performance so that a given compound or character was at times rendered correctly and other times incorrectly. It was, also, discerned that the identified errors bear a significant resemblance to orthographic errors prevalent in human-generated texts. This phenomenon is attributable to the predominantly open-source monolingual and bilingual corpora used for training these online systems, which comprise texts authored and translated by diverse individuals across varied contexts and registers. Considering the increasing reliance on online MT systems, such orthographic and punctuation errors possess the potential to gradually permeate into the texts generated by Persian-speaking communities, potentially undermining the APLL's ongoing orthographic standardization efforts. One proposed solution for this problem entails the development, either by the MT providers themselves or independent developers, of dedicated or integrated auxiliary systems designed to scrutinize and post-edit translations prior to final user delivery, specifically evaluating their conformity with the approved orthographic guidelines and standard punctuation conventions.

Cite this The Author(s): Medadian, Gh. (2025): Investigation of the correspondence between the translations produced by ChatGPT, DeepSeek, Google Translate, Microsoft Translator and Reverso, and the approved orthography of the Academy of Persian. Language and Literature, No. 2, Vol.15 , Serial No. 29, Autumn & Winter - (193-227). DOI: 10.22059/jolr.2025.400687.666934.

Publisher: University of Tehran Press.
https://jolr.ut.ac.ir/article_104422.html?lang=en

1. Introduction

This study investigates a critical gap in computational linguistics and Persian language processing: the adherence of machine translation (MT) outputs to standardized orthographic guidlines. It conducts a systematic evaluation of English-to-Persian journalistic translations generated by five prominent online MT systems (ChatGPT, DeepSeek, Google Translate, Microsoft Translator, and Reverso) against the standards of the Academy of Persian Language and Literature. The research is guided by two primary inquiries: first, to categorize orthographic errors in MT outputs, and second, to classify punctuation mark errors. The objective is to identify and taxonomize these errors, providing a resource for computational linguists, MT developers, and translation studies pedagogues. Mitigating such errors is critical for enhancing post-editing efficiency. Furthermore, by analyzing outputs trained on human-generated corpora, this study establishes a framework for differentiating between errors inherited from training data and those intrinsic to machine processing algorithms.

2. Literature Review

Scholarly inquiry has extensively documented non-compliance with official Persian orthography across diverse textual domains, including academic writing, journalism, legal documents, and media subtitles. A synthesis of this research reveals a consistent taxonomy of errors, primarily comprising: (1) erroneous intra-word (half-)spacing in compound lexical items, (2) incorrect application of secondary Persian characters (e.g., hamze), and (3) errors in the use of punctuation marks. The transdomain prevalence of these identical error typologies suggests systemic cross-contamination of non-standard orthographic practices within the Persian linguistic ecosystem. While human translation errors have been preliminarily examined, a salient lacuna exists regarding the systematic analysis of orthographic and punctuation errors in Persian texts generated by MT systems. This research directly addresses this gap, enabling a comparative analysis between machine and human error profiles.

3. Materials and Methods

A corpus of six English-language news reports (4,473 words) was compiled from English newspapers (The Independent, The Guardian, The Financial Times) on a single date (January 7, 2025). These texts were translated by the five selected the MT systems, generating a Persian corpus of 28,524 words. A rigorous, multi-stage manual analysis was employed to identify and codify orthographic and punctuation errors in the MT output. The analytical protocol involved three iterative review rounds to ensure reliability and achieve data saturation. Orthographic error identification was anchored in the Academy's Official Persian Orthography Guide, with the Academy's Orthographic Dictionary serving as a secondary reference. For punctuation, lacking explicit Academy guidelines, the standards of Cintas and Remael (2020) were used to detect errors. Spacing accuracy was verified using both manual inspection and automated validation via non-printing character display in Microsoft Word. Identified errors were subsequently classified based on syntactic and morphological criteria, and their distribution was comparatively analyzed across the five MT systems.

4. Results and Discussion

The analysis yielded a tripartite error taxonomy: (1) Spacing errors within compound words and structures, (2) Errors in the writing mandatory secondary characters, and (3) punctuation mark errors. A comparative analysis revealed significant inter-system disparity. DeepSeek demonstrated markedly superior orthographic correspondence, exhibiting only 4 of the 16 identified error types. In contrast, ChatGPT exhibited 12 error types, while Google Translate, Microsoft Translator, and Reverso each exhibited the full spectrum of 16 error types. Notably, DeepSeek was the sole system to produce zero punctuation errors.

The distribution and nature of these errors are not stochastic but are directly correlated with the non-standard orthographic patterns prevalent in the web-scraped monolingual and parallel corpora used for training MT systems. This is particularly evident in the consistent erroneous application of spaces within functional, closed-set lexical items (e.g., compound prepositions

and compound conjunctions), indicating these forms have been algorithmically learned from human sources. To enhance output standardization, a dual-path solution is proposed. Primarily, the curation of high-quality, normatively compliant training data is essential. A more immediately actionable solution is the implementation of a dedicated "Target Language Control and Editing Module" for post-processing MT output. This module would apply rule-based normalization to correct predictable errors in spacing, punctuation, and use of secondary Persian characters. A concomitant issue is the Academy's own orthographic guide, which permits variant spellings (e.g., connected vs. half-space-separated forms), thereby institutionalizing inconsistency. A move towards a more definitive, less permissive standard is recommended to facilitate machine normalization and overall orthographic uniformity.

5. conclusion

This study confirms that contemporary MT systems propagating a significant volume of orthographic and punctuation errors in English-to-Persian translation, with error profiles are deeply reflective of the non-standard practices found in their human-curated training data. The performance of DeepSeek indicates that architectural or training data differences can substantially impact orthographic output quality. While the study's scope is limited to a specific genre and a fixed set of systems, the shared reliance of the MT field on similar deep-learning architectures and vast, uncurated web corpora allows for cautious generalization of these findings. Future research must expand to other genres, a broader array of systems, and automated subtitle generation platforms. The aggregation of such empirical findings is vital for two parallel objectives: guiding the development of more robust and linguistically accurate MT systems, and informing evidence-based reforms of the Persian orthographic standards by the Academy of Persian Language and Literature.

دانشگاه تهران

پژوهش‌های زبانی

شاپای الکترونیکی: ۲۶۷۶-۳۳۶۲

https://jolr.ut.ac.ir/

بررسی تطابق ترجمه‌های چت‌جی‌پی‌تی، دیپ‌سیک، گوگل، مایکروسافت و ریورسو با معیارهای دستور خط مصوّب فارسی و اصول نگارش نویسه‌های سجاوندی

غلامرضا مدادیان[1]✉

۱. نویسنده مسئول، استادیار مطالعات ترجمه، گروه زبان انگلیسی ، دانشگاه حضرت معصومه (س)، قم، ایران. رایانامه: gh.medadian@hmu.ac.ir

اطلاعات مقاله

نوع مقاله:
علمی – پژوهشی
تاریخ دریافت:
۱۴۰۴/۰۵/۲۳

تاریخ بازنگری:
۱۴۰۴/۰۶/۰۷

تاریخ پذیرش:
۱۴۰۴/۰۶/۱۷

تاریخ انتشار:
۱۴۰۴/۰۶/۳۰

واژه‌های کلیدی: خطای رسم‌الخطی، خطاهای نویسه‌های سجاوندی، دستور خط مصوّب، چت‌جی‌پی‌تی، دیپ‌سیک، گوگل ترنسلیت، مایکروسافت ترنسلیتر، ریورسو

چکیده

ازآنجایی‌که تطابق ترجمه‌ها با اصول و ویژگی‌های خطی زبان مقصد یکی از ابعاد مهم کیفیت آنهاست، در این مطالعه تبعیّت ترجمه‌های انگلیسی–به–فارسی پنج سامانهٔ برخط (در ژانر متون مطبوعاتی) با دستورالعمل‌های دستور خط مصوب فرهنگستان زبان و ادب فارسی و اصول نگارش نویسه‌های سجاوندی بررسی شد. برای این منظور چندین متن برگرفته از متون انگلیسی منتشره در روزنامه‌های انگلیسی‌زبان جمع‌آوری شد و توسط پنج سامانه برخط (چت‌جی‌پی‌تی، دیپ‌سیک، گوگل ترنسلیت، مایکروسافت ترنسلیتر و ریورسو) به فارسی ترجمه شد. با بررسی دقیق ترجمه‌ها به‌صورت کلمه به کلمه بر مبنای دستورالعمل‌های دستور خط مصوّب، مجموعاً ۱۶ نوع خطای رسم‌الخطی و سجاوندی در ترجمه‌ها شناسایی گردید، که به سه دستهٔ عمده تقسیم گردیدند: خطاهای (نیم)فاصله‌گذاری بین اجزاء کلمات و ترکیبات، خطاهای نگارش نویسه‌های ثانوی و خطاهای فاصله‌گذاری نویسه‌های سجاوندی. طبق نتایج، گوگل، مایکروسافت و ریورسو تمامی این خطاها را تولید کرده بودند. دیپ‌سیک و چت‌جی‌پی‌تی نیز به‌ترتیب ۴ و ۱۲ نوع از این خطاها را در برونداد خود داشتند. به‌علاوه، نتایج نشان داد که عملکرد هیچ‌یک از سامانه‌ها کاملاً یک‌دست نیست به‌طوری‌که گاهی یک سامانهٔ واحد یک ترکیب یا نویسه واحد را صحیح نگارش کرده بود و گاهی ناصحیح. همچنین، مشخص شد که خطاها مشابهت بسیار بالایی با خطاهای رسم‌الخطی موجود در متون انسانی دارند. علت این امر آن است که متونی که برای آموزش سامانه‌های ترجمه استفاده می‌شود عموماً متون تک‌زبانه یا دوزبانهٔ منبع‌بازی هستند که توسط نویسندگان و مترجمان در بسترها و بافت‌های مختلف تولید شده‌اند. با توجه به استفادهٔ روزافزون از سامانه‌های ترجمه، این خطاها می‌توانند به‌مرور به رسم‌الخط احاد جامعهٔ فارسی‌زبان نفوذ کنند و تلاش‌های معیارسازی خط توسط فرهنگستان زبان را تحت تأثیر قرار دهند. یک راه‌حل رفع این مشکل آن است که سامانه‌های مجزا یا تجمیع‌شده‌ای توسط متولیان این سامانه‌ها (یا توسعه‌دهندگان مستقل) ایجاد شود تا قبل‌از تحویل ترجمه به کاربر نهایی آن را از منظر سازگاری با دستور خط مصوب و اصول نگارش نویسه‌های سجاوندی بررسی و پس‌ویرایش نماید.

استناد: مدادیان، غلامرضا (۱۴۰۳): بررسی تطابق ترجمه‌های چت‌جی‌پی‌تی، دیپ‌سیک، گوگل، مایکروسافت و ریورسو با معیارهای دستور خط مصوّب فارسی و اصول نگارش نویسه‌های سجاوندی. پژوهش‌های زبانی، سال ۱۶، شمارهٔ ۱، بهار و تابستان، پیاپی ۳۰– (۲۲۷–۱۹۳). DOI: 10.22059/jolr.2025.400687.666934.

ناشر: مؤسسهٔ انتشارات دانشگاه تهران

https://jolr.ut.ac.ir/article_104422.html?lang=en

۱. مقدمه

بیش از هشتاد سال از پیدایش سامانه‌های ترجمهٔ ماشینی اولیه می‌گذرد و این ماشین‌ها از مسیر بسیار پرفراز و نشیبی عبور کرده‌اند به‌طوری‌که در برهه‌هایی حتی تا مرز کنارگذاشته شدن کامل نیز پیش رفته‌اند. اولین سیستم‌های ترجمهٔ ماشینی[1] چیزی بیش از لغت‌نامه‌های دوزبانه نبودند، در ادامه و در تلاش برای فهماندن زبان به رایانه، سیستم‌های ترجمهٔ ماشینی پیشرفته‌تر رویکردهای مبتنی‌بر-قاعده را برای ترجمهٔ بین‌زبانی پیش گرفتند و موفقیت‌هایی را هم کسب کردند، اما با محدودیت‌هایی هم روبرو بودند (به‌ویژه، دشواری و کندی توسعهٔ آنها و مانع معنایی). در ادامه، محققان برای فائق‌آمدن بر این محدودیت‌ها، رویکرد آماری و نمونه‌محور به ترجمهٔ ماشینی را پیشنهاد و عملیاتی کردند؛ هردوی این رویکردها مبتنی‌بر-داده بودند. درنهایت، درحال‌حاضر سیستم‌های ترجمهٔ ماشینی با ورود به یک مرحلهٔ جدید، از روش‌های یادگیری عمیق و شبکه‌های عصبی[2] برای ترجمهٔ متون بین زبان‌ها استفاده می‌کنند.

درطول تاریخچهٔ ترجمهٔ ماشینی ارزیابی کیفیت عملکرد آنها همواره موضوع بسیار مهمی بوده است. بُعدی که بیش از هرچیز در کیفیت‌سنجی[3] ترجمهٔ ماشینی مورد توجه بوده است «فهم» درست متن مبدأ توسط ماشین و بازآفرینی محتوا و معنای آن در زبان مقصد بوده است (کمابیش به‌مانند ترجمهٔ انسانی). بااین‌حال، یکی دیگر از جنبه‌های مهم کیفیّت ترجمه (چه انسانی و چه ماشینی) تطابق و هماهنگی آن با قواعد و ویژگی‌های صوری زبان مقصد است. این بُعد کمتر مورد توجه قرار گرفته است چراکه به نظر می‌رسد تمرکز محققان حوزهٔ ترجمهٔ ماشینی بر دیگر مشکلات ترجمهٔ (رفع مانع معنایی، شناختی، نحوی[4] و غیره) باعث شده است که تطابق ترجمه با ویژگی‌های صوری زبان مقصد به حاشیه رانده شود. بااین‌حال، ویژگی‌های صوری زبان برای ماشین‌ها اهمیّت زیادی دارند و باید در تحقیقات بیش از پیش مدنظر قرار گیرند.

خطّ یکی از مهمترین ابعاد صوری هر زبانی است و هر زبان قواعد و ویژگی‌های خطی و رسم‌الخطی کمابیش خاص خود را داراست. این قواعد مجموعاً «دستور خط»[5] یک زبان را تشکیل می‌دهند. جدیدترین دستور خط رسمی فارسی در سال ۱۴۰۲ توسط

1. machine translation
2. deep learning and neural networks
3. assessment
4. Semantic, congnitive, syntactic
5. orthography

فرهنگستان زبان و ادب فارسی منتشر شده است و دستورالعمل‌های کمابیش مشخصی (و به‌زعم نگارنده این سطور، بسیار خلاصه‌ای) را برای خط معیار فارسی ارائه می‌دهد.

تاکنون پژوهش‌های بسیاری به بررسی تطابق انواع متون فارسی انسانی با دستور خط فارسی پرداخته‌اند. تک مطالعاتی نیز هماهنگی ترجمه‌های انسانی با این دستور را بررسی کرده‌اند (بخش ۲ را ببینید). نتایج این مطالعات نشان می‌دهد که انواع متون انسانی فارسی آنچنان که باید با قواعد این دستور مطابقت ندارند و خطاهای رسم‌الخطی متنوعی در آنها شناسایی، مستند و ارائه شده است. خطاهای شناسایی‌شده در حوزه‌های مانند (نیم)فاصله‌گذاری[1]، پیوسته‌نویسی کلمات و ترکیبات، نگارش نویسه‌های ثانوی فارسی و کتابت نویسه‌های سجاوندی گزارش شده‌اند. بااین‌حال، علی‌رغم اقبال روزافزون به استفاده از ترجمه‌های ماشینی تولیدی توسط سامانه‌های برخط نزد کاربران فارسی‌زبان (بالاخص، در چند سال اخیر)، تاکنون هیچ مطالعه‌ای به بررسی تطابق این ترجمه‌ها با دستور خط مصّوب فرهنگستان نپرداخته است.

به‌منظور پرداختن به این شکاف پژوهشی مهم، این مطالعه در پی آن است که تطابق ترجمه‌های مطبوعاتی انگلیسی-به-فارسی تولیدشده توسط چندین سامانهٔ ترجمهٔ ماشینی برخط کمابیش مطرح (چت‌جی‌پی‌تی، دیپ‌سیک، مترجم گوگل، مترجم مایکروسافت و ریورسو) را با دستور خط مصّوب فرهنگستان بررسی کند. این پژوهش در پی آن است که با مبنا قرار دادن دستور خط مصّوب فرهنگستان زبان و ادب فارسی و اصول نگارش نویسه‌های سجاوندی به دو پرسش زیر پاسخ دهد؟

۱- چه نوع خطاهای رسم‌الخطی در ترجمه‌های تولیدشده توسط سامانه‌های ترجمهٔ ماشینی وجود دارد؟

۲- خطاهای نگارش نویسه‌های سجاوندی موجود در ترجمه‌های تولیدشده توسط سامانه‌های ترجمهٔ ماشینی کدامند؟

هدف آن است که خطاهای رسم‌الخطی و خطاهای نگارش نویسه‌های[2] سجاوندی در ترجمه‌های تولیدی آنها شناسایی شده و در اختیار محققان، مدرسان، توسعه‌دهندگان سامانه‌ها و دیگر تصمیم‌گیران مرتبط قرار گیرد. جلوگیری از بروز این خطاها یا (حداقل) کاهش آنها در ترجمه‌ها می‌تواند در کاستن از زمان مورد نیاز برای پس‌ویرایش ترجمه‌ها مؤثر باشد. ازآنجایی‌که سامانه‌های ترجمهٔ ماشینی عمدتاً براساس متون یا

1. (half)spacing
2. orthography and punctuation errors

ترجمه‌های انسانی (ترازبندی‌شده یا موازی)[1] موجود (به‌ویژه، در اینترنت) آموزش داده می‌شوند، تحقیق حاضر می‌تواند زمینه را برای مقایسه خطاهای رسم‌الخطی انسانی و ماشینی نیز فراهم کند.

۲. پیشینهٔ پژوهش

پژوهش‌های بسیاری روی خطاهای رسم‌الخطی متون فارسی (عمدتاً بر اساس معیارهای دستور خط مصّوب فرهنگستان زبان و ادب فارسی) انجام شده است. این بررسی‌ها انواع متون فارسی را در طیفی از محیط‌ها و بافت‌ها (ازجمله، مقالات و پایان‌نامه‌های علمی، زیرنویس‌های شبکه‌های تلوزیونی، انواع پیکره‌های متنی، متون مطبوعاتی، آراء و احکام قضایی، و برخی کُتُب درسی) مورد مطالعه قرار داده‌اند. در این بخش مروری بر این پژوهش‌ها و نتایج آنها خواهیم داشت.

مقالات و پایان‌نامه‌های علمی (ذخیره‌شده در پیکره‌های رایانه‌ای) بیش از متون دیگر توسط پژوهشگران مورد بررسی قرار گرفته‌اند. کلاهدوزان، معینی، پاپی و ذوالفقاری (۱۳۸۳) پایان‌نامه‌های ارشد و دکترای دانشجویان پزشکی را مطالعه می‌کنند و در آن‌ها ایرادات متعددی در رابطه با کتابت صحیح نویسهٔ همزه، پیوسته‌نویسی/جدانویسی کلمات و نگارش نویسه‌های سجاوندی گزارش می‌کنند. احمدی‌نسب (۱۳۹۴) گزارش می‌کند که مجلات علمی نمایه‌شده در نمایهٔ آی‌اس‌سی دربارهٔ پیوسته‌نویسی/جدانویسی کلمات، استفاده از فاصلهٔ کامل بین کلمات و نگارش صورت صحیح کلمات بیشترین حساسیّت را از خود نشان می‌دهند. درمقابل، کمترین حساسیّت آنها دربارهٔ حرکت‌گذاری کلمات، کتابت واژه‌های قرضی خارجی و کلمات دارای مصوّت‌های کوتاه و نگارش کسرهٔ حالتِ اضافه است. آخشیک (۱۳۹۵) با بررسی تعدادی از مقالات پژوهشی فارسی درمی‌یابد که ایرادات (نیم)فاصله‌گذاری بین اجزاء سازندهٔ کلمات و خطاهای مربوط به درج نویسهٔ همزه دارای فراوانی بالایی در این متون هستند. آخشیک و فتاحی (۱۳۹۱) با بررسی تیتر پایان‌نامه‌ها در پایگاه «علوم و فناوری اطلاعات ایران» و «مرکز منطقه‌ای اطلاع‌رسانی علوم و فن‌آوری» گزارش می‌کنند که بیشتر متون علمی در رعایت اصول (نیم)فاصله‌گذاری بین عناصر تشکیل‌دهندهٔ کلمات ایراداتی دارند. رنجبر، عباس‌پور، ستوده و مولودی (۱۳۹۸) متون اس‌آی‌دی، مگ‌ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فنون و پایان‌نامه‌های دانشگاه شیراز را از منظر رسم‌الخطی بررسی

1.aligned or parallel

می‌کنند و گزارش می‌کنند که فقط حدود ۲۳٪ متون، از اصول و قواعد فرهنگستان پیروی کرده‌اند. آنها معتقدند که علی‌رغم آشنایی نویسندگان با سازوکار نیم‌فاصله‌گذاری در محیط رایانه، تقریباً ۶۲ درصد نویسندگان بین اجزای سازندهٔ کلمات و ترکیبات نیم‌فاصله‌گذاری می‌کنند. به‌طورکلی، بررسی‌هایی که روی رسم‌الخط متون علمی فارسی انجام شده است حاکی از آن است که نویسندگان (و/یا ویراستاران) این متون به‌طور کامل از دستورالعمل‌های رسم‌الخط فرهنگستان تبعیّت نمی‌کنند.

برخی از پژوهشگران به بررسی رسم‌الخط مطبوعات کثیرانتشار پرداخته‌اند. برای نمونه، ذوالفقاری (۱۳۸۶) ایرادات رسم‌الخطی ۲۲ روزنامه و مجلهٔ کثیرالانتشار فارسی معاصر را مطالعه می‌کند. نتایج وی نشان می‌دهد که تنها چند مجله تاحدی از قواعد رسم‌الخط مصّوب پیروی می‌کنند. او بیشترین ایرادات را در حوزهٔ روی‌هم‌نویسی کلمات طولانی و نگارش نویسهٔ همزه گزارش می‌کند. ستوده و هنرجویان (۱۳۹۱) متون موجود در پیکرهٔ روزنامهٔ همشهری را بررسی می‌کنند و گزارش می‌کنند که نویسندگان این مقالات عمدتاً از درج نویسه‌های ثانوی (انواع همزه) خودداری می‌کنند یا به‌جای آنها از نویسه‌های اصلی فارسی استفاده می‌کنند (برای نمونه، نویسهٔ «و» به‌جای «ؤ»). این یافته‌ها نشان می‌دهد که زبان فارسی مطبوعاتی فارسی نیز به‌مانند متون علمی آن با رسم‌الخط معیار فاصله دارد.

متون حقوقی فارسی نیز از منظر مطابقت با رسم‌الخط معیار مورد بررسی پژوهشگران بوده است. شیعه‌علی (۱۳۹۲) و بساک و همکاران (۱۳۹۲) با بررسی متون آراء قضایی انواعی از خطاهای رسم‌الخطی و نگارش نویسه‌های سجاوندی را در آنها شناسایی می‌کنند و استدلال می‌کنند که این ایرادات می‌توانند تبعات مهمی را برای افراد درگیر در دعاوی داشته باشند. در همین راستا، اسدالهی و آذرنیوار (۱۴۰۰) نیز در یک مطالعهٔ جدیدتر با بررسی کل متن قانون مدنی ایران، ایراداتی مانند عدم رعایت اصول جدانویسی/پیوسته‌نویسی (برای مثال، «معذالک*» و «هیچوجه*»)، نگارش نویسهٔ «همزه» روی کرسی دندانه («ـئـ») به‌جای نویسهٔ «ـیـ» («*جائز*» به‌جای «*جایز*») را گزارش می‌کنند. یافته‌های این پژوهش‌ها نشان می‌دهند متون حقوقی فارسی نیز (علی‌رغم حساسیت بالایی که برای طرفین مرافعات دارند) رسم‌الخط واحدی ندارند و ایرادات رسم‌الخطی و سجاوندی در آنها قابل‌مشاهده است.

خطاهای رسم‌الخطی در شبکه‌های تلوزیونی فارسی‌زبان ملی نیز مورد بررسی قرار گرفته‌اند. سمایی (۱۳۸۸) گزارش می‌کند که زیرنویس‌های شبکه‌های تلوزیونی فارسی

از اصول دستور خط مصوّب فرهنگستان به‌خوبی تبعیت نمی‌کنند و کتابت نویسه‌های سجاوندی نیز در آنها وضعیت مناسبی ندارد. کشاورز و ستاری (۱۳۹۰) نیز با بررسی زیرنویس‌های چندین شبکهٔ تلوزیونی فارسی‌زبان گزارش می‌کنند که اصول درج نویسه‌های سجاوندی فارسی و کتابت صورت صحیح کلمات آنچنان که شایسته است در آنها رعایت نمی‌شود و شبکه‌های مختلف در این زمینه‌ها یکدست و واحد عمل نمی‌کنند. مدرس خیابانی (۱۳۹۷) با بررسی زیرنویس‌های شبکه‌های خبر و آی‌فیلم صداوسیما ایرادات فاصله‌گذاری برون و درون‌کلمه‌ای (همان (نیم)فاصله‌گذاری) را در آنها گزارش می‌کند. درنهایت، احمدی‌نسب و همکاران (۱۴۰۱) در یک مطالعهٔ جدیدتر ایرادات رسم‌الخطی زیرنویس‌های چندین شبکهٔ تلویزیونی فارسی‌زبان را بررسی می‌کنند. نتایج آنها نشان می‌دهد که اصول پیوسته‌نویسی/جدانویسی آنچنان که شایسته است در آنها رعایت نمی‌شود. یافته‌های این تحقیقات نیز نشان می دهد صداوسیما نیز، به‌عنوان یک نهاد کاملاً دولتی، در پیروی از دستورخط معیار کاستی‌هایی دارد و عملکرد آن در شبکه‌های مختلف یکسان نیست.

کُتُب درسی از مهمترین متون هر زبانی هستند چراکه طیف وسیعی از گویشوران زبان‌ها با آنها در ارتباط قرار می‌گیرند. بنابراین، این حوزهٔ تحقیقاتی بسیار مهم است. دراین‌رابطه، هاشمی (۱۳۹۰) با بررسی کامل متن کُتب دورهٔ تحصیلی ابتدایی (سال ۱۳۸۷) گزارش می‌کند که در آنها پیوسته‌نویسی/جدانویسی افعال مرکب به‌صورت یکدستی انجام نشده است به‌طوری‌که چندین صورت مختلف برای افعال مرکب واحد در این متون قابل مشاهده است. علی‌رغم اهمیّت بالای کُتُب درسی، این تحقیق تنها تحقق موجود در این بافت در حوزهٔ زبان فارسی است و لازم است کُتُب مقاطع مختلف تحصیلی (به‌ویژه، سال‌های اخیر) از منظر رسم‌الخطی مورد بررسی قرار بگیرند.

تحقیقات متعددی نیز به شناسایی و طبقه‌بندی خطاهای رسم‌الخطی و نقطه‌گذاری سجاوندی در برونداد مترجم‌های ماشینی در حیطهٔ زبان‌های دیگر پرداخته‌اند. برای نمونه، پوپوویچ (۲۰۱۸) خطاهای رسم‌الخطی رایج در زبان‌های اسلاوی را دسته‌بندی می‌کند و نشان می‌دهد که خطاهای مربوط به نگارش حروف کوچک و بزرگ و خطاهای املایی ناشی از صرف پیچیده این زبان‌ها، از شایع‌ترین خطاها در برونادهای ماشینی هستند (Popović). در مطالعه‌ای جدیدتر، بورخارت و همکاران (۲۰۲۲) با تمرکز بر خطاهای واژگانی در ترجمهٔ ماشینی، طبقه‌بندی‌ای را ارائه می‌کنند که به‌طور خاص خطاهای نگارش اسامی خاص و اصطلاحات فنی را در برونادهای ماشینی در بر

می‌گیرد (Burchardt et al.). افزون‌براین، گوزمان و همکاران (۲۰۱۹) در پژوهشی بین‌المللی، با ارزیابی ترجمه‌های چندزبانه، به طبقه‌بندی خطاهای نقطه‌گذاری در آنها می‌پردازند (ازجمله جای‌گذاری نادرست ویرگول و نقطه) و تأثیر این خطاها بر وضوح معنایی را کمّی‌سازی می‌کنند (Guzmán et al.). به‌طورکلی، این تحقیقات بر وجود انواع خطاهای رسم‌الخطی و خطاهای نگارش نویسه‌های سجاوندی در ترجمه‌های ماشینی به زبان‌های مختلف صحّه می‌گذارند و بر ضرورت درنظرگرفتن معیارهای ارزیابی دقیق‌تر برای سنجش کیفیّت صوری خروجی مترجم‌های ماشینی تأکید می‌کنند.

برخی محققان نیز به بررسی عملکرد موتورهای جستجو در پردازش و بازیابی اطلاعات به زبان‌های غیرانگلیسی پرداخته‌اند. این موضوع با ویژگی‌های رسم‌الخطی زبان‌های مختلف و قواعد درست‌نویسی و تبعات این ویژگی‌ها برای جستجو و بازیابی اطلاعات در محیط‌های رایانه‌ای ارتباط دارد. مقداد[1] (۲۰۰۵) با بررسی شیوهٔ نگارش عربی در اینترنت و موتورهای جستجو گزارش می‌کند که موتورهای جستجو در پردازش جستارهای عربی عملکرد ضعیفی دارند و اغلب از درک ریخت‌شناسی منحصربه‌فرد این زبان ناتوانند. این مشکل عمدتاً به‌دلیل عدم پردازش صحیح اعراب، ریشهٔ کلمات و مترادف‌ها است. بررسی هامو[2] (۲۰۰۹) نشان می‌دهد که افزودن اعراب به متون عربی می‌تواند به‌طورچشمگیری اثربخشی موتورهای جستجو در بازیابی اسناد را بهبود بخشد. تاث[3] (۲۰۰۶) عملکرد موتورهای جستجو در پردازش جستارهای انگلیسی را در مقایسه با زبان مجاری بررسی می‌کند و گزارش می‌کند که بازیابی اطلاعات به زبان انگلیسی کارآیی بیشتری دارد. او این شکاف عملکردی را به پیچیدگی زبانی و چالش‌های ریخت‌شناسی بیشتر در زبان مجاری نسبت می‌دهد. مانز و ریکه[4] (۲۰۰۲) در بررسی خود نشان می‌دهند که استفاده از تحلیل ریخت‌شناسی سطحی، مانند ریشه‌یابی ابتدایی، اثربخشی بازیابی اطلاعات برای زبان‌های هلندی، آلمانی و ایتالیایی را بهبود می‌بخشد. لازارینیس[5] (۲۰۰۷ و ۲۰۰۸) عملکرد قابلیت جستجوی وبسایت‌های تجارت الکترونیک یونانی را ارزیابی می‌کند و گزارش می‌کند که آن‌ها عملکرد ضعیفی دارند و اغلب در پردازش ویژگی‌های اولیهٔ زبان یونانی (برای نمونه، اعراب) ناتوان هستند.

1. Moukdad
2.Hammo
3.Toth
4. Monz & Rijke
5. Lazarinis

لواندوفسکی[1] (۲۰۰۸) نشان می‌دهد که موتورهای جستجو در یافتن نتایج مرتبط به زبان‌های خارجی مشکل دارند و علت اصلی آن عوامل وابسته به کاربر از قبیل زبان رابط بین کابر و موتور جستجو است. درنهایت، ژانگ و سویو[2] (۲۰۰۷) قابلیت‌های چندزبانه موتورهای جستجو را ارزیابی کرده و گزارش می‌کنند که اگرچه آن‌ها خدمات اولیه‌ای مانند ترجمهٔ رابط کاربری و ترجمهٔ کلیدواژه‌ها را ارائه می دهند، اما عملکردشان ناهماهنگ است. بخشی از این ناهماهنگی به شیوهٔ نگارش کلمات در زبان‌ها و خطاهای نگارشی در آنها مرتبط است. به‌طور کلی، این تحقیقات نشان می‌دهند که موتورهای جستجو در پردازش زبان‌های غیرلاتین (عربی، مجاری و یونانی) عملکرد ضعیفی دارند.

بااینکه می‌توان گفت یکی از ابعاد مهم مهارت مترجمان تسلط آنها بر زبان مقصد و ویژگی‌های مختلف آن است (ازجمله، ویژگی‌های صوری و رسم‌الخط آن)، رعایت قواعد صحیح رسم‌الخطی در متون ترجمه‌شده کمتر مورد بررسی قرار گرفته. در تنها مطالعهٔ انجام‌شده در این حیطه، مدادیان (۱۴۰۳) خطاهای رسم‌الخطی و نگارش نویسه‌های سجاوندی را در ترجمه‌های دانشجویان کارشناسی مترجمی زبان بررسی می‌کند و خطاهای آنها را در سه دستهٔ عمده قرار می‌دهد: خطاهای (نیم)فاصله‌گذاری بین اجزای کلمات و ترکیبات، خطاهای نگارش نویسه‌های ثانوی و خطاهای نگارش نویسه‌های سجاوندی. لازم است یادآور شویم که تاکنون هیچ مطالعه‌ای بر روی خطاهای رسم‌الخطی متون فارسی ترجمه‌شده توسط ماشین یا سامانه‌های برخط انجام نشده است. این در حالی است که استفاده روزافزون از ماشین برای ترجمه به‌صورت روزافزون در حال افزایش است و در سال‌های اخیر افزایشی تصاعدی و چشم‌گیر داشته است. تحقیق حاضر گامی در جهت پر کردن این خلاء پژوهشی بااهمیّت خواهد بود.

به‌طور کلی یافته‌های تحقیقات پیشین روی متون فارسی نشان می‌دهد که اقشار مختلف فارسی‌زبانان (در محیط‌ها، بافت‌ها و رسانه‌های گوناگون) آنچنان که شایسته است از اصول، معیارها و دستورالعمل‌های دستورخط معیار فارسی پیروی نمی‌کنند و خطاها و ایرادات رسم‌الخطی فراوانی در متون تولیدی آن‌ها قابل‌مشاهده است. این خطاها در سه دسته‌بندی عمدهٔ قرار می‌گیرند: (۱) خطاهای (نیم)فاصله‌گذاری بین عناصر سازندهٔ کلمات و ترکیبات، (۲) خطاهای کتابت نویسه‌های ثانوی و (۳) خطاهای کتابت نویسه‌های سجاوندی. نکته مهم دیگر آن است که خطاهایی که محققان گوناگون

1.Lewandowski
2. Zhang & Suyu

از متون محیط‌ها و بافت‌های مختلف گزارش می‌کند (علی‌رغم برخی تفاوت‌ها) مشابهت بالایی با یکدیگر دارند. این حاکی از آن است که خطاهای رسم‌الخطی تمایل دارند که در جامعهٔ فارسی‌زبان از یک محیط به محیط دیگر نشت کنند و با یکدیگر هم‌افزایی متقابل دارند. برای مثال، صورت‌هایی ناصحیحی که یک فرد در روزنامه‌ها و زیرنویس‌های شبکه‌های تلوزیونی مشاهده می‌کند، درنهایت در مقالات علمی‌ای که او کتابت می‌کند فرصت بروز و وقوع مجدد پیدا می‌کنند؛ درمقابل، افرادی که مقالات علمی را می‌خوانند ممکن است خطاهای رسم‌الخطی مشاهده‌شده در مقالات علمی را در زیرنویس‌ها بازتولید کنند و این چرخه همواره تکرار می‌شود (تا زمانی‌که رسم‌الخط صحیح و معیار بتواند در تمامی محیط‌ها و بافت‌ها غالب شود و این چرخهٔ معیوب قطع گردد).

۳.روش تحقیق

در این مطالعهٔ شش گزارش تحلیلی مطبوعاتی انگلیسی‌زبان (روی‌هم دارای ۴۴۷۳ کلمه) برگرفته از روزنامه‌های ایدیپندنت[1]، گاردین و فایننشال تایمز در تاریخ ۷ ژانویه ۲۰۲۵ جمع‌آوری گردید. سپس، در همان تاریخ این متون توسط سامانه‌های هوش مصنوعی چت‌جی‌پی‌تی، دیپ‌سیک، گوگل ترنسلیت، مایکروسافت و ریورسو[2] به‌صورت جداگانه به فارسی برگردان و ترجمه شد. سامانه‌های چت، مایکروسافت و گوگل آمریکایی هستند، سامانه دیپ‌سیک چینی است و ریورسو فرانسوی می‌باشد. چت‌جی‌پی‌تی و دیپ‌سیک سامانه‌های مولد تبدیل‌گر تمرین‌دیده[3] و مدل‌های زبانی عظیم[4] هستند که به درخواست کاربر ترجمه نیز انجام می‌دهند. سه سامانهٔ دیگر سامانه‌های اختصاصی ترجمه هستند. تمامی سامانه‌ها کمابیش از رویکردهای یادگیری عمیق برای ترجمه استفاده می‌کنند (این مختصر مجالی برای پرداختن به جزئیات معماری این سامانه‌ها و تفاوت‌ها و شباهت‌های آنها نیست و هدف از بررسی حاضر صرفاً شناسایی انواع خطاهای رسم‌الخطی ترجمه‌های تولیدی آنهاست. بااین‌تفاسیر، خوانندگان علاقمند می‌توانند برای مطالعه در این زمینه‌ها به منابع ذیل مراجعه کنند: واسوانی و همکاران (۲۰۱۷) اصول ترجمهٔ ماشینی این پنج سیستم و سامانه‌های

1.Independent, Gaurdian and Financial Times
2. https://chatgpt.com; https://chat.deepseek.com; https://translate.google.com; https://translator.microsoft.com; https://www.reverso.net/text-translation
3. GPTs
4. LLMs

تبدیل‌گر[1] مشابه را شرح می دهند (Vaswani et al.)؛ دلوین و همکاران (۲۰۱۹) و گزارش تحقیقات مایکروسافت (۲۰۲۳) تکنیک‌های استفاده‌شده در سامانه‌هایی مانند مایکروسافت را توضیح می دهند (Devlin et al; Microsoft Research)؛ رادفورد (۲۰۱۹) اصول ترجمهٔ سامانه‌هایی مانند چت‌جی‌پی‌تی و مدل‌های زبانی عظیم را شرح می‌دهد (Radford et al.)؛ فدوس و همکاران (۲۰۲۱) سیستم تبدیل‌گر گوگل را توصیف می‌کنند (Fedus et al.)).

در ادامه، ترجمه‌های تولیدی هر پنج سامانه (روی‌هم ۲۸۵۲۴ کلمه) خط‌به‌خط و کلمه‌به‌کلمه مورد بررسی و خوانش قرار گرفت تا (الف) خطاهای رسم‌الخطی و (ب) خطاهای نگارش علائم سجاوندی موجود در آنها شناسایی و ثبت شود. برای کسب اطمینان از دقت کار، خوانش و بررسی متون سه مرتبه انجام شد. در دور دوم نیز خطاهایی (هرچند بسیار اندک) شناسایی شد که در دور اول از نظر پنهان مانده بودند. در دور سوم خطای جدیدی مشاهده نشد و جمع‌آوری داده پایان یافت.

معیار شناسایی خطاهای رسم‌الخطی، دستورالعمل‌ها و نمونه‌های موجود در آخرین ویراست دستور خط فارسی (۱۴۰۲)[2] مصّوب فرهنگستان زبان و ادب فارسی بود.(۱) در مواردی که دربارهٔ صورت صحیح برخی از کلمات و ترکیبات فارسی به‌کاررفته در ترجمه‌ها ابهام یا شکی وجود داشت، از فرهنگ املایی فرهنگستان (صادقی و زندی‌مقدم، ۱۳۹۴)[3] نیز برای کسب اطمینان مشورت گرفته شد. دستور خط و فرهنگ املایی فرهنگستان برای برخی کلمات فارسی بیش از یک صورت صحیح قائل است (*واگن‌چی/واگنچی؛ واج‌ارایی/واج‌آرائی*)؛ چنین مواردی در بررسی و شناسایی خطاها لحاظ گردید. ازآنجایی‌که رسم‌الخط فرهنگستان به نویسه‌های سجاوندی فارسی نپرداخته، برای شناسایی خطاهای نگارش نویسه‌های سجاوندی از دستورالعمل‌های سینتاس و رِمائل (۲۰۲۰)[4] کمک گرفته شد.

برای تشخیص فاصله‌گذاری بین اجزاء کلمات و ترکیبات، محقق، علاوه‌بر بررسی چشمی و دستی متون، با فعال‌سازی ابزار نمایش نویسه‌های غیرقابل‌چاپ[5] در نرم‌افزار مایکروسافتWord به‌دقت فاصله‌ها و نیم‌فاصله‌ها را به‌صورت خودکار بررسی کرد تا

1. transformer
2. https://apll.ir/wp-content/uploads/2023/07/Dastour-e-Khat-17.04.1402.pdf
3. http://apll.ir/wp-content/uploads/2018/10/F-E-1394.pdf
4. Cintas & Remael
5.non-printing characters

خطای انسانی احتمالی در شناسایی فاصله‌ها و نیم‌فاصله‌ها پوشش داده شود. دست آخر، خطاها پس‌از شناسایی و استخراج، براساس مقوله‌های (عمدتاً) نحوی و ساخت‌واژی (اسم، صفت، قید، حرف اضافه، حرف ربط و غیره) دسته‌بندی نهایی شدند و توزیع آنها در سامانه‌ها با یکدیگر مقایسه شد.

۴.یافته‌ها

پس‌از بررسی دقیق ترجمه‌های تولیدی سامانه‌ها به شیوه‌ای که در بالا توصیف شد و شناسایی خطاهای رسم‌الخطی و خطاهای نگارش نویسه‌های سجاوندی، خطاهای شناسایی‌شده به سه دستهٔ کلی (هر یک با زیرمجموعه‌های متعدد) طبقه‌بندی شدند: (۱) خطاهای (نیم)فاصله‌گذاری بین اجزای تشکیل‌دهندهٔ کلمات (غیربسیط) و ترکیبات، (۲) خطاهای نگارش نویسه‌های ثانویهٔ اجباری و (۳) خطاهای نگارش نویسه‌های سجاوندی. در این بخش به بررسی این خطاها و زیرمجموعه‌های متعدد آنها خواهیم پرداخت.

۱-۴.خطاهای (نیم)فاصله‌گذاری بین اجزای تشکیل‌دهندهٔ کلمات (غیربسیط) و ترکیبات (کد ۱۰۰۰)

این دسته حاوی ۹ زیرمجموعه است، که در ادامه معرفی می‌شوند. برای هر نوع از این خطاها از یک کد که با عدد ۱۰۰۰ آغاز می شود استفاده می‌کنیم.

۱-۱-۴. نگارش فاصلهٔ کامل بین اجزاء ترکیبات اسمی به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۱)

ترکیبات اسمی از دو یا چند جزء مستقل یا وابسته تشکیل شده‌اند و در زمان تلفظ درنگ قابل‌توجهی بین اجزاء آنها قابل استماع نیست. این ترکیبات به‌صورت یک واحد مستقل واژگانی عمل می‌کنند. طبق دستورخط مصّوب، اجزاء ترکیبات اسمی باید با درج نیم‌فاصله از یکدیگر منفک شوند (در مواردی صورت پیوسته‌نویسی‌شدهٔ این ترکیبات (*کتاب‌خانه/کتابخانه*) نیز صحیح است یا حتی ترجیح دارد). بااین‌حال، ترکیبات اسمی متعددی در ترجمه‌ها شناسایی گردید که بین اجزاء آنها «فاصلهٔ کامل» درج شده بود به‌طوری‌که از منظر صوری دو یا چند کلمه مجزا به‌نظر می‌رسیدند. در ۱۰۰۱ نمونه‌هایی از این نوع خطا آمده است.

۱۰۰۱- *تخم مرغ*، کمک هزینه*، مدیر اجرایی*، خط مقدم*، دو سوم* (چت)؛ عقب نشینی*،[۲] نتیجه گیری*، سرمایه گذاری*، بهره وری*،*

همسطح سازی، رتبه بندی*، دبیر کل*، نخست وزیر*، نرم افزار* (گوگل)؛ لابی گری*، دانش آموزان*، همه گیری*، سه شنبه*، سخت کوشی* (مایکروسافت)؛ توسعه دهنده*، پاسخ دهندگان*؛ رمز ارز* (ریورسو)؛*

با این‌وجود، سامانه‌ها در موارد متعددی اجزای ترکیبات اسمی را به‌درستی نیم‌فاصله‌گذاری کرده بودند؛ این موضوع عدم یک‌دستی در عملکرد سامانه‌ها در نگارش این ترکیبات را نشان می‌دهد.

۲-۱-۴. نگارش فاصلهٔ کامل بین اجزاء ترکیبات وصفی به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۲)

به‌مانند ترکیبات اسمی، ترکیبات وصفی نیز از دو یا چند جزء مستقل یا وابسته تشکیل شده‌اند، به‌صورت یک واحد مستقل واژگانی عمل می‌کنند و درنگ آوایی چندانی بین اجزاء آنها قابل‌تشخیص نیست. طبق دستورخط، بین این اجزاء نیز باید نیم‌فاصله‌گذاری شود، اما در ترجمه‌ها موارد متعددی مشاهده شد که بین اجزاء سازنده «فاصلهٔ کامل» درج گردیده بود. خطاهای مربوط به فاصله‌گذاری ترکیبات وصفی به سه زیرمجموعهٔ عمده قابل‌تقسیم است: الف- ترکیبات وصفی مشتق از افعال مرکب/عبارات فعلی، ب- ترکیبات وصفی دارای پیشوندواره شمارشی و ج- دیگر ترکیبات وصفی. نمونه‌هایی از این خطاها در ۱۰۰۲ آمده است. البته، در موارد متعددی نیز سامانه‌ها ترکیبات وصفی را به‌درستی نیم‌فاصله‌گذاری کرده بودند؛

۱۰۰۲- *منصوب شده*، ۵۰ نفره*، منحصر به فرد* (چت)؛ دو ماهه*، رمزنگاری شده*، منبع باز*، (گوگل)؛ قریب به اتفاق*، کوتاه مدت*(مایکروسافت)؛ طولانی مدت*، کم درآمد*، حیات بخش* (ریورسو).

۳-۱-۴. نگارش فاصلهٔ کامل بین اجزاء ترکیبات قیدی به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۳)

ترکیبات قیدی نیز (به‌مانند ترکیبات اسمی و وصفی) از دو یا چند جزء مستقل یا وابسته تشکیل شده‌اند، به‌صورت یک واحد مستقل واژگانی عمل می‌کنند و بین اجزای سازندهٔ آنها درنگ آوایی قابل‌توجهی قابل شناسایی نیست. معمولاً، یکی از این اجزاء حرف اضافه است و به‌دنبال آن یک صفت و یک پسوند می‌آید (*به + راحت + ی/به‌راحتی*). طبق دستورخط، بین این اجزاء باید نیم‌فاصله‌گذاری شود، اما ترکیبات قیدی متعددی در ترجمه‌ها مشاهده گردید که بین اجزاء آنها «فاصلهٔ کامل» کتابت

شده بود. ۱۰۰۳ نمونه‌هایی از این ترکیبات را ارائه می‌کند. در موارد معدودی هم سامانه‌ها این ترکیبات را صحیح نیم‌فاصله‌گذاری کرده بودند.

۱۰۰۳- *به تنهایی*، به سادگی* (چت)؛ به ویژه* (گوگل و مایکروسافت)؛ بدون شک*، به مراتب* (ریورسو)

۴-۱-۴. نگارش فاصلهٔ کامل بین پیشوند(واره) و تکواژ آزاد کلمات به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۴)

طبق دستورخط، باید بین اجزای کلمات غیربسیط متشکل از پیشوند(واره) و تک‌واژ مستقل نیم‌فاصله‌گذاری شود اما کلمات غیربسیط متعددی در ترجمه‌ها مشاهده گردید که در آنها بین پیشوند(واره) (تصریفی یا اشتقاقی) و تکواژ مستقل «فاصلهٔ کامل» کتابت شده بود. ۱۰۰۴ نمونه‌های از این نوع خطای رسم‌الخطی را ارائه می‌کند. در موارد متعددی نیز سامانه‌ها کلمات غیربسیط دارای پیشوند(واره) را به‌صورت صحیح نیم‌فاصله‌گذاری کرده بودند.

۱۰۰۴- *با تجربه*، پر درآمد*، می آورد* (چت)؛ *بی نظیر*؛ کم دستمزد* (گوگل) بی سابقه* با انگیزه* (مایکروسافت و ریورسو)؛

۵-۱-۴. نگارش فاصلهٔ کامل بین پسوند(واره) و تکواژ آزاد کلمات به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۵)

طبق دستور خط، باید بین اجزای کلمات غیربسیط متشکل از پسوند(واره) و تکواژ مستقل نیز نیم‌فاصله‌گذاری شود (به‌مانند پیشوند(واره)+تکواژ). البته، در مواردی پیوسته‌نویسی اجزاء هم صحیح است (*سال‌ها/سال‌ها*). بااین‌حال، کلمات متعددی در ترجمه‌ها مشاهده گردید که در آنها با نقض اصل نیم‌فاصله‌گذاری، بین پسوند و تکواژ مستقل «فاصلهٔ کامل» کتابت شده بود. ۱۰۰۵ نمونه‌های از این نوع خطا را ارائه می‌کند. ترجمه‌های بعضی از سامانه‌ها (به‌ویژه دیپ‌سیک) هیچ نمونه‌ای از این نوع خطا مشاهده نگردید. همچنین، دیگر سامانه‌ها نیر در موارد بسیاری چنین کلماتی را به‌درستی نیم‌فاصله‌گذاری کرده بودند.

۱۰۰۵- *اولین ها*؛ شده اند*، منطقه ای*، رقابتی ترین* (گوگل)؛ *قسمت هایی*، منطقه ای* (مایکروسافت)؛ *حرفه شان*، داده اند*، هسته ای* (ریورسو)

۶-۱-۴. نگارش فاصلهٔ کامل قبل و بعداز «واو» در مرکب‌های عطفی به‌جای نیم‌فاصله‌گذاری (کد ۱۰۰۶)

طبق دستورخط مصّوب، باید بین اجزاء مرکب‌های عطفی نیم‌فاصله‌گذاری شود (یا در مواردی پیوسته‌نویسی شوند)، اما مرکب‌هایی در ترجمه‌ها شناسایی شد که این اصل را نقض کرده بودند. ۱۰۰۶ نمونه‌هایی از موارد شناسایی‌شده در ترجمه‌ها را نشان می‌دهد. ترکیب «جستجو» نیز در ترجمه‌ها مشاهده گردید، اما ازنظر فرهنگستان (فرهنگ املایی، ۱۳۹۴، ص. ۲۱۵) این‌صورت نیز به‌مانند «*جست‌وجو*» صحیح است.

۱۰۰۶- *دست و پنجه** (چت و گوگل)؛ *حل و فصل** (ریورسو)

۷-۱-۴. نگارش فاصلهٔ کامل بین اجزاء تشکیل‌دهندهٔ «حروف اضافهٔ مرکبِ دوجزئی» به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۷)

«حروف اضافهٔ مرکب دوجزئی» به مجموعه‌ای بسته از کلمات کاربردی که اعضای آن معدود است تعلق دارند. این حروف مرکب از یک حرف اضافه (*به، در، از و بر* و غیره)، که می‌تواند قبل یا بعداز اسم بیاید، و یک اسم ساخته می‌شوند و دستور خط آنها را برای سهولت درک عموم گویشوران فارسی‌زبان «حروف اضافه مرکب» نام‌گذاری می‌کند. برخی از این ترکیبات در عمل حروف اضافه هستند (*قبل‌از، بعداز، دربارهٔ*، غیره) اما برخی تنها در ساختمان خود یک حرف اضافه دارند (*براساس، به‌عنوان، به‌سبب، به‌دلیل، ازنظر، ازسوی*، غیره).[۳] این ترکیبات نیز یک واحد مستقل واژگانی را تشکیل می‌دهند و از منظر آوایی بین اجزاء آنها درنگ قابل‌توجهی وجود ندارد. طبق نظر دستورخط، باید بین دو جزء سازندهٔ این ترکیبات نیز نیم‌فاصله‌گذاری شود، اما نمونه‌های متعددی در ترجمه‌ها شناسایی گردید که بین اجزاء «فاصلهٔ کامل» درج شده بود. ۱۰۰۷ نمونه‌هایی مستخرج از ترجمه‌ها را نشان می‌دهد.

۱۰۰۷-*به خاطر**، *به دلیل**، *به عنوان**، *به طرز**، *به طور**، *به سمت**، *در طول**، *در مورد**، *در عرض**، *در طی**، *از لحاظ**، *از جمله**، *پس از**، *پیش از**، *قبل از**، *از سوی**، *از طرف**، *از نظر**، *بر اساس**، *در کنار** (چت، دیپ‌سیک، گوگل، مایکروسافت، ریورسو)

تقریباً تمامی حروف اضافهٔ مرکبی که در ترجمه‌های هردو سیستم شناسایی گردید ناصحیح فاصله‌گذاری شده بودند.

۸-۱-۴.نگارش فاصلهٔ کامل بین اجزاء «حروف ربطِ مرکبِ دو یا چندجزئی» به‌جای نیم‌فاصله‌گذاری بین آنها یا پیوسته‌نویسی (تمام یا برخی) از اجزاء آنها (کد ۱۰۰۸)

طبق دستور خط، برخی از این حروف در ساختمان خود دارای «که» و برخی فاقد «که» هستند (*بااین‌که، بااین‌وجود*). همچنین، در ترکیب آنها حروف اضافه و اسم نیز به چشم می‌خورد. درهرصورت، تمامی آنها به طبقهٔ واژگان کاربردی (نه محتوایی) تعلق دارند و یک واحد واژگانی مستقل به‌حساب می‌آیند. بین اجزاء این ترکیبات درنگ آوایی قابل‌توجهی قابل‌استماع نیست. طبق نظر دستورخط، باید بین اجزاء این حروف نیز (مانند حروف اضافه مرکب) نیم‌فاصله‌گذاری شود، اما نمونه‌های متعددی در ترجمه‌ها شناسایی گردید که بین اجزاء سازندهٔ آنها «فاصلهٔ کامل» درج گردیده بود یا کمابیش پیوسته‌نویسی شده بودند. ۱۰۰۸ نمونه‌های از این خطا را ارائه می‌کند. تمامی حروف ربطِ مرکبِ شناسایی‌شده در سامانه‌ها (به‌مانند حروف اضافه مرکب شناسایی‌شده در سامانه‌ها) از منظر فاصله‌گذاری ناصحیح نگارش شده بودند.

۱۰۰۸- *با این حال**، *علاوه بر**، *در عین حال**، *در همین حال**، *زمانی که**، *زمانیکه**، *از آنجایی که**، *در حالی که**، *در حالیکه**، *جایی که**، *همانطور که**

۹-۱-۴. نگارش فاصلهٔ کامل بین اجزاء ترکیبات عربی‌الاصل حاوی الف و لام به‌جای نیم‌فاصله‌گذاری بین آنها (کد ۱۰۰۹)

این ترکیبات نیز در فارسی واحدهای واژگانی مستقل هستند و درنگ آوایی ناچیزی بین اجزاء سازندهٔ آنها وجود دارد. طبق دستور خط، بین اجزاء چنین ترکیباتی باید نیم‌فاصله‌گذاری انجام شود، در تنها موردی از این نوع ترکیبات که در ترجمه‌ها مشاهده شد این قاعده نقض شده بود. ۱۰۰۹ این نمونه را نشان می‌دهد. صورت صحیح این ترکیب «*فارغ‌التحصیلان*» است. تنها دیپ‌سیک این ترکیب را صحیح تولید کرده بود.

۱۰۰۹- *فارغ التحصیلان**

۴-۲. خطاهای نگارش نویسه‌های ثانوی اجباری (کد ۲۰۰۰)

خط معاصر فارسی علاوه‌بر استفاده از نویسه‌های اولیه (*ا، ب، ج، د*، و غیره) از تعداد معدودی از نویسه‌های ثانویی نیز بهره می‌برد. ازنظر رسم‌الخط مصوّب، برخی از این نویسه‌های ثانوی، اجباری و برخی اختیاری هستند. در ترجمه‌های بررسی‌شده حتی یک

مورد نگارش علائم ثانویه اختیاری (تشدید و حرکات ضمه، فتحه و کسره) در کلمات مشاهده نگردید. در رابطه با نگارش نویسه‌های اجباری ((همزه، یای کوتاه و تنوین)، ۴ نوع خطا در ترجمه‌ها شناسایی شد، که در این قسمت آنها را معرفی می‌کنیم.

عدم نگارش نویسهٔ «یای کوتاه» (ء) بر روی «های غیرملفوظ» (کد ۲۰۰۱)

طبق دستور خط، نویسهٔ «ی» درصورتی‌که بعداز «های» غیرملفوظ پایانی بیاید به‌صورت نویسهٔ یای کوتاه «ء» بر روی آن نگارش می‌شود (*نامهٔ، خانهٔ*)، اما در ترجمه‌های سامانه‌ها موردی مشاهده نشد که این اصل در آن رعایت شده باشد. جالب است که در چنین کلماتی حتی نویسه یای بلند به‌عنوان یک جایگزین قدیمی و منسوخ هم نگارش نشده بود (برای نمونه، *خانواده‌ی**). ۲۰۰۱ مواردی از این نوع خطا را در سامانه‌ها نشان می‌دهد.

۲۰۰۱- الف- *خانواده** *دیوید، مجموعه** *کامل؛ محاکمه** *دوماهه** *پرتنش* (چت)؛ *جلسه** *امضای خصوصی*، به گفته دادگاه*، نویسنده** *کتاب* (گوگل)؛

عدم نگارش «همزهٔ میانیِ ماقبلِ مفتوح» روی کرسی «الف» (کد ۲۰۰۲)

در ترجمه‌ها کلماتی شناسایی شدند که در آنها «همزهٔ میانیِ ماقبلِ مفتوح» روی کرسی «الف» حذف شده بود. ۲۰۰۲ مواردی از این نوع خطا را نشان می‌دهد.

۲۰۰۲- *تامین** (گوگل، مایکروسافت و ریورسو)؛ *رای** (گوگل)؛ *تاثیر** (ریورسو)

گوگل و ریورسو یک کلمه مشخص را یک‌بار صحیح و یکبار ناصحیح نگارش کرده بودند (*تامین*/تأمین* (گوگل)؛ *تاثیر*/تأثیر* (ریورسو))، که نایکدستی عملکرد این سامانه‌ها در نگارش این نوع همزه را نشان می‌دهد. نکته قابل‌توجه آن است که در ترجمهٔ گوگل این موضوع در بافت یک پاراگراف واحد روی داده بود. در چت و دیپ‌سیک این خطا مشاهده نشد.

عدم نگارش «همزهٔ میانیِ ماقبلِ مضموم» روی کرسی «واو» (کد ۲۰۰۳)

کلماتی نیز در ترجمه‌های سامانه‌ها شناسایی شدند که در آنها همزهٔ میانیِ ماقبلِ مضموم روی کرسی «واو»، که نویسه‌ای اجباری است، حذف شده بود. ۲۰۰۳ مواردی از این نوع خطا را نشان می‌دهد. در رابطه با این خطا، سامانهٔ ریورسو یک کلمهٔ مشخص را یک‌بار صحیح و یکبار ناصحیح نگارش کرده بود (*سوال*/سؤال: ریورسو*).

۲۰۰۳- *موسسه** (چت؛ گوگل، مایکروسافت)؛ *سوال** (دیپ‌سیک؛ گوگل، مایکروسافت، ریورسو)؛ *روسا** (گوگل، مایکروسافت)

عدم نگارش تنوین نصب در قیود عربی‌الاصل (کد ۲۰۰۴)

تنوین نیز یکی از نویسه‌های ثانوی اجباری خط فارسی است و در قیودی که فارسی از عربی وام گرفته است وجود دارد. مشاهده شد که در برخی از این قیود تنوین حذف شده است. البته، گاهی هم در سامانه‌ها این نویسه درج شده بود.

۲۰۰۴- *نسبتا** (گوگل، مایکروسافت)؛ *متعاقبا**؛ *مستقیما**؛ *عمدا**؛ *فورا**، *قبلا**، *تقریبا** (مایکروسافت)؛ *کاملا**، *اساسا** (ریورسو)؛ *ظاهرا**، *واقعا** (مایکروسافت، ریورسو)

۳-۴.خطاهای نگارش نویسه‌های سجاوندی (کد ۳۰۰۰)

نویسه‌های سجاوندی از نویسه‌های اجباری هستند و خط فارسی معیار نیز نمی‌تواند و نباید از این امر مستثنی باشد. بااین‌حال، دستورخط دربارهٔ این نویسه‌ها دستورالعملی ارائه نکرده است. در بررسی ترجمه‌ها ۳ نوع خطا در حوزهٔ نگارش این نویسه‌ها مشاهده گردید، که در اینجا معرفی می‌گردند. لازم به ذکر است که سامانه‌ها در موارد بسیاری نیز این نویسه‌ها را درست درج کرده بودند.

نگارش فاصلهٔ کامل بین برخی نویسه‌های سجاوندی (نقطه، ویرگول، گیومه پایانی و غیره) و نویسهٔ قبل‌از آن‌ها (کد ۳۰۰۱)

طبق دستورالعمل سینتاس رمائل (۲۰۲۰) برخی نویسه‌های سجاوندی (نقطه، ویرگول،کاما، نقطه‌ویرگول، دو نقطه، پرانتز و گیومهٔ پایانی و غیره) باید بدون فاصله از نویسهٔ قبل‌از خود نگارش شوند، اما در ترجمه‌ها مواردی مشاهده شد که این قاعده نقض شده بود. ۳۰۰۱ نمونه‌هایی از این خطا را نشان می‌دهد. تنها در ترجمه‌های دیپ‌سیک نمونه‌ای از این خطا مشاهده نشد.

۳۰۰۱-

*... بسیاری از افراد با مشکل مواجه‌اند .** (چت)

*دکتر مری بوستد ، دبیر کل مشترک اتحادیه ملی آموزش، گفت ...** (گوگل)

*کریگ رایت ، خالق ارز رمزنگاری شده خودخوانده بیت کوین ، دستور پرداخت میلیاردها دلار را صادر کرد.** (مایکروسافت)

نگارش نویسهٔ نقل‌قول انگلیسی ("..." یا '...') به‌جای نویسهٔ نقل‌قول فارسی («») (کد ۳۰۰۲)

خط فارسی از نویسهٔ گیومه برای نقل‌قول مستقیم (یا تأکید و برجسته‌سازی) استفاده می‌کند اما موارد متعددی (در ردهٔ کلمه، عبارت و جمله) در ترجمه‌ها مشاهده شد که این اصل نقض شده بود. ۳۰۰۲ نمونه‌هایی از این نوع خطا را ارائه می‌کند. تنها در ترجمه‌های تولیدی دیپ‌سیک موردی از این خطا مشاهده نشد.

-۳۰۰۲

رایت گفت: "اگر مستقیماً از من می‌پرسید، بله." (چت)

... باید به معلمان موقت و "غیرمتخصصان" تکیه کنند. (گوگل)

مانشاوس با استفاده از قسمت هایی از "ماجرای ساتوشی"، استدلال کرد. (مایکروسافت)

نگارش خط تیره[1] (-) یا ان‌دش[2] (–) به‌جای ام‌دش (—)[3] (کد ۳۰۰۳)

نمونه‌هایی در متون ترجمه‌شده دیده شد که در آنها سامانه‌ها نویسهٔ سجاوندی ام‌دش (—)، که اساساً به‌جای و در نقش پرانتز به کار می رود، را به نویسه‌های نزدیک و شبیه به آن (خط تیره (-) و ان‌دش (–)) تغییر داده بودند. این دو نویسهٔ اخیر کاربردهای متفاوتی مانند ترکیب کلمات با یکدیگر (خط‌تیره) و نشان‌دادن بازه‌های زمانی و مسافتی (ان‌دش) دارند. این نوع خطا در چت و دیپ‌سیک مشاهده نشد. در دیگر سامانه‌ها نیز گاهی این نویسه به‌درستی در ترجمه‌ها بازتولید شده بود.

-۳۰۰۳

... اثبات مالکیت رایت بر کلیدهای خصوصی ساتوشی - حرکتی که بسیاری از منتقدان رایت می‌گویند اختلاف ... را حل می‌کند - کافی نیست. * (خط‌تیره (-) به‌جای ام‌دش (—): گوگل)

فشاری که بر استانداردهای زندگی وارد شده – بزرگترین بحران از زمان جنگ جهانی دوم – معیشت معلمان را تهدید می‌کند. * (ان‌دش (–) به‌جای ام‌دش (—): گوگل)

پس از یک جلسه امضای خصوصی - با هدف اثبات مالکیت کلیدهای خصوصی ساتوشی - با توسعه دهنده بیت کوین گاوین اندرسن در سال ۲۰۱۶، درد عاطفی شدید ... را متحمل شد. * (خط تیره (-) به‌جای ام‌دش (—): ریورسو)

1. Hyphen
2. En dash
3. Em dash

۴-۴. مقایسهٔ خطاهای رسم‌الخطی و نویسه‌های سجاوندی در سامانه‌ها

جدول ۱ توزیع خطاهای شناسایی‌شده در ترجمه‌ها را در پنج سامانهٔ مورد بررسی در این مطالعه نشان می‌دهد. با یک نگاه مشخص است که سامانهٔ دیپ‌سیک با اختلاف قابل‌توجه از دیگر سامانه‌ها کمترین انواع خطاهای رسم‌الخطی را در ترجمه‌های خود تولید می‌کند (تنها ۴ نوع خطا از بین ۱۶ نوع خطا). چت‌جی‌پی‌تی، که مانند دیپ‌سیک یک سامانهٔ مولد-تبدیل‌گر تمرین‌دیده است، با ۱۲ نوع خطا در جایگاه دوم قرار دارد. درواقع، از این منظر، عملکرد دیک‌سیک ۳ برابر بهتر از چت‌جی‌پی‌تی است. گوگل، مایکروسافت و ریورسو، به‌عنوان سامانه‌های اختصاصی ترجمه، هر سه مشترکاً هر ۱۶ نوع خطا را در ترجمه‌های خود تولید می‌کنند و از این منظر بر یکدیگر برتری ندارند.

نکتهٔ قابل‌توجه دیگر آن است که دیپ‌سیک، به‌عنوان جدیدترین سامانه بین این ۵ سامانه، هیچ‌یک از سه نوع خطای «نگارش نویسه‌های سجاوندی» را در ترجمه‌های خود تولید نمی‌کند. درمقابل، گوگل ۲ نوع از این خطاها را دارد و دیگر سامانه‌ها هر سه نوع خطا را تولید می‌کنند. این موضوع هماهنگی بهتر ترجمه‌های تولیدی دیپ‌سیک با اصول نگارش نویسه‌های سجاوندی را نشان می‌دهد.

جدول ۱: توزیع مقابله‌ای انواع خطاها در سامانه‌ها

	خطاهای رسم‌الخطی	چت‌جی پی‌تی	دیپ سیک	گوگل	مایکرو سافت	ریورسو
خطاهای فاصله‌گذاری (کد ۱۰۰۰)	فاصلهٔ کامل بین اجزاء ترکیبات اسمی (کد ۱۰۰۱)	√	×	√	√	√
	فاصلهٔ کامل بین اجزاء ترکیبات وصفی (۱۰۰۲)	√	×	√	√	√
	فاصلهٔ کامل بین اجزاء ترکیبات قیدی (۱۰۰۳)	√	×	√	√	√
	فاصلهٔ کامل بین پیشوند(واره) و تک‌واژ آزاد (۱۰۰۴)	√	×	√	√	√
	فاصلهٔ کامل بین پسوند(واره) و تک‌واژ آزاد (۱۰۰۵)	×	×	√	√	√
	فاصلهٔ کامل بین اجزاء ترکیبات عطفی دارای واو (۱۰۰۶)	√	×	√	√	√
	فاصلهٔ کامل بین اجزاء حروف اضافهٔ مرکبِ (۱۰۰۷)	√	√	√	√	√
	فاصلهٔ کامل بین اجزاء حروف ربطِ مرکبِ (۱۰۰۸)	√	√	√	√	√
	فاصلهٔ کامل بین اجزاء	√	×	√	√	√

	ترکیبات عربی‌الاصل دارای الف و لام (۱۰۰۹)					
اجباری	عدم نگارش «یای کوتاه» (۲۰۰۱)	√	√	√	√	√
خطاهای نویسهٔ‌های ثانوی	عدم نگارش «همزهٔ» روی کرسی «الف» (۲۰۰۲)	×	×	√	√	√
	عدم نگارش «همزهٔ» روی کرسی «واو» (۲۰۰۳)	√	√	√	√	√
	عدم نگارش تنوین (۲۰۰۴)	×	×	√	√	√
خطاهای نویسه‌های سجاوندی	نگارش فاصلهٔ کامل بین نویسه سجاوندی و نویسهٔ قبلی (۳۰۰۱)	√	×	√	√	√
	نگارش نویسهٔ نقل‌قول انگلیسی به‌جای نویسهٔ گیومه فارسی (۳۰۰۲)	√	×	√	√	√
	نگارش خط تیره یا ان‌دش به‌جای ام‌دش (۳۰۰۳)	×	×	√	√	√
	جمع کل	**۱۲**	**۴**	**۱۶**	**۱۶**	**۱۶**

۵. بحث

پس‌از معرفی و ارائهٔ نمونه برای خطاهای رسم‌الخطی و خطاهای نگارش نویسه‌های سجاوندی در بخش یافته‌ها، در این بخش به بحث دربارهٔ این خطاها خواهیم پرداخت. هدف آن است که منشاء آنها را واکاوی کنیم، آنها را با خطاهای انسانی گزارش‌شده در تحقیقات پیشین قیاس کنیم و راه‌حل‌هایی را برای کاهش رخداد آنها در سامانه‌ها پیشنهاد کنیم.

نتایج به‌وضوح نشان می‌دهد که سامانه‌ها «تمامی» حروف اضافهٔ دوجزئی (*به‌عنوان، به‌ازای، به‌سبب و غیره*) و «حروف ربط دو/چندجزئی (*بااین‌وجود، درصورتی‌که، چنان‌که، به‌این‌ترتیب، به‌این‌علت و غیره*) را در ترجمه‌ها ناصحیح نیم‌فاصله‌گذاری می‌کنند (یعنی، به‌جای نیم‌فاصله، فاصلهٔ کامل بین اجزاء آنها درج می‌شود). با توجه به معماری سامانه‌ها (که جملگی براساس نوعی از روش‌های یادگیری عمیق هستند و کمابیش از پیکره‌های متنی موازی و عظیم به‌عنوان منابع داده‌ای برای آموزش و یادگیری استفاده می‌کنند)، این یافته حاکی از آن است که نویسندگان و/یا مترجمان فارسی‌زبان (و حتی غیرفارسی‌زبان آشنا با فارسی) که متون تک‌زبانه و/یا ترجمه‌های ترازشدهٔ آنها برای آموزش/یادگیری این سامانه‌ها به کار گرفته شده، با دستورالعمل‌های رسم‌الخط مصوّب

دربارهٔ نگارش این دو نوع حروف هماهنگی ندارند. به‌بیان‌دیگر، خطاهای آنها در این زمینه در ترجمه‌های تولیدی سامانه‌ها منعکس شده است. این نوع حروف از کلمات کاربردی (نه محتوایی) فارسی هستند و اساساً به مجموعه‌ای بسته با تعداد اعضای معدود تعلق دارند. با توجه به این ویژگی، خطاهایی مربوط به این حروف مرکب به‌سادگی توسط سامانه‌ها قابل‌تصحیح است، به‌شرطی‌که صورت صحیح آنها به‌طریقی در اختیار سامانه‌ها قرار بگیرد. طبق یافته‌ها، تمامی سامانه‌ها نویسه‌های ثانوی اختیاری (تشدید و حرکات ضمه، فتحه وکسره) را در ترجمه‌های تولیدی خود نگارش نمی‌کنند و آنها را از کلمات حذف می‌کنند. ازآنجایی‌که رفتار رسم‌الخطی سامانه‌ها بازتابی از رفتار رسم‌الخطی نویسندگان و مترجمان است، می‌توان چنین استنباط کرد که مترجمان، به‌طورخاص، و فارسی‌زبانان، به‌طورعام، تمایلی به نگارش این نویسه‌های ثانوی اختیاری در خط فارسی ندارند (حتی در مواردی که درج آنها می‌تواند از معنای کلمات هم‌نویسه رفع ابهام نماید(*معین/معیّن*)) و بنابراین نمونه‌ای از نگارش آنها در ترجمه‌های ماشینی مشاهده نمی‌شود.

طبق یافته‌ها، تمامی سامانه‌ها در نگارش نویسهٔ «یای کوتاه» (ء) به‌عنوان یک نویسهٔ ثانوی اجباری بر روی «های غیرملفوظ» نیز کاملاً ناکام هستند و حتی به‌جای آن نویسهٔ یای بلند (ی)، که اکنون منسوخ در نظر گرفته می‌شود، نیز درج نمی‌کنند. گرچه طبق یافته‌های پیشین، فارسی‌زبانان به درج یای کوتاه روی های غیرملفوظ (به‌مانند دیگر نویسه‌های ثانوی) کم‌توجه هستند (بخش دو را ببینید)، اما درج این نویسه باعث تسهیل خوانش متون فارسی کلمات می‌شود. برای غیرفارسی‌زبانان نیز درج این نویسه اهمیت دوچندان دارد. بنابراین، مهم است که سامانه‌ها بتوانند این نویسه را در کلمات به‌درستی کتابت کنند. واقعیت آن است که تصحیح خطای «عدم درج یای کوتاه» برای سامانه‌های مدرن مبتنی‌بر-یادگیری عمیق اصلاً ساده نیست، چراکه سامانه باید (برای نمونه) در ابتدا ترکیبات صفت‌وموصوفی که در آنها صفت به های غیرملفوظ ختم شود (برای مثال، *ترجمهٔ شفاهی*) را در متن شناسایی کند تا بتواند بر روی «های غیرملفوظ» نویسهٔ یای کوتاه نگارش نماید. به‌عنوان نمونهٔ دیگر، سامانه باید بتواند رابطهٔ ملکیت بین دو اسم را تشخیص دهد تا بتواند روی های غیرملفوظ پایانی در اسم اول یای کوتاه درج کند (*نامهٔ رئیس‌جمهور*). در مواردی که دو یا چند اسم مختوم به های غیرملفوظ به دنبال هم بیایند (برای مثال، *مدرسهٔ ترجمهٔ شفاهی دانشگاه*)، تشخیص نیاز به درج یای کوتاه بر روی کلمات برای سامانه حتی سخت‌تر هم خواهد شد.

برای اینکه سامانه‌ها بتوانند نیاز به درج یای کوتاه در کلمات را تشخیص دهند (حداقل) لازم است که دارای مجموعهٔ بزرگی از لغات برچسب‌گذاری‌شده باشند و یک واحد برچسب‌گذاری و تحلیل نحوی برخط در آنها تعبیه شده باشد. این واحد یکی از عناصر اصلی ترجمهٔ ماشینی مبتنی‌بر-قاعدهٔ سنتی است و سامانه‌های جدید ترجمه (چه آماری، چه عصبی، چه مبتنی-برنمونه و چه مدل‌های زبانی مانند سامانه‌های مولد و تبدیل‌گر[1]) اساساً آن را به‌دلیل محدودیت‌های فراوانی که دارد کنار گذارده‌اند.

می‌توان از یک دید متفاوت نیز به عدم درج یای کوتاه در ترجمه‌ها نگاه کرد. درواقع، به‌نظر می‌رسد که نگارش این نویسهٔ ثانوی در میان نویسندگان و مترجمان محبوبیّت ندارد. حال سؤال این است که آیا فرهنگستان باید اساساً بر الزام نگارش آن اصرار کند یا اینکه ترجیح عامهٔ فارسی‌زبانان معاصر را بپذیرد و آن را در ویراست‌های آینده خود یک نویسهٔ ثانوی اختیاری اعلام کند. ناگفته نماند که تغییر زبانی یک پدیدهٔ طبیعی و شناخته‌شده در زبان‌های بشری است و ممکن است در گذر زمان عنصری از یک زبان حذف شود یا به آن اضافه شود و یا کاربرد آن دستخوش تغییر گردد.

ازآنجایی‌که نویسه‌های سجاوندی نیز ذاتاً مجموعه‌بسته بوده و دارای نقش کاربردی هستند، برای آنها نیز می‌توان قواعد مشخصی را در «بخش کنترل و ویرایش زبان مقصد» سامانه‌های ترجمه تعریف کرد تا از رخداد خطاهای نگارش نویسه‌های سجاوندی در ترجمه‌ها جلوگیری شود. درواقع، ساده‌ترین خطاهایی که می‌توان آنها را در ترجمه‌های تولیدی سامانه‌ها قبل‌از تحویل ترجمه به کاربر تصحیح کرد همین خطاهای نویسه‌های سجاوندی است.

نتایج نشان داد که گاهی سامانه‌ها یک کلمه مشخص را در یک متن واحد (و حتی یک بند واحد) یک‌بار به‌صورت صحیح و یک‌بار به‌صورت ناصحیح نگارش کرده بودند (*نرم‌افزار/نرم افزار**: گوگل)؛ *تامین**/*تأمین*: گوگل؛ *سال ها**/*سال‌ها/سال‌ها*: گوگل؛ *ارائه شده**/*ارائه‌شده*: گوگل؛ *سوال**/*سؤال*، *تاثیر**/*تأثیر*: ریورسو). این از یک نایکدستی فاحش در عملکرد سامانه‌ها پرده برمی‌دارد. البته، با توجه به معماری مبتنی‌بر-یادگیری عمیق غالب سامانه‌ها، قابل‌درک است که آنها ممکن است هریک از این صورت‌ها را از یک منبع متفاوت الگوبرداری و بازیابی کرده باشند. یک راه‌کار برای حل مشکلاتی از این دست آن است که خروجی سامانه‌های مترجم درنهایت توسط «واحد کنترل و ویرایش زبان مقصد» سامانه بررسی، خطایابی و اصلاح شود. البته، لازم است که این

1. GPTs and transformers

واحد با داده‌ها، صورت‌ها و حتی قواعد صحیح آموزش داده شود. این امر مستلزم آن است که سامانه‌های جدید داده‌محور و مدل‌های زبانی عظیم در مرحله نهایی و قبل‌از تولید ترجمه به‌عنوان خروجی سیستم به معماری مبتنی‌بر-قاعده سنتی رجوع کنند. راه‌کار دیگر آن است که به‌اندازه کافی داده‌های منطبق با رسم‌الخط صحیح فارسی در اختیار سامانه‌ها قرار گیرد تا بتوانند با الگوبرداری از آنها صورت‌های صحیح کلمات را در ترجمه‌ها بازتولید کنند.

سامانه‌های مورد بررسی در این مطالعه اساساً از متون تک‌زبانهٔ دسترسی‌باز موجود در اینترنت، متون منتشره در شبکه‌های اجتماعی، زیرنویس‌ها و متون رسانه‌ای یا انواع متون دوزبانه (برای نمونه، متون وب‌سایت‌های چندزبانه) که عمدتاً به‌صورت خودکار و بدون دخالت انسان متخصص و زبان‌شناس ترازبندی می‌شوند در مرحلهٔ آموزش خود استفاده می‌کنند. امروزه، حافظه‌های ترجمه (که در ردهٔ جمله تراز[1] می‌شوند) نیز در اینترنت در دسترس هستند و یکی دیگر از منابع احتمالی و در دسترس برای آموزش این سامانه‌ها هستند. در نتیجهٔ استفاده از متون دسترسی‌باز فوق‌الذکر، این سامانه‌ها الگوهای (صحیح یا ناصحیح) رسم‌الخطی و نگارش نویسه‌های سجاوندی را از رسم‌الخط مترجمان و/یا نویسندگان همین متون پالایش‌نشده یادگیری(عمیق) می‌کنند. بنابراین، می‌توان گفت که یافته‌های مطالعهٔ حاضر، تا حدی خطاهای رسم‌الخطی و سجاوندی طیف گسترده (و البته غیرقابل‌تفکیکی) از نویسندگان و مترجمان (حرفه‌ای و غیرحرفه‌ای) را منعکس می‌کند.

با قیاس خطاهای رسم‌الخطی انواع متون فارسی انسانی (بخش ۲ را ببینید) و خطاهای شناسایی‌شده در سامانه‌ها، می‌توان استدلال کرد که این سامانه‌ها و خطاهای رسم‌الخطی و سجاوندی آنها بازتاب کمابیش دقیقی از خطاهای عموم نویسندگان و مترجمان فارسی‌زبان است. البته، معدود خطاهایی در متون انسانی گزارش شده است که نمونه‌ای از آنها در ترجمه‌های این دو سامانه یافت نشد. برای نمونه، مدادیان (۱۴۰۳) درج نیم‌فاصله بین اجزاء ترکیبات صفت‌وموصوفی و مضاف‌مضاف‌الیهی را در ترجمه‌های علمی دانشجویان مقطع کارشناسی گزارش می‌کند، اما در سامانه‌های مورد بررسی خطایی از این دست مشاهده نشد. می‌توان استنباط کرد خطاهایی که در متون فارسی انسانی مشاهده شده‌اند اما در این تحقیق مشاهده نشدند، احتمالاً در متون

1. align

انسانی عمومیّت و فراوانی بسیار پایینی دارند؛ درنتیجه، سامانه‌ها در مرحلهٔ آموزش از آنها الگوبرداری نمی‌کنند و در خروجی آنها بازتولید خطاهای انسانی‌ای که فراوانی پایینی دارند را شاهد نیستیم.

در همین راستا، می‌توان دید که شباهت بسیار نزدیکی بین یافته‌های مطالعهٔ حاضر و یافته‌های بررسی مدادیان (۱۴۰۳) که خطاهای رسم‌الخطی دانشجویان مترجمی زبان انگلیسی را بررسی کرده است وجود دارد. با توجه به محبوبیّت و همه‌گیری استفاده از سامانه‌های ترجمهٔ ماشینی، احتمال دارد که دانشجویان شرکت‌کننده در آن تحقیق نیز از یکی از همین سامانه‌ها برای انجام تمام یا بخشی از تکلیف پژوهشی خود استفاده کرده باشند (کما اینکه وی خود اشاره می‌کند که چند ترجمه را به ظن استفادهٔ دانشجویان از سامانه‌های ترجمه از تحقیق کنار گذارده است). شایان‌ذکر است که سال ۱۴۰۳ سالی بود که تقریباً در تمامی محافل علمی ایران (و جهان) سخن از هوش مصنوعی و کاربردهای متنوع آن (ازجمله ترجمه) بود و افراد به‌صورت گسترده شروع به استفاده از آن برای اهداف مختلف کردند. همچنین، ممکن است دانشجویان از این سامانه‌ها برای ترجمه کمک گرفته باشند و سپس خروجی آنها را قبل‌از تحویل به محقق پس‌ویرایش کرده باشند به‌طوری‌که محقق متوجه استفادهٔ آنها از این سامانه‌ها نشده باشد. بنابراین، شباهت بین یافته‌ای مطالعهٔ حاضر و بررسی مذکور قابل توضیح است.

نتایج مطالعهٔ حاضر نشان داد که ترجمه‌های تمامی سامانه‌ها کمابیش ناهماهنگی‌هایی با دستورالعمل‌های دستور خط مصّوب و اصول نگارش نویسه‌های سجاوندی دارند و از این منظر به درجات مختلفی از پس‌ویرایش انسانی نیاز دارند. با توجه به حجم فزایندهٔ استفادهٔ از این سامانه‌ها در سال‌های اخیر برای ترجمهٔ انواع متون، می‌توان گفت که اگر خطاهای رسم‌الخطی آنها به طرقی برطرف نشوند، می‌توانند به‌مرور زمان و با افزایش تماس احاد فارسی‌زبانان (چه نویسنده و چه مترجم) با خروجی این سامانه‌ها به رسم‌الخط عموم فارسی‌زبانان نَشت کنند. این نشت می‌تواند بر مترجمانی که در تماس مستقیم با این سامانه‌ها هستند و برای ترجمهٔ تکالیف خود (به‌صورت روزافزون) از سامانه‌ها کمک می‌گیرند تأثیر رسم‌الخطی منفی بگذارد. به‌علاوه، استفاده از این سامانه‌ها با وضع موجود، می‌تواند رسم‌الخط خوانندگان فارسی‌زبانی که خروجی کار مترجمان را مطالعه می‌کنند را نیز متأثر کند.

می‌توان راه‌کارهایی را برای هماهنگ‌سازی ترجمه‌های تولیدی سامانه‌ها با

دستورالعمل‌های فرهنگستان و اصول نگارش نویسه‌های سجاوندی پیشنهاد داد تا از وقوع خطاهای رسم‌الخطی در خروجی ترجمه‌ای این سامانه‌ها جلوگیری کرد و نیاز به پس‌ویرایش آنها توسط انسان را کاهش داد. راه‌حل اول آن است که متون تک‌زبانه یا دوزبانهٔ مورد نیاز برای (پیش)آموزش سامانه‌ها را از منظر رسم‌الخطی و کاربست نویسه‌های سجاوندی برپایهٔ دستورالعمل‌های معیار اصلاح کنیم. بهترین حالت آن است که این کار به‌صورت دستی توسط انسان انجام شود، اما با توجه به حجم بالای متون مورد نیاز برای (باز)آموزش سامانه‌ها، انجام این کار با استفاده از یک سامانهٔ مجزا که رسم‌الخط مصّوب را آموزش دیده باشد عملی‌تر به نظر می‌رسد. راه‌کار دیگر آن است که خروجی ترجمه‌ای این سامانه‌ها را توسط یک «بخش کنترل و ویرایش زبان مقصد» ابتکاری در سامانه‌ها اصلاح کنیم تا با رسم‌الخط معیار منطبق شود. همچنین، توسعه‌دهندگان می‌توانند برای آموزش «بخش کنترل و ویرایش زبان مقصد» یا همان «بخش نرمالیزر»[1]، از یک پیکرهٔ کوچک‌تر اما پالایش و تصحیح‌شده توسط انسان مطابق با رسم‌الخط معیار استفاده کنند. به‌علاوه، گنجاندن (استفاده از) فرهنگ املایی فرهنگستان (صادقی و زندی مقدم، ۱۳۹۴) در «بخش کنترل و ویرایش زبان مقصد» نیز می‌تواند هماهنگی آن با دستورالعمل‌های رسم‌الخط معیار را تقویت کند. البته، استفاده همزمان، ترکیبی و ابتکاری از این راهکارها می‌تواند ضعف‌های احتمالی تک‌تک آنها را پوشش دهد. لازم به ذکر است که بیش از یک دهه از انتشار فرهنگ املایی فرهنگستان می‌گذرد و به به‌روزرسانی نیاز دارد کما اینکه ناهماهنگی‌هایی بین صورت‌های مد نظر آن و دستور خط فرهنگستان وجود دارد (مدادیان (۱۴۰۳) را ببینید).

ازآنجایی‌که ممکن است (حداقل در حال‌حاضر) توسعهٔ واحد ترجمهٔ انگلیسی-به-فارسی سامانه‌های ترجمه برای توسعه‌دهندگان آنها دارای جذابیت تجاری نباشد، محققان (و شرکت‌های) ایرانی مسئول می‌توانند «بخش کنترل زبان فارسی» را به‌صورت مستقل از سامانهٔ اصلی به‌صورت یک سامانهٔ بومی مجزا طراحی و توسعه دهند به‌طوری‌که کاربران بتوانند خروجی سامانه‌ها را برای اصلاح رسم‌الخطی و نگارش نویسه‌های سجاوندی به این سامانه واگذار کنند یا اینکه این سامانه‌ها متن موردنظر کاربر را دریافت کنند، آن را برای ترجمه به یکی از سامانه‌های ترجمه ارسال کنند و

1. normalizer

خروجی آنها را دریافت کرده و پس‌از کنترل و اصلاح رسم‌الخطی (به‌صورت محلی) آن را به کاربر نهایی خود تحویل دهند. البته درحال‌حاضر سامانه‌هایی برای اصلاح خطاهای املایی و سجاوندی وجود دارد (پاکنویس، ویراستیار، و حتی گرامرلی[1]). بااین‌حال، لازم است هماهنگی آنها با رسم‌الخط معیار مورد بررسی قرار گیرد. البته، تجربهٔ شخصی نگارنده حاکی از آن است که سامانه‌های موجود هماهنگی ضعیفی با رسم‌الخط معیار دارند.

درنهایت، نقدی بر دستورخط مصّوب داریم و پیشنهاداتی نیز ارائه می‌شود. یکی از مشکلات دستورخط مصّوب اخیر و ویراست‌های قبلی آن دوراهی «پیوسته‌نویسی-نیم‌فاصله‌گذاری» کلمات و ترکیبات است. بااین‌که مشخصاً ترجیح دستورخط مصّوب (۱۴۰۲) و فرهنگ املایی آن (۱۳۹۴) استفاده از نیم‌فاصله بین اجزاء سازندهٔ غالب کلمات و ترکیبات است، گاهاً صورت‌های پیوسته‌نویسی‌شده هم همزمان صحیح قلمداد می‌شوند (*کتاب‌خانه/کتابخانه؛جست‌وجو/جستجو*). استدلال آن است که این صورت‌ها در میان فارسی‌زبان در گذشته و حال مقبولیّت داشته و دارند. جالب است که صورت‌های پیوسته‌نویسی‌شده کلمات و ترکیبات آنچنان در ترجمه‌های تولیدی سامانه‌ها نیز مشاهده نگردید که خود نشان می‌دهد جامعهٔ مترجمان و نویسندگان فارسی زبان معاصر (که متون تولیدی آنها برای آموزش سامانه‌ها به کار گرفته می‌شود) نیز اقبال چندانی به صورت‌های پیوسته در خط فارسی ندارند.

نظر نگارندهٔ این سطور آن است که رویکرد فرهنگستان در به‌رسمیّت‌شناختن بیش از یک صورت برای کلمات واحد خود مانعی دربرابر معیارسازی رسم‌الخط فارسی است. پیشنهاد می‌شود که پیوسته‌نویسی کلاً از رسم‌الخط فارسی نوین کنار گذاشته شود این کار هم خط فارسی را در میان‌مدت یکدست و معیار می‌کند و هم یادگیری آن را برای فارسی‌زبانان و غیرفارسی‌زبانان تسهیل می‌کند. همچنین، مشکلات ذخیره و بازیابی اطلاعات از متون فارسی در محیط‌های رایانه‌ای— به‌ویژه، پیکره‌های متنی— کمتر می‌شود. پیشنهاد دوم آن است که فرهنگ املایی فرهنگستان (۱۳۹۴) نیز با توجه به پیشنهاد قبل بازنویسی و به‌روزرسانی شود. به‌نظر می‌رسد که رویکرد محافظه‌کارانهٔ فرهگستان مبنی بر برگزیدن طریق مصالحه در انتخاب صورت صحیح کلمات و پذیرش بیش از یک صورت صحیح برای برخی کلمات و ترکیبات، درنهایت باعث می‌شود که معیارسازی رسم‌الخط فارسی و املای واژگان آن با چالش‌های بیشتری روبرو شود.

1. https://paknevis.ir/; https://virastyar.ir/; https://www.grammarly.com/

با توجه به یافته‌های این تحقیق و پژوهش‌های پیشین، به‌نظر می‌رسد که همچنان فرهنگستان زبان و ادب فارسی مسیری طولانی برای معیار کردن خط فارسی در میان احاد فارسی‌زبانان دارد و باید برنامه‌ریزی طولانی‌مدت برای این کار انجام شود و هماهنگی خوبی بین این نهاد و دیگر نهادهای کشور ایجاد شود (پیشنهادهای مدادیان (۱۴۰۳) مجموعه‌ای از راه‌کارها را برای نیل به این مقصود ارائه داده است). خطاهای شناسایی‌شده در سامانه‌ها آینهٔ بسیار مناسبی از رسم‌الخط فارسی‌زبانان/غیرفارسی‌زبانان آشنا با فارسی (مترجم و غیرمترجم) در سرتاسر جهان است و می‌تواند در اختیار فرهنگستان قرار گیرد تا با لحاظ آنها دستورالعمل‌های خود را در ویراست‌های جدید خود به‌روزرسانی کند.

۶. نتیجه‌گیری

در این بررسی خطاهای رسم‌الخطی و خطاهای نگارش نویسه‌های سجاوندی در ترجمه‌های مطبوعاتی انگلیسی-به-فارسی پنج سامانهٔ برخط، که از روش‌های مختلف یادگیری عمیق (عصبی، آماری، مبتنی‌بر-نمونه) و داده‌های دوزبانهٔ موازی عظیم برای تولید ترجمه استفاده می‌کنند، مورد بررسی قرار گرفت. نتایج سه دستهٔ عمده از خطاها را در متون ترجمه‌شده نمایان کرد که شباهت قابل‌توجهی به خطاهای شناسایی‌شده در متون و ترجمه‌های انسانی دارند. با توجه به الگوبرداری و یادگیری (ماشینی) این سامانه‌ها از متون انسانی، این شباهت منطقی و قابل‌انتظار است. در میان این سامانه‌ها، دیپ‌سیک از منظر انواع خطاهای تولیدی با اختلاف زیاد بهترین عملکرد رسم‌الخطی را داشت.

علی‌رغم اینکه این مطالعه بر روی تعداد کلمات معدودی در یک ژانر متنی مشخص و تنها در پنج سامانه انجام شد، می‌توان نتایج آن را با احتیاط به دیگر سامانه‌های مشابه نیز تعمیم داد چراکه دیگر سامانه‌ها نیز اساساً از روش‌های یادگیری عمیق مبتنی‌بر داده‌های متنی عظیم (تک‌زبانه و دوزبانه) برای آموزش و الگوبرداری بهره می‌برند و احتمالاً از متون دسترسی‌باز کمابیش یکسان (عمدتاً موجود در اینترنت) برای آموزش خود استفاده می‌کنند. بااین‌حال، لازم است برای تأیید و تقویت نتایج این مطالعه، تحقیقات مشابهی بر روی سامانه‌های دیگر و با لحاظ تعداد کلمات بیشتر و در ژانرهای متنی متنوع انجام شود. همچنین، می‌توان مطالعاتی روی سامانه‌هایی که زیرنویس خودکار تولید می‌کنند (یوتیوب، توئیتر، اینستاگرام و غیره) نیز انجام داد. درنهایت، می‌توان با تجمیع یافته‌های این تحقیقات، گام‌های مؤثرتری در جهت بهبود

خروجی ترجمه‌ای و غیرترجمه‌ای سامانه‌های برخط از منظر رسم‌الخطی و سجاوندی برداشت. همچنین، می‌توان با در نظر گرفتن یافته‌های تجمیعی این تحقیقات، بازنگری‌هایی را که مطابق با ترجیحات جمع کثیری از فارسی‌زبانان باشد در ویراست(های) بعدی دستور خط فرهنگستان انجام داد.

پی‌نوشت‌ها

۱.ویراست اول دستور خط فرهنگستان در سال ۱۳۸۰ منتشر شد و در چاپ چهارم خود در سال ۱۳۸۴ نیز مورد بازنگری‌هایی قرار گرفت. جدیدترین ویراست رسم‌الخط با گذشت بیش‌از دو دهه از انتشار ویراست اول با لحاظ نظرات جمع کثیری از اساتید فرهنگستان و دانشگاه‌های (داخل و خارج) کشور، مدیران و دست‌اندرکاران چاپ، نشر و ویراستاری کشور در سال ۱۴۰۲ با بازنگری‌هایی منتشر شده است.

۲.درج علامت * در انتهای کلمه بدان معناست که آن کلمه از منظر رسم الخطی و طبق دستور خط مصّوب فرهنگستان ناصحیح است.

۳.یکی از داوران محترم مقاله به‌درستی اشاره می‌کند که برخی از این نمونه‌ها دارای نویسهٔ «نیم‌فاصله» نیستند بلکه دو جزء آنها، طبق گفتهٔ دستور خط مصّوب، «بی‌فاصله» از یکدیگر نگارش شده‌اند (*ازنظر، ازسوی، براساس*). در توضیح باید اشاره کنیم که چه بین این اجزاء نیم‌فاصله درج کنیم و چه آنها را بی‌فاصله نگارش کنیم تفاوتی در ظاهر و صورت نهایی آنها ایجاد نمی‌شود. به‌بیان‌دیگر، صورت نیم‌فاصله‌نویسی شده و بی‌فاصله‌نویسی‌شدهٔ آنها از یکدیگر قابل‌تشخیص نیست. درواقع، نیازی به درج نیم‌فاصله بین این اجزاء نیست و خود آنها به‌صورت طبیعی به‌دلیل ویژگی‌های نویسه‌های تشکیل‌دهنده خود امکان چسبیدن به یکدیگر را ندارند.

منابع

احمدی‌نسب، ف.، کاظمی‌فرد، ا. و عظیمی‌فرد، ف. (۱۴۰۱). ارزیابی و رتبه‌بندی میزان التزام شبکه های رسانهٔ ملی به دستورخط فارسی مصّوب فرهنگستان زبان و ادب فارسی با کاربست رهیافتی نوین از نظریهٔ تصمیم‌گیری‌های چندمعیاره. *پردازش و مدیریت اطلاعات*، *۳۷*(۴)، ۱۱۲۷-۱۱۵۲. https://doi.org/10.35050/JIPM010.2022.005

احمدی‌نسب، فاطمه (۱۳۹۴). لزومِ کاربرد دستور خط مصوّبِ فرهنگستان زبان و ادب فارسی (در نشریاتِ علمیِ فارسی، جهت ترویج خط و زبان فارسی به‌عنوان زبان علم). *مجموعهٔ مقاله‌های دهمین همایش بین‌المللی ترویج زبان و ادب فارسی*. دانشگاه محقق اردبیلی.
https://www.sid.ir/paper/843605/fa

اسداللهی، خ. و آذرنیوار، ل. (۱۴۰۰). آسیب شناسی نگارشی و ویرایشی قانون مدنی با تکیه بر دستور زبان فارسی. *پژوهش‌های دستوری و بلاغی*، *۱۱(۲۰)*، ۲۸۷-۳۱۵.
doi: 10.22091/jls.2022.8066.1386

آخشیک، س. (۱۳۹۵). خط و خطا: بازتاب دشواری‌های نگارش کلمه در بازیابی اطلاعات بانک نشریات کشور (مگ ایران). *اولین کنفرانس بین المللی بازیابی تعاملی اطلاعات*.
https://civilica.com/doc/572879

آخشیک، س. و فتاحی، ر. (۱۳۹۱). تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی. *کتابداری و اطلاع‌رسانی*، *۱۵*(۳)، ۹-۳۰.
https://lis.aqr-libjournal.ir/article_42907.html

بساک، ح.، سعادت‌زاده، م. و بساک، ح. (۱۳۹۲). نقد و بررسی نگارشی، ویرایشی و دستوری آرای دادگاههای عمومی جزایی مشهد و شناسایی ضعف‌ها و عوامل اثرگذار بر آنها. *رویهٔ قضایی*، (۴-۵)، ۱۱-۳۶

ذوالفقاری، ح. (۱۳۸۶). *آسیب‌شناسی زبان مطبوعات، فصلنامه علمی رسانه، ۱۸* (۴). ۹-۴۲.
https://dor.isc.ac/dor/20.1001.1.10227180.1386.18.4.1.0

رنجبر، ا.، عباس پور، ج.، ستوده، ه و مولودی، ا. س. (۱۳۹۸). بررسی میزان انطباق رفتار نگارشی نگارندگان و کاربران پایگاه‌های اطلاعات علمی فارسی با دستورالعمل‌های مصوّب فرهنگستان زبان و ادب فارسی در ارتباط با پیوسته‌نویسی، نزدیک‌نویسی و جدانویسی کلمات .*کتابداری و اطلاع‌رسانی، ۲۲(۳)، ۱۶۴-۱۸۷.*
doi: 10.30481/lis.2019.53904

ستوده، ه. و هنرجویان، ز. (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آن بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات. *کتابداری و اطلاع‌رسانی، ۱۵* (۴)، ۵۹-۹۲.
https://lis.aqr-libjournal.ir/article_42651.html

سمایی، ف. (۱۳۸۸). *بررسی مسائل زبان‌شناختی زیرنویس‌های شبکهٔ خبر*. تهران: مرکز تحقیقات صدا وسیما.

شیعه علی، ف. (۱۳۹۲). *نقد و بررسی آیین نگارشی آرای دادگاههای عمومی وحقوقی* مشهد. پایان نامه کارشناسی ارشد رشته زبان و ادبیات فارسی. دانشگاه پیام نور واحد مشهد.

صادقی، ع. ا. و زندی‌مقدم، ز. (۱۳۹۴). *فرهنگ املایی خط فارسی براساس دستور خط فارسی*. نشر آثار: تهران.

فرهنگستان زبان و ادب فارسی (۱۴۰۲). *دستور خط فارسی*. نشر آثار: تهران

کشاورز، پ. و ستاری، ع. (۱۳۹۰). *کاربرد خط فارسی در تلویزیون*. تهران: مرکز تحقیقات صدا و سیما.

کلاهدوزان، ا.، معینی، م.، پاپی، ا.، عسگری، غ. و ذولفقاری، ب. (۱۳۸۳). بررسی توزیع فراوانی عدم رعایت اصول و قواعد نگارش فارسی در پایان‌نامه‌های کارشناسی ارشد ناپیوسته و دکترای دانشکده‌های پزشکی و داروسازی در سال ۱۳۷۸-۷۹. *مدیریت اطلاعات سلامت، ۱* (۲)، ۵۰-۵۶.
https://him.mui.ac.ir/article_10859.html

مدادیان، غ. (۱۴۰۳). بررسی خطاهای رسم‌الخطی و نشانه‌گذاری سجاوندی در ترجمه‌های علمی ترجمه‌آموزان مقطع کارشناسی برپایهٔ دستور خط مصوّب فرهنگستان زبان و ادب فارسی. *پژوهش‌نامه آموزش زبان فارسی به غیرفارسی‌زبانان، ۱۳(۱)، ۱۶۱-۲۰۴.*
https://doi.org/10.30479/jtpsol.2024.20476.1669

مدرس خیابانی، ش. (۱۳۹۷). آسیب‌شناسی نگارشی زیرنویس‌ها در شبکه‌های تلویزیونی خبر و آی‌فیلم: پژوهشی پیکره‌بنیاد. *رسانه‌های دیداری و شنیداری، ۱۲* (۲۷)، ۳۱-۶۰.
https://dorl.net/dor/20.1001.1.26454696.1397.12.27.3.4

هاشمی، س. ح. (۱۳۹۰). بررسی کتاب‌های درسی سال تحصیلی ۱۳۸۷-۸۸ دوره ابتدایی از نظر میزان هماهنگی در جدایی یا اتصال پایه واژه‌های فعلی و غیرفعلی در کلمات مرکب. *پژوهش زبان و ادبیات فارسی*، ۹ (۲۱/۳)، ۱-۱۰. https://www.sid.ir/paper/56353/fa

References

Academy of Persian Language and Literature (2023). *Persian Orthography*. Nashr-e Asar: Tehran. [in Persian]

Ahmadinasab, F., Kazemifard, A., & Azimifard, F. (2022). Evaluation and Ranking of the National Media Networks' Adherence to the Persian Orthography Approved by the Academy of Persian Language and Literature Using a Novel Approach of Multi-Criteria Decision-Making Theory.

Information Processing and Management, 37(4), 1127-1152. [in Persian] https://doi.org/10.35050/JIPM010.2022.005

Ahmadinasab, Fatemeh (2015). The Necessity of Applying the Orthography Approved by the Academy of Persian Language and Literature (in Persian Scientific Journals, to Promote Persian Script and Language as the Language of Science). *Collection of Articles from the 10th International Conference on the Promotion of Persian Language and Literature*. University of Mohaghegh Ardabili. [in Persian] https://www.sid.ir/paper/843605/fa

Akhshik, S. (2016). Script and Error: The Reflection of Difficulties in Word Writing on Information Retrieval in the Country's Magazines Database (Mag Iran). *The First International Conference on Interactive Information Retrieval*. [in Persian] https://civilica.com/doc/572879

Akhshik, S., & Fattahi, R. (2012). Analyzing the Challenges of Conjoined and Separate Writing of Persian Words in Information Storage and Retrieval in Databases. *Library and Information Science*, 15(3), 9-30. [in Persian] https://lis.aqr-libjournal.ir/article_42907.html

Asadollahi, Kh., & Azarniyavar, L. (2021). Pathological Study of the Writing and Editing of the Civil Code Based on Persian Grammar.*Grammatical and Rhetorical Research*, 11(20), 287-315. [in Persian] doi: 10.22091/jls.2022.8066.1386

Bassak, H., Sa'adatzadeh, M., & Bassak, H. (2013). A Critical Study of the Writing, Editing, and Grammatical Aspects of Verdicts from Public Criminal Courts of Mashhad and Identification of Their Weaknesses and Influential Factors. *Judicial Procedure*, (4-5), 11-36. [in Persian]

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, P., & van den Heuvel, H. (2022). A taxonomy of terminological errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 47–56). European Association for Machine Translation. https://aclanthology.org/2022.eamt-1.6

Cintas, J. D., & Remael, A. (2020). *Subtitling: Concepts and practices*. Routledge, London and New York.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. 10.30479/jtpsol.2024.20476.1669

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, *22*(120), 1–39.

Guzmán, F., Chen, X., & Orăsan, C. (2019). A multifaceted comparison of translation paradigms and their effects on punctuation. *Machine Translation*, 33(3), 205–230.https://doi.org/10.1007/s10590-019-09232-x

Hammo, B. H. (2009). Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. *Information Retrieval 12*(3): 300-323. http://link.springer.com/article/10.1007/s10791-008-9081-9.

Hashemi, S. H. (2011). An Investigation of Textbooks for the 2008-09 Academic Year in Elementary Education Regarding the Level of Consistency in the Separation or Connection of Verbal and Non-Verbal Stems in Compound Words. *Research in Persian Language and Literature*, 9(3/21), 1-10. [in Persian] https://www.sid.ir/paper/56353/fa

Keshavarz, P., & Sattari, A. (2011). *The Use of Persian Script on Television*. Tehran: Radio and Television Research Center. [in Persian]

Kolahdoozan, A., Moeini, M., Papi, A., Asgari, Gh., & Zolfaghari, B. (2004). Investigating the Frequency Distribution of Non-compliance with Persian Writing Principles and Rules in Master's and PhD Theses of Medical and Pharmacy Schools in the Academic Year 1999-2000. *Health Information Management*, 1(2), 50-56. [in Persian] https://him.mui.ac.ir/article_10859.html

Lazarinis, F. (2007). At the sharp END evaluating the searching capabilities of commerce websites in a non-English language A Greek case study. *Online Information Review, 31*(6): 881-891. http://www.emeraldinsight.com/journals.htm?articleid=1640585.

Lazarinis. (2008). Improving concept-based web image retrieval by mixing semantically similar Greek queries. *Program: electronic library and information systems, 42*(1), 56-67. http://www.emeraldinsight.com/journals.htm?articleid=1674242.

Lewandowski, D. (2008). Problems with the use of Web search engines to find results in foreign languages. *Online Information Review 32*(4): 668-672. http://www.emeraldinsight.com/journals.htm?articleid=1747662.

Madadian, Gh. (2024). A Study of Orthographic and Punctuation Errors in the Translations of Undergraduate Translation Students Based on the Orthography Approved by the Academy of Persian Language and Literature. *Journal of Teaching Persian to Speakers of Other Languages*, 13(1), 161-204. [in Persian] https://doi.org/10.30479/jtpsol.2024.20476.1669

Microsoft Research. (2023). Z-Code++: A Pre-trained Language Model Optimized for Abstractive Summarization. *Microsoft Research Blog*

Modarres Khiabani, Sh. (2018). A Pathological Study of Subtitles on News and iFilm TV Channels: A Corpus-Based Research. *Audiovisual Media*, 12(27), 31-60. [in Persian] https://dorl.net/dor/20.1001.1.26454696.1397.12.27.3.4

Monz, C. & De Rijke, M. (2002). *Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian*. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross Language Evaluation Forum, CLEF 2001, Darmstadt, Germany.

Moukdad, H. (2005). Lost in cyberspace: How Do Search Engines Handle Arabic Queries? *The international information & library review*, *37*(4): 237-394. https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/334/282

Popović, M. (2018). Error classification and analysis for machine translation quality assessment. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 195–204). European Association for Machine Translation. https://aclanthology.org/W18-1920/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Ranjbar, A., Abbas Pour, J., Sotoudeh, H., & Mauludi, A. S. (2019). Investigating the Level of Compliance of the Writing Behavior of Authors and Users of Persian Scientific Information Databases with the Guidelines Approved by the Academy of Persian Language and Literature Regarding Conjoined, Hyphenated, and Separate Writing of Words. *Library and Information Science*, 22(3), 164-187. [in Persian] doi: 10.30481/lis.2019.53904

Sadeghi, A. A., & Zandimoghadam, Z. (2015). *Persian Orthographic Dictionary Based on Persian Orthography*. Nashr-e Asar: Tehran. [in Persian]

Samaie, F. (2009). *An Investigation of the Linguistic Issues in the Subtitles of the News Network*. Tehran: Radio and Television Research Center. [in Persian]

ShiaAli, F. (2013). *A Critical Study of the Writing Style of Verdicts from Public and Civil Courts of Mashhad* (Master's thesis in Persian Language and Literature). Payame Noor University, Mashhad Branch. [in Persian]

Sotoudeh, H., & Hanarjuyan, Z. (2012). A Review of the Difficulties of the Persian Language in the Digital Environment and Its Impact on the Effectiveness of Automatic Text Processing and Information Retrieval. *Library and Information Science*, 15(4), 59-92. [in Persian] https://lis.aqr-libjournal.ir/article_42651.html

Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search Engine for Various Queries. *Libri, 56*, 38-47. https://www.degruyter.com/document/doi/10.1515/LIBR.2006.38/html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998-6008.

Zhang, J., & Suyu, L. (2007). Multiple language supports in search engines. *Online Information Review, 31*(4): 516-532. http://www.emeraldinsight.com/journals.htm?articleid=1621798.

Zolfaghari, H. (2007). Pathology of the Language of the Press. *Scientific Quarterly of Media*, 18(4), 9-42. [in Persian] https://dor.isc.ac/dor/20.1001.1.10227180.1386.18.4.1.0.