عنوان زیرپروژه:

مقدمه‌ای بر طراحی و ایجاد خطایاب املایی صرفی زبان فارسی

فهرست مطالب

عنوان — شماره صفحه

1. **مقدمه** 4

2. **اشتباهات املایی** 6

2-1. هم‌آواها 7

3. **کارهای انجام گرفته** 10

4. **چالش‌های زبان فارسی** 13

4-1. مشکلات رسم‌الخط و دستور نگارش 13

4-2. حروف هم‌شکل 15

4-3. ریخت‌شناسی 16

4-3-1. قوانین ساخت و صرف افعال 16

4-3-2. قوانین ترکیب وندها 16

4-3-3. قوانین فاصله‌گذاری 17

4-4. هم‌آواها 22

4-5. توزیع انواع خطاهای املایی 22

5. **روش‌های مرتب‌سازی و انتخاب پیشنهادات** 24

5-1. روش ساده فاصله حروف 24

5-2. بسامد کلمات 25

5-3. فاصله ویرایشی کمینه 25

5-3-1. روش همینگ 26

5-3-2. روش لونشتاین 26

5-3-3. روش دامرو-لونشتاین 27
5-3-4. روش وگنر-فیشر 28
5-3-5. روش هیرشبرگ 28

عنوان **شماره صفحه**

5-3-6. روش اوکونن 29
5-3-7. روش جرو-وینکلر 29
5-3-8. محاسبه فاصله میان حروف 29

6. نتیجه‌گیری 32

7. مراجع 33

1. مقدمه

با آغاز استفاده‌ی عمومی از رایانه در جامعه‌ی بشری، چالش‌های بسیاری پیرامون پردازش زبان‌های طبیعی به وجود آمد. خطا در کار با زبان‌های طبیعی امری است ناگزیر که متخصصان را بر آن داشت برای رفع مشکلات موجود در این زمینه کوشش‌هایی را انجام دهند. از جمله‌ی این مسائل خطاهای املایی و نگارشی کاربران رایانه‌ای است که می‌تواند حاصل دلائل مختلفی مانند اشتباه در تحریر و یا کم‌اطلاعی کاربران از زبان مورد استفاده باشد. بسیاری از خطاها از دید کاربران پنهان می‌ماند و بسیاری دیگر خطاهایی هستند که کاربران به اشتباه گمان بر درست بودنشان دارند. انسان‌ها معمولا در هنگام گفتگو و یا نوشتن اشتباهاتی را در چهار سطحِ 1) لغوی[1]، 2) نحوی[2]، 3) معنایی[3]، و 4) مبتنی بر زمینه[4]، انجام می‌دهند [1, 2]. غلط‌های املایی در زمینه‌های مختلفی از جمله، 1) غلط‌های تایپی در زمینه نمایه‌سازی و سیستم‌های بازیابی اطلاعات [3]، 2) تشخیص اشتباه متون نوشته شده [4, 5]، 3) غلط‌های املایی در متون علمی و دانشگاهی [6]، و 4) غلط‌های املایی ناخودآگاه در نوشتار کودکان [7, 8].

غلط‌یاب‌های املایی معمولا به تصحیح غلط‌های تایپی[5] و نگارشی[6] می‌پردازند [9]. غلط‌های تایپی معمولا ناشی از اشتباهات معمول تایپی است تا ناشی از عدم آگاهی نگارنده. به عنوان مثال غلط املایی ممکن است به علت جابجایی یک حرف با حروف کناری خود در صفحه‌کلید به وجود آمده باشد [3, 6, 9]. غلط‌های نگارشی به علت ناآگاهی نگارنده از قوانین و واژگان زبان مانند حدس زدن املای کلمه، نگارش کلمه از روی تلفظ آن و یا انتخاب کلمه اشتباه (مثلا استفاده از «همچنین» به جای «همچنان») به وجود می‌آیند [9].

---

[1] Lexical

[2] Syntactic

[3] Semantic

[4] Contextual

[5] Typographical

[6] Orthographical

در این مستند به بررسی روش‌های خطایابی املایی، چالش‌های زبان فارسی در زمینه اشتباهات املایی می‌پردازیم.

2. اشتباهات املایی

تحلیل و بررسی اشتباهات تایپی در پرونده‌های متنی بسیار بزرگ [3, 10] منتج به این گشت که بخش اعظمی (80 درصد تا 90 درصد) از کلمات غلط با معادل صحیح املایی خود در 1) حذف شدن یک حرف، 2) تعویض یک حرف با حرفی دیگر، 3) جابجایی دو حرف کناری از یک کلمه، و 4) درج یک حرف اضافی در کلمه، تفاوت دارند [11]. محققان همچنین دریافتند که معمولا حرف اول از کلمات صحیح تایپ می‌شوند [12, 13]. برای روشن‌تر شدن بحث به بررسی چهار نوع خطای فوق در مورد کلمه «صلح» خواهیم پرداخت:

1) حذف یک حرف — صح
2) تعویض یک حرف — صاح
3) درج یک حرف — صلخح
4) جابجایی دو حرف مجاور — صحل

در حقیقت در تمامی مثال‌های فوق فاصله ویرایش[1] معدل یک است. فاصله ویرایشی نمایانگر حداقل تعداد درج حروف، حذف حروف، تعویض حروف و یا جابجایی حروف مجاور است که برای تبدیل یک رشته (کلمه) به رشته‌ای دیگر مورد نیاز است [14].

یکی از راه‌های سنتی در اصلاح اشتباهات تایپیT ایجاد تغییرات در کلمه دارای اشتباه املایی و جستجوی کلمات جدید حاصله در لیست واژگان صحیح زبان است تا کلمات صحیح محتمل پیدا شوند [15, 16]. در مرحله بعدی تشخیص این که هر یک از این کلمات استخراج شده چه میزان برای انتخاب به عنوان بهترین پیشنهاد برای کلمه‌ی اشتباه مناسب هستند، اهمیت می‌یابد [17].

[1] Edit Distance

سیاست‌های گوناگونی برای چگونگی انتخاب بهترین پیشنهادات موجود است. از جمله‌ی این روش‌ها می‌توان به: 1) استفاده از بسامد کلمات [4]، 2) فاصله‌ی نویسه‌ها[1] [18]، 3) بسامد پیشنهادات و میزان رواج گونه‌های مختلف خطا [10]، 4) احتمال ارضاء قوانین خطا [12]، و 5) روش‌های ترکیبی، اشاره کرد.

2-1. هم‌آواها

هم‌آواها[2] کلماتی هستند که همانند یکدیگر تلفظ می‌شوند. هم‌آواها ممکن است املای یکسانی داشته باشند – مانند شیر (حیوان) و شیر (نوشیدنی)، و یا املای متفاوتی داشته باشند – مانند *اسیر* (زندانی) و *اثیر* (صدای قلم بر کاغذ). هم‌آواها ممکن ایت کلماتی صحیح باشند و یا کلماتی با املای اشتباه - مانند *اهراز* به جای *احراز*.

این گونه از اشتباهات املایی بیشتر هنگامی که افراد به زبانی غیر از زبان مادریشان می‌نویسند رخ می‌دهد. زبان فارسی حجم بسیاری از حروف با آوای یکسان را شامل می‌شود که یکسان تلفظ می‌شوند اما نگارشی متفاوت دارند. این حروف در حقیقت هم‌آوا هستند اما هم‌شکل[3] نیستند. وجود تعداد زیادی از این حروف در زبان فارسی تعداد اشتباهات املایی هم‌آوا را در زبان فارسی حتی برای افراد تحصیل کرده که فارسی زبان مادریشان هست، نیز افزایش می‌دهد. این حروف هم‌آوا به شرح زیر طبقه‌بندی می‌شوند:

1) حروف خانواده الف — {«آ» و «ا»}
2) حروف خانواده همزه — {«ء» و «أ» و «و» و «ؤ» و «ی» و «ئ»}
3) حروف خانواده ته — {«ت» و «ط»}
4) حروف خانواده سین — {«س» و «ص» و «ث»}
5) حروف خانواده هه — {«ه» و «ح»}
6) حروف خانواده زین — {«ز» و «ض» و «ظ» و «ذ»}
7) حروف خانواده قاف — {«ق» و «غ»}

[1] Character Distance
[2] Homophone
[3] Homograph

بنابر این کلمه‌ای مانند «دغدغه» می‌تواند به سه گونه‌ی «دقدغه»، «دغدقه» و «دقدقه» نوشته شود و یا کلمه‌ی «استثنایی» می‌تواند تعداد بسیار زیادی - در حدود 432 عدد، کلمه‌ی هم‌آوا داشته باشد.

با رجوع به تحلیل‌های پیکره‌ای، مطالعات آماری و تجارب عملی در غلط‌یابی املایی در زبان فارسی، فاصله ویرایشی پیشنهادی در تولید پیشنهادات برای یک کلمه‌ی اشتباه معمولی معادل 2 در نظر گرفته شود و برای کلمات هم‌آوای دارای اشتباه املایی بدون محدودیت. بنابراین توصیه می‌شود که پیشنهادات هم‌آوا بدون محدودیت جداگانه محاسبه و تولید شوند. همچنین ابتدا پیشنهادات با فاصله ویرایشی 1 ایجاد شوند و مورد بررسی قرار گیرند و سپس در صورت عدم حصول نتیجه‌ی مطلوب به تولید پیشنهادات با فاصله ویرایشی 2 اقدام گردد [19].

نکته‌ی دیگری که باید مورد نظر قرار گیرد، پیچیدگی محاسباتی و زمانی بالای تولید پیشنهادات با این روش‌ها است که از مرتبه‌ی $O(n^2)$ است. بنابر این نکته‌ای مانند طول کلمه اشتباه، تعداد حروف یک زبان و فاصله ویرایشی مورد نظر تاثیر بسیار زیادی بر حجم محاسبات و زمان پاسخ‌گویی[1] سیستم خواهد داشت که حتی با بهینه‌سازی روش‌ها هم نمی‌توان کاهش چشم‌گیری در زمان پاسخ گویی ایجاد نمود. بنابراین سعی بر کاهش هر یک از مولفه‌های موثر مذکور می‌تواند بهترین راه برای کاهش این پیچیدگی‌ها باشد.

تعیین یک طول بیشینه برای کلمات دارای اشتباه املایی که اقدام بر اصلاح آن‌ها می‌شود ضروری است. با توجه به میانگین طول کلمات در زبان فارسی و انحراف از معیار طول کلمات، می‌توان یک طول کلمه بیشینه را انتخاب نمود (بین 13 تا 17 حرف) که اگر کلمه‌ای با بیش از این تعداد حرف در لیست کلمات صحیح زبان موجود نبود، دیگر اقدام به تصحیح آن نشود و تنها غلط بودن آن گزارش گردد. البته می‌توان در نسخه‌های پیشرفته‌تر تنها به تصحیح این کلمات در فاصله ویرایشی 1 و با انتخاب گزینشی برخی حروف و نه همه‌ی آن‌ها، اقدام به تولید پیشنهادات کرد تا این گونه حجم محاسبات کاهش یابد.

نکته‌ی دیگری که بسیار مورد اهمیت است، ساختار داده‌ی[2] نگهداری کلمات صحیح زبان است. این ساختار می‌باید سرعت جستجوی بسیار بالایی و ترجیحا از مرتبه $O(1)$ داشته باشد. همچنین در

[1] Response Time

[2] Data Structure

نسخه‌های تعاملی، می‌باید از به روز رسانی زنده[۱] پشتیبانی کند. حجم و فضای مورد نیاز این ساختار داده بر روی رسانه‌های فیزیکی پایا[۲] و نیز رسانه‌های فرار[۳] (مانند حافظه) از اهمیت بالایی برخوردار است و می‌باید در حدی معقول (کمتر از 200 مگابایت برای کلیه‌ی کلمات و دیگر اطلاعات) باشد و هرچه این فضا کاهش یابد، پایائی و قالبیت اطمینان و کیفیت خدمات غلط‌یابی در سطوح بهتری قابل تضمین خواهد بود. ارائه خدمات تکمیل خودکار[۴] کلمه توسط این سامانه نیز مطلوب است.

[۱] Live Update

[۲] Persistence

[۳] Volatile

[۴] Auto Complete

3. کارهای انجام گرفته

پیلوتی و همکارش [20] با تحلیل روان‌شناختی بر روی مسئله خطاهای املایی به این امر اشاره کرده‌اند که انسان‌ها معمولاً عاجز از پیدا کردن خطاهای تایپی و نگارشی‌ای هستند که خودشان در متن مرتکب شده‌اند. دلیل اصلی که آن‌ها بر آن تأکید داشتند، رویکرد بالا به پائین افراد نسبت به متون خودشان بوده است. در نتیجه همین مسائل خطایاب‌های املایی هم‌زمان با به وجود آمدن ویرایش‌گرهای متن‌های رایانه‌ای به وجود آمدند. دامرا [3] جزء اولین کسانی بود که در این زمینه به پژوهش پرداخت و پس از او بسیاری از افراد در این زمینه روش‌های مختلفی را محک زدند که کوکیچ [21] در مقاله‌ی خود به صورت مشروح به این فعالیت‌ها اشاره کرده است. کوکیچ در مقالهٔ خود خطاهای املایی را به سه گونه‌ی 1) مجزا[1]، 2) غیر مجزا[2] و 3) حساس به متن[3] تقسیم‌بندی کرده است. معمولاً در مورد خطاهای نوع دوم و سوم مبحث تحلیل نحوی و معنایی هم به میان می‌آید ولی در مورد خطاهای نوع اوّل بیشترین اولویت در پیدا کردن تصحیح‌های املایی مناسب و پس از آن رده‌بندی آن‌ها است؛ به گونه‌ای که تصحیحی که واقعاً دارای صحت باشد؛ در بالاترین رده قرار بگیرد. معیار اصلی که در بسیاری از خطایاب‌ها مدنظر قرار گرفته این است که تصحیح‌ها یا پیشنهادهای درست در رده‌ی اوّل پیشنهادها به کاربران باشد و درصد حضور پیشنهادهای درست در رده‌ی اوّل جزء مهمترین معیارها بوده است. کما اینکه گارفینکل و همکارانش [22] در خطایابی که طراحی کرده بودند، هدف را بر همین معیار قرار دادند که پیشنهاد درست در بالاترین ردهٔ ممکن قرار بگیرد. البته در آن کار با معیارهای حاشیه‌ای دیگری همچون پیچیدگی زمانی دسترسی به واژه‌نامه، گذاشتن حد برای ارائه‌ی تصحیح‌های با فاصلهٔ تصحیح بالاتر از یک و میزان حضور پیشنهاد درست در بین پیشنهادها هم در نظر گرفته شده بود. کارهای دیگری هم روی پیدا کردن تصحیح‌ها با استفاده از روش‌هایی مانند شبکه‌های عصبی مصنوعی [23]، کانال نویزی [24] و مدل‌های چند-وزنی [23, 25] انجام شده است.

[1] Isolated

[2] Non-word/Isolated

[3] Context dependent

در زبان‌های بر پایه‌ی خط عربی مانند فارسی و اردو هم کارهایی انجام شده است. از جملهٔ این کارها، کاری است که نسیم و همکارش [26] برای زبان اردو (پاکستانی) از روش‌های نزدیکیِ صوت[1] و شکل[2] استفاده کرده‌اند. آن‌ها در این روش به فراخوانیِ[3] 94.6% رسیده بودند.

در مورد زبان فارسی هم کارهایی انجام شده است که از آن جمله می‌توان به کار درّی نوگورانی و صبوریان [27] به طراحی خطایابی در زبان فارسی پرداخته‌اند امّا در کاری که انجام داده بودند نه در مورد جزئیات روش اشاره‌ای و نه نتیجه‌ای را اعلام کردند.

البته قاسمی‌زاده و همکارانش [28] به روشی یادگیرنده به استفاده از ساختار درخت ترنری اشاره کردند که داده‌هایشان را روی مجموعهٔ آزمایشی دو زبان فارسی و انگلیسی مورد آزمون قرار دادند. که با توجه به آن به دقتی[4] بین 75% تا 85% رسیدند. البته داده‌های آزمون آن‌ها برای زبان فارسی به صورت واژگان خطادار جدا از هم بوده است و تعریفی که آن‌ها از دقت داده بودند با تعریف معمول دارای تفاوت‌های جزئی بوده است.

مختاری‌پور و همکارش [29] ریشه‌یابی برای زبان فارسی پیشنهاد داده بودند که ادعا می‌کردند برای ریشه‌یابی واژگان فارسی کارایی مناسبی دارد.

کومیو و همکارش [30] بر مبنای چنین تفکری، فکر طراحی یک خطایاب را مطرح کردند که بدون داشتن واژه‌نامه و با وجود یک پیکره‌ی متنی قابل به‌روزآوری بر مبنای بسامد حضور واژگان در آن پیکره و همچنین توابع احتمالی امکان هر نوع خطا به تصحیح خطاهای املایی بپردازد.

رسولی و مینایی [31] در مقاله‌شان پیشنهاد طراحی یک خطایاب پویا در زبان فارسی بر مبنای متن مورد استفاده هر کاربر داده بودند و مبنای آن‌ها بر این بود که علاوه بر احتمال خطا، به واژگان درون متن وزن بیشتری برای پیشنهاد صحیح بودن بدهند. نقصی که در آن کار وجود داشت، محور قرار دادن واژگان درون متن به تنهایی بود. این امر باعث می‌شد که بسیاری از پیشنهادهای صحیح نادیده انگاشته شوند و بسیاری از پیشنهادهایی که ربط زیادی به اصل واژه نداشتند، صرفاً به دلیل حضور در متن حاضر در رده‌های بالا قرار بگیرند.

---

[1] Soundex

[2] Shapex

[3] Recall

[4] Precision

کاشفی و همکارانش در غلط‌یاب‌شان با نام «ویراستیار»، یک غلط‌یاب املایی محاوره‌ای فارسی ارائه کردند که تنها کار فارسی قابل توجه در خصوص غلط‌یابی املایی در لایه‌ی ریخت‌شناسی[۱] زبان فارسی است. این غلط‌یاب علاوه بر امکان تصحیح غلط‌های املایی مجرد، هم‌آواها و هم‌شکل، به اصلاح آگاه از زمینه‌ی[۲] غلط‌های املایی، اصلاح تکرار تایپی، اصلاح کلیه‌ی حالات فاصله‌گذاری اشتباه در ریخت‌شناسی فارسی، اصلاح کاربردهای نادرست «شبه-فاصله[۳]»، و اصلاح با کمک ریشه‌یابی[۴] کلمات با پس‌وند[۵] و پیش‌وند[۶] و ریشه‌یابی افعال است. کاشفی و همکارانش [19] همچنین شیوه‌های نوینی از مرتب‌سازی پیشنهادات برای زبان فارسی با کارایی بسیار بالا در زبان فارسی ارائه داده‌اند.

---

۱ Morphology

۲ Context-aware

۳ Pseudo -space

۴ Stemming

۵ Prefix

۶ Suffix

4. چالش‌های زبان فارسی

هر زبانی ویژگی‌ها، محدودیت‌ها و توانایی خاص خود را دارد. در حوزه‌ی غلط‌یابی املایی مطالعه مدل زبانی و خصوصیات ویژه‌ی آن به منظور طراحی و اجرای غلط‌یاب مناسب آن زبان امری ضروری است. زبان فارسی پنج چالش عمده را در این حوزه شامل می‌شود: 1) مشکلات خط و دستور نگارش، 2) وجود حروف هم‌شکل بسیار، 3) ریخت‌شناسی پیچیده، 4) تعداد زیاد غلط‌های ناشی از هم‌آوایی، و 5) توزیع متفاوت گونه‌های مختلف خطاهای املایی.

4-1. مشکلات رسم‌الخط و دستور نگارش

یکی از مشکلات اساسی در زبان فارسی وجود رویکردها و معیارهای متفاوت و بعضاً متضاد در مورد نحوه‌ی نگارش است؛ همچنین برای طراحی یک خطایاب مناسب برای زبان فارسی نیاز به یک ریشه‌یاب واژگانی است تا بتوان اشتقاق‌های مختلف برای هر واژه را شناخت اما موارد استثناء که با ریشه‌یابی ماشینی مطابقت ندارند در فارسی بسیار است. مشکل اساسی‌ای که در زبان فارسی وجود دارد وجود کاربران بسیاری است که بر اساس زبان فارسی معیار و دستور زبان و نگارش مصوب فرهنگستان علوم، نگارش نمی‌کنند. به عنوان نمونه می‌توان به آمار هنگفت وبلاگ‌ها و روزنوشت‌های اینترنتی موجود در بین ایرانیان فارسی‌زبان اشاره کرد. نه تنها مشکل خارج از معیار نویسی در بین عامه‌ی کاربران وجود دارد، بلکه به دلیل اختلاف نظر و سلیقه بین نویسندگان و محققان بسیاری از کاربران در زبان فارسی از نوع نگارش و واژگانی استفاده می‌کنند که به زعم زبان معیار نادرست است ولی عموم افراد آن را درست می‌دانند. در ضمن به دلیل گستردگی ورود واژگان بیگانه و جدید به زبان فارسی امکان استفاده از یک واژه‌نامه در یک خطایاب املایی می‌تواند باعث بروز مشکلاتی شود.

همان‌طور که عنوان شد، به دلیل تغییر دستور خط در نظام آموزش کشور، اکثر نویسندگان معمولی زبان فارسی، با یکدیگر توافقی در نحوه‌ی نگارش کلمات ندارند. در نتیجه متونی که توسط این گروه تولید می‌گردد (که عمدتاً شامل متون غیر رسمی وب‌نوشت‌های فارسی است)، هیچ سبک و قانون خاصی را دنبال نمی‌کنند.

از این دسته از کاربران که بگذریم، کاربرانی را خواهیم داشت که در مکان‌های غیر رسمی مشغول تایپ متون رایانه‌ای هستند. این کاربران اکثراً دستور خط فارسی را در اختیار دارند اما متنی که هر کدام از آن‌ها تولید می‌کند از لحاظ کیفیت ویرایشی، با دیگری متفاوت است. در واقع خیلی از آن‌ها خیلی از اصول اصلی ویرایشی را با وجود آگاهی از آن رعایت نمی‌کنند. ساده‌ترین مثال، قانون استفاده از تنوین است. البته این میزان گستردگی در رعایت نکردن این قانون، شاید دلیلی بر لزوم بازنگری در این قانون باشد. در این زمینه فصول بعدی توضیحات کامل‌تری خواهند داشت.

ویرایش‌گران حرفه‌ای که جمعیت بسیار اندکی دارند، با وجود آشنایی با دستور خط فارسی و اصول ویراستاری، توافق کلی در نگارش لغات ندارند. مثلاً ممکن است که یک ویراستار خاص تمام "آن‌ها"های خود را به صورت "آنها" بنویسد (یا بالعکس) زیرا دستور خط فارسی در این زمینه توصیه‌ی خاصی ندارد. این کار باعث می‌شود تا نتیجه‌ی تحلیل متنی که یک ویراستار حرفه‌ای به رایانه داده است، مانند نتیجه‌ی تحلیل متن مشابهی که ویراستار حرفه‌ای دیگری آن را تولید کرده است نباشد.

به طور خلاصه از دید رایانه‌ای، می‌توان ایرادهایی که بر دستور خط فارسی فرهنگستان زبان و ادب فارسی وارد است را چنین دسته‌بندی کرد:

- بازگذاشتن دست نویسندگان در فاصله‌گذاری میان کلمات.
- عدم وجود دستورالعمل قطعی برای استفاده از نیم‌فاصله.
- عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات؛ استفاده از دستورالعمل مبتنی بر لغت (مانند تک‌هجایی بودن، بسیط‌گونه بودن).

البته حتی اگر تمام این مشکلات حل شوند و ابهامات برطرف گردد، همچنان در زبان فارسی کاربرانی خواهند بود که لغات را خارج از این استاندارد می‌نویسند که بدون تحولات بنیادین در زبان فارسی مشکلات عدم هم‌خوانی همواره وجود خواهند داشت.

2-4. حروف هم‌شکل

اکثر حروف زبان فارسی چهار حالت متفاوت دارند: 1) فرم مجرد (مانند "ک")، 2) فرم آغازین (مانند "کـ")، 3) فرم میانی (مانند "کـ")، و 4) فرم پایانی (مانند "ک"). برخی از حروف فارسی در چند و یا تمامی حالات هم‌شکل هستند. حروف هم‌شکل فارسی شامل موارد زیر هستند:

1) الف و همزه در تمامی حالات – {«ا» و «أ»}
2) ب و پ در تمامی حالات – {«ب» و «پ»}
3) ت و ث در تمامی حالات – {«ت» و «ث»}
4) جیم، چه و حه در تمامی حالات – {«ج» و «چ» و «ح»}
5) حه و خه در تمامی حالات – {«ح» و «خ»}
6) دال و ذال در تمامی حالات – {«د» و «ذ»}
7) را و زین و ژه در تمامی حالات – {«ر» و «ز» و «ژ»}
8) طا و ظا در تمامی حالات – {«ط» و «ظ»}
9) صاد و ضاد در تمامی حالات – {«ص» و «ض»}
10) عین و غین در تمامی حالات – {«ع» و «غ»}
11) کاف و گاف در تمامی حالات – {«ک» و «گ»}
12) فه و قاف در حالات میانی و آغازین– {«فـ» و «قـ» و «ف» و «ق»}

حروف فوق، حروفی هستند که از نظر ظاهری خصوصا در متون تایپی، بسیار شبیه به هم هستند و هنگامی که از قلم‌های در اندازه کوچک و یا قلم‌هایی با طراحی ضعیف استفاده شود، حتی در بازخوانی متون توسط انسان نیز قابل تشخیص نیستند. اهمیت این چالش با دقت بر این نکته که اکثر این حروف مشابه، حروف مجاور یکدیگر نیز هستند که احتمال رخداد جابجایی آن‌ها به جای یکدیگر را افزایش می‌دهد.

این نوع از اشکالات خصوصا در کاربردهای خطایابی به عنوان پس‌پردازشی برای نویسه‌خوانی نوری متون فارسی از اهمیت ویژه‌ای برخوردار است و می‌باید این‌گونه حروف با سیاست‌هایی خاص، جای‌گذاری و غلط‌یابی شوند.

3-4. ریخت‌شناسی

از چالش‌های مهم دیگر فارسی، ریخت‌شناسی پیچیده و مبهم آن است. ریخت‌شناسی فارسی با رویکرد مشکلات املایی چالش‌هایی از قبیل: 1) قواعد فعلی، 2) قواعد وندی[1]، و 3) قواعد فاصله‌گذاری، را موجب می‌شود.

4-3-1. قوانین ساخت و صرف افعال

عدم وجود قوانین تعریف شده‌ی مشخص برای ساخت و صرف افعال فارسی، وجود افعال بسیار پیچیده، وجود افعال چندجزئی و مرکب و نیز موارد خاص و استثنا که از قوانین صرف فعلی تبعیت نمی‌کنند، ریشه‌یابی افعال فارسی را مشکل می‌سازد. از این رو جهت حفظ صحت و دقت کارکرد غلط‌یابی، می‌باید تمامی حالات صرف کلیه‌ی افعال در لیست واژگان زبان قرار گیرند. اما چون تعداد این افعال بسیار زیاد می‌شود (تا 5 برابر کل کلمات مناسب زبان برای غلط‌یابی) این امر خود چالشی دگر است که استفاده ترکیبی از ریشه‌یاب‌های قدرتمند و مطمئن برای بخش اعظمی از افعال و نیز قراردادن افعال خاص در واژه‌نامه فعلا بهترین راه‌کار به نظر می‌رسد.

4-3-2. قوانین ترکیب وندها

مبحث وندها در فارسی از دیگر مشکلات ریخت‌شناسی زبان فارسی است. زبان فارسی به نسبت زبان انگلیسی شامل تعداد زیاد از پس‌وندها است، که این وندها در برخی موارد هنگام ترکیب با کلمه اصلی، موجب ایجاد تغییرات ساختاری در کلمه پایه می‌شوند. از طرفی پس‌وندهای فارس می‌توانند با یکدیگر ترکیب شوند که این قوانین ترکیب بسیار پیچیده است و دارای حالات خاص بسیاری است. تعداد این

[1] Affix

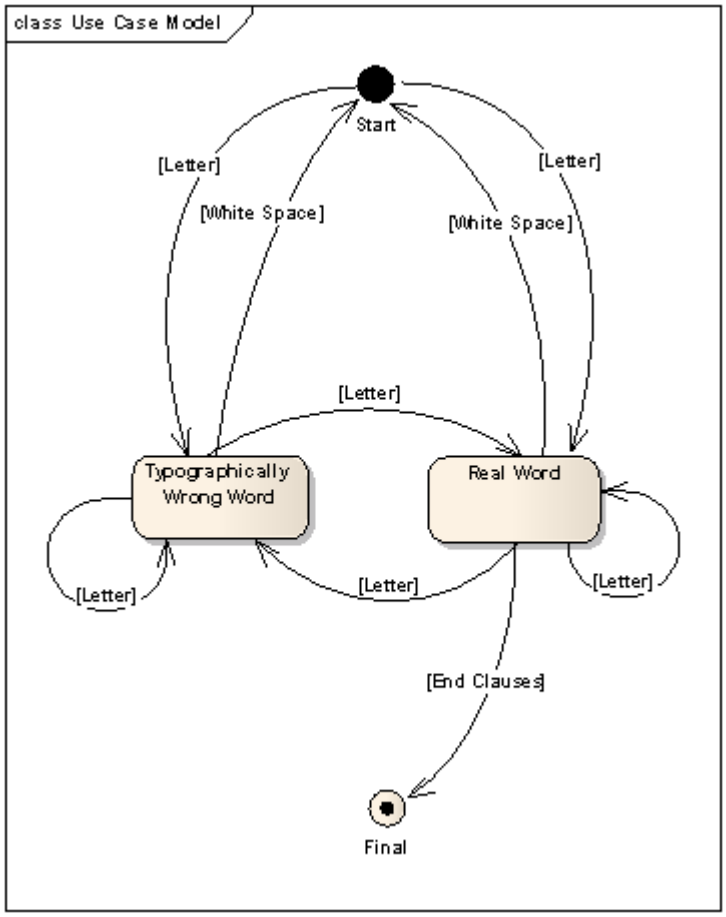
ترکیبات پس‌وندی در زبان فارسی حدود 250 مورد است که با در نظر گرفتن حالات خاص و حالات محاوره‌ای، از این تعداد نیز فراتر خواهند رفت.

استفاده از وندها در زبان فارسی محدوده لغات زبان را می‌تواند تا 250 برابر افزایش دهد که ساختن، شناسایی و قرار دادن کلیه‌ی این کلمات در لیست واژگان امری تقریبا غیر ممکن است و همچنین چنین لیست حجیمی عملا محاسبات را با چالش‌های بسیار جدی مواجه می‌سازد. از این رو استفاده از ابزارهای خودکار جداسازی وند[1]، توصیه می‌شود. اما چالش دیگری که در این زمینه به وجود می‌آید این است که کلیه‌ی کلمات زبان همه‌ی وندها و ترکیب‌هایشان را نمی‌پذیرند و یا اصلا وندی نمی‌پذیرند (مانند افعال) بنابراین در استفاده از ابزارهای خودکار جداسازی وند نیز می‌باید دقت نظر داشت و حداقل کلمات موجود در لیست واژگان را از لحاظ پذیرش و یا عدم پذیرش وندها تفکیک نمود.

4-3-3. قوانین فاصله‌گذاری

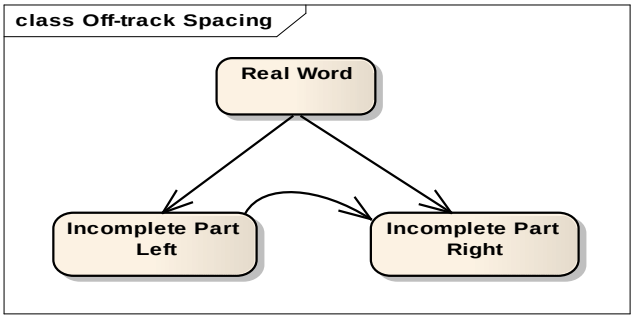
در مدل زبان فارسی، نوشتار زبان فارسی از هم‌آیی حروف جهت ایجاد کلمات، فاصله برای تفکیک کلمات و علائم نشانه‌گذاری پایان دهنده برای اتمام جملات استفاده می‌شود. در این میان معمولا هم‌آیی اشتباه حروف موجب ایجاد غلط‌های املایی در کلمات می‌شوند و کلماتی خارج از کلمات زبان را ایجاد می‌کنند. اما نکته‌ای دیگر که در لایه‌ی ریخت‌شناسی زبان می‌تواند غلط‌های املایی را موجب شود. اشتباه در هم‌آیی فاصله و حروف است به گونه‌ای که این هم‌آیی اشتباه، کلمات صحیحی را دچار اشکال نماید. مدل ساخت کلمات و جملات فارسی در شکل 1 منعکس گردیده است.

[1] Affix-stripper



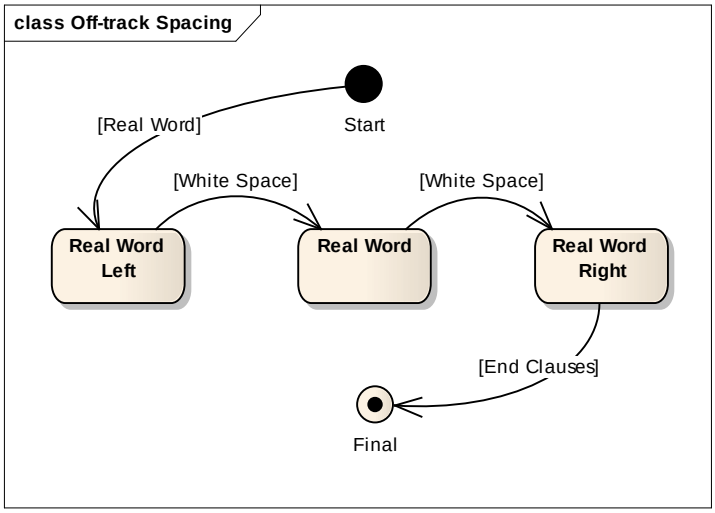
شکل 1. مدل ساخت کلمات و جملات زبان فارسی

حال فرض کنیم که یک کلمه از 2 بخش *زیرکلمه* سمت چپ و *زیرکلمه* سمت *راست* همان گونه کـه در شکل 2 مشخص گردیده، تشکیل شده باشد.



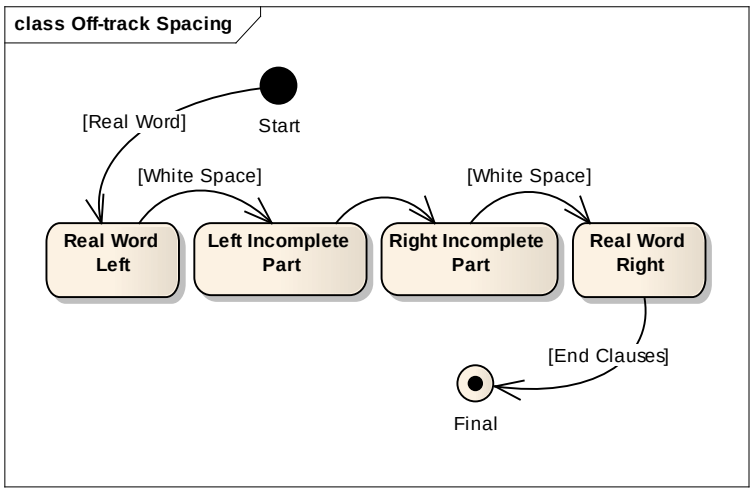
شکل 2

همچنین در یک مدل از هم‌آیی سه کلمه صحیح، مدل زبان را می‌توان همانند شکل 3 متشکل از سه *کلمه صحیح سمت چپ*، *کلمه صحیح میانی* و *کلمه صحیح سمت راست*، دانست.



شکل 3

حال اگر *کلمه صحیح میانی* را با معادل آن در شکل 2 جای‌گذاری کنیم یک مدل هم‌آیی 4تایی از کلمات را در مدل صحیح زبانی آن گونه که در شکل 4 انعکاس یافته، خواهیم داشت.



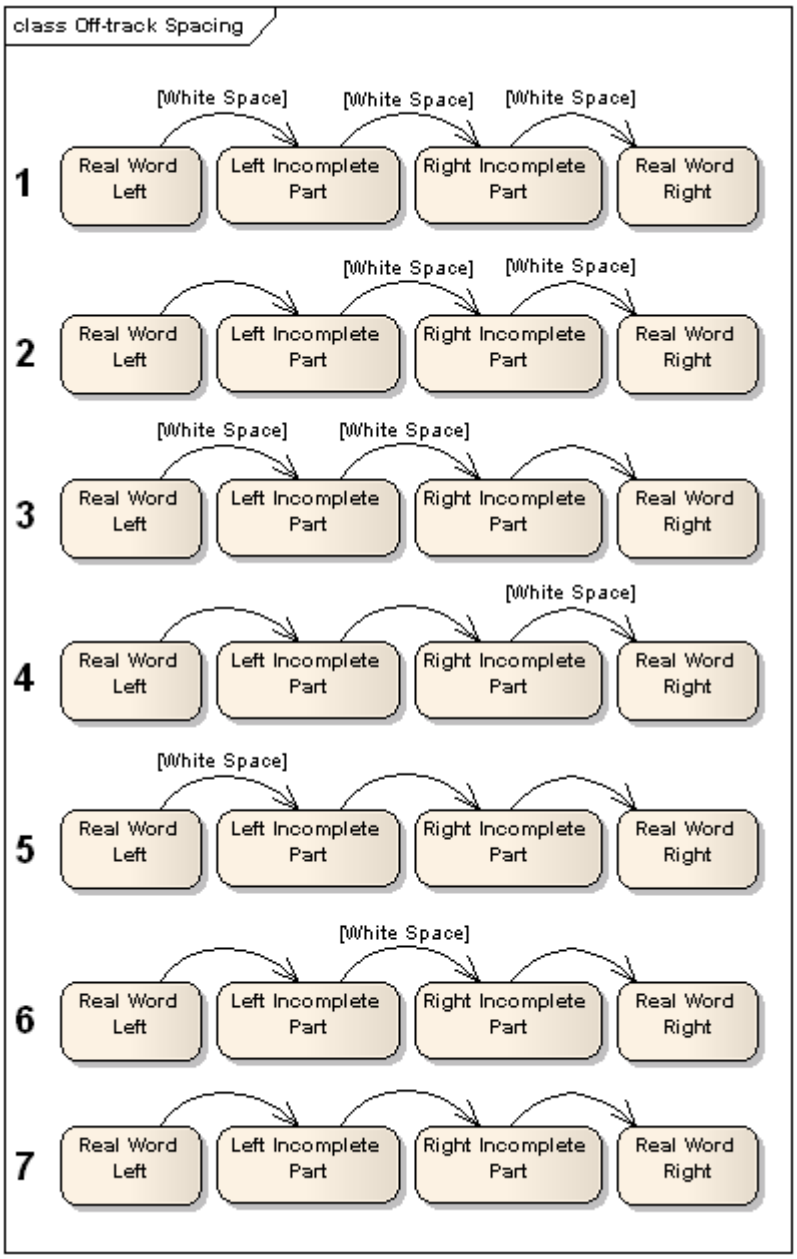
شکل 4

در مدل شکل 4، سه حالت برای جابجایی فاصله به وجود می‌آید:

1) مابین کلمه صحیح سمت چپ و زیرکلمه سمت چپ
2) مابین کلمه زیرکلمه سمت چپ و زیرکلمه سمت راست
3) مابین زیرکلمه سمت راست و کلمه صحیح سمت راست

بنابراین هشت (2^3) حالت از فاصله‌گذاری برای یک مدل هم‌آیندی سه کلمه‌ای را می‌توان برای زبان متصور شد که تنها یک حالت از این هشت حالت صحیح است. شکل 5 هفت حالت ناصحیح ممکن از فاصله‌گذاری اشتباه در مدل هم‌آیندی 3تایی زبان فارسی را نشان می‌دهد.

اشکالات فاصله‌گذاری اشتباه و ترکیب آن‌ها با چالش‌های شبه-فاصله، عمده (حدود 80%) خطاهای املایی زبان فارسی را شامل می‌شود [19] که در روش‌های معمول خطایابی قابل تصحیح و در برخی موارد حتی قابل تشخیص نیز نیستند. بنابراین در نظر گرفتن چالش فاصله‌گذاری اشتباه در زبان فارسی از نکات بسیار مهم در طراحی و ایجاد خطایاب‌های املایی صرفی و حتی نحوی است.


class Off-track Spacing
1
2
3
4
5
6
7
[White Space]
Real Word Left
Left Incomplete Part
Right Incomplete Part
Real Word Right

شکل 5

4-4. هم‌آواها

میتون [16] در تحقیقات خویش نشان داد که تنها محاسبه‌ی فاصله‌ی ویرایشی برای هم‌آواها مناسب و کافی نیست. او همچنین دریافت که در 56% مواقع یک هم‌آوا بهترین پیشنهاد برای یک تصحیح یک کلمه‌ی دارای خطای املایی است، البته اگر که هم‌آوای کلمه در لیست پیشنهادات موجود باشد.

با توجه به تعداد و گستردگی کلمات هم‌آوا در زبان فارسی و همچنین مطالعات صورت گرفته در پیکره‌های متنی بزرگ فارسی [19]، احتمال این که پیشنهاد مناسب برای کلمه‌ی دارای اشتباه املایی یک هم‌آوای آن باشد، در زبان فارسی بیش از 76% درصد است که این احتمال بسیار بیش‌تر از احتمال مطالعه شده توسط میتون برای زبان انگلیسی است.

با توجه به این قضیه، تعداد بسیار هم‌آواها در زبان فارسی حتی می‌تواند افراد تحصیل کرده‌ی فارسی زبان را دچار مشکل ساخته و اشتباهات املایی بسیاری از این دست را موجب شود. بنابراین توجه به اصلاح املایی هم‌آواها به صورت مجزا از سیاست‌های کلی غلط‌یابی امری دیگر است که در طراحی غلط‌یاب‌های املایی زبان فارسی می‌باید مورد توجه ویژه قرار گیرد.

5-4. توزیع انواع خطاهای املایی

بررسی توزیع انواع خطاهای املایی تایپی می‌تواند منتج به استخراج احتمال رخداد هر گونه از خطا شامل درج، حذف، تعویض و جابجایی حروف، در زبان فارسی شود. یک بررسی [9] در متون انگلیسی این توزیع را این‌گونه مطالعه کرده است:

1) حذف یک حرف – 37.6%
2) درج یک حرف – 23.5%
3) تعویض یک حرف – 24.8%
4) جابجایی دو حرف مجاور – 14.1%

مطالعه‌ای دیگر [16] و باز برای زبان انگلیسی این توزیع را ین گونه محاسبه نموده است:

1) حذف یک حرف – 33%
2) درج یک حرف – 19%
3) تعویض یک حرف – 42%
4) جابجایی دو حرف مجاور – 6%

محاسبه‌ی این توزیع برای زبان فارسی امری ضروری است که می‌تواند احتمال رخداد هر نوع از خطا را مشخص سازد. برای مثال احتمال رخداد خطای حذف یک حرف در توزیع فوق 14% بیشتر از خطای درج یک حرف اضافی است. استفاده از این احتمالات می‌تواند نحوه‌ی انتخاب میان پیشنهادات برای یک کلمه را به شکل مطلوبی بهینه سازد.

5. روش‌های مرتب‌سازی و انتخاب پیشنهادات

تولید پیشنهادات و همچنین پیشنهاد مطلوب و افزودن آن به لیست پیشنهادات، تازه ابتدای راه خطایابی است. یک غلط‌یاب می‌باید پیشنهادات مطلوب‌تر را از میان شاید صدها پیشنهاد گزینش کرده و لیستی کوچکتر (معمولا 7تایی) از پیشنهادات را ارائه کند. از طرفی دقت و قدرت یک غلط‌یاب هنگامی مشخص می‌شود که پیشنهاد مطلوب سر لیست پیشنهادات و یا هر چه نزدیک‌تر به ابتدای لیست باشد [15 ,11]. بنابراین مرتب‌سازی پیشنهادات امری بسیار ضروری و پیچیده در غلط‌یابی املایی است. در این بخش به راه‌کارهای رایج در این زمینه می‌پردازیم.

5-1. روش ساده فاصله حروف

در روش فاصله حروف[1] ابتدا حروف دو کلمه به صورت متناظر مورد بررسی قرار می‌گیرند. در صورت عدم تطابق یک امتیاز منفی محاسبه می‌شود. سپس حروف دو کلمه دو به دو، مقایسه می‌شوند و مانند مورد قبل در صورت عدم تطابق یک امتیاز منفی دیگر نیز محاسبه می‌شود. سپس اگر حروف اول دو کلمه نیز متفاوت بودند، بار دیگر امتیازی منفی محاسبه می‌گردد.

به عنوان مثال برای دو کلمه‌ی «تقلب» و «تفلاب» این امتیازات این‌گونه محاسبه می‌شوند:

3 امتیاز منفی برای عدم هم‌خوانی حروف متقابل

4 امتیاز منفی برای عدم هم‌خوانی دو به دوی حروف متقابل (مثلا عدم هم‌خوانی «تق» با «تف» و «قل» با «فل»)

و چون حروف اول یکسان هستند امتیاز منفی دیگری محاسبه نمی‌شود، بنابر این امتیاز منفی محاسبه شده برای این دو کلمه 7 است.

حال اگر همین امتیاز را برای دو کلمه‌ی «تالاب» و «تفلاب» محاسبه کنیم:

[1] Letter Distance

1 امتیاز منفی برای عدم هم‌خوانی حروف متقابل

2 امتیاز منفی برای عدم هم‌خوانی دو به دوی حروف متقابل

که در نهایت امتیاز منفی در این حالت معادل 3 می‌شود که با مقایسه به حالت قبل می‌توان نتیجه گرفت «تالاب» پیشنهاد بهتری برای «تفلاب» است نسبت به «تقلب».

5-2. بسامد کلمات

در این روش با استفاده از بسامد تکرار کلمات در مدل زبانی، میان پیشنهادات مواردی که بسامد بیشتری دارند انتخاب می‌شوند. مشکل این روش محاسبه صحیح بسامد کلمات است که بسیار وابسته به پیکره‌های متنی مورد استفاده است و از طرفی برخی کلمات مانند حروف اضافه و برخی افعال، بسامد بسیار بالایی پیدا می‌کنند در حالی که احتمال اینکه اشتباه نوشته شوند پایین است.

5-3. فاصله ویرایشی کمینه

در این روش‌ها از فاصله ویرایشی به عنوان معیار تشابه میان کلمات استفاده می‌شود [14]. این روش‌ها شامل هفت راه‌کار اصلی هستند: 1) روش فاصله‌ی همینگ[۱]، 2) روش فاصله‌ی لونشتاین[۲]، 3) روش فاصله دامرو-لونشتاین[۳]، 4) روش فاصله وگنر-فیشر[٤]، 5) روش فاصله هیرشبرگ[٥]، 6) روش اوکونن[٦]، و 7) روش جرو-وینکلر[۷].

[۱] Hamming

[۲] Levenshtein

[۳] Dameru-Levenshtein

[٤] Wagner-Fischer

[٥] Hirschberg

[٦] Ukkonen

[۷] Jaro-Winkler

5-3-1. روش همینگ

فاصله همینگ دو کلمه با طول مشابه شامل تعداد حروف متناظر نامشابه است [32]. فرمول (1) روش محاسبه فاصله همینگ را نشان می‌دهد. به عنوان مثال فاصله همینگ دو کلمه «امید» و «نماد»، 2 است که برای تبدیل آن به فرم نرمال شده در بازه‌ی 0 تا 1، این فاصله به طول کلمات تقسیم می‌گردد. این معیار میزان تفاوت را نشان می‌دهد جهت محاسبه معیار مشابهت، همانند فرمول (2)، این مقدار را از 1 کم می‌کنیم؛ بنابراین میزان تشابه همینگ دو کلمه فوق $1 - \frac{2}{4} = 0.5$ است.

$$f_h(i) = 0 \; \boldsymbol{if} \; (q_i = l_j) \; \boldsymbol{else} \; f_h(i) = 1 \qquad (1)$$

$$Similarity = 1 - \frac{HammingDistance\,(A,B)}{max(|A|,|B|)} \qquad (2)$$

5-3-2. روش لونشتاین

فاصله‌ی لونشتاین میان دو رشته از کمینه‌ی تغییرات لازم برای تبدیل یک رشته به دیگری محاسبه می‌شود. این تغییرات شامل، درج یک حرف اضافه، حذف یک حرف و یا تعویض دو حرف است [33]. برخلاف روش همینگ، این روش می‌تواند بر کلمات با طول متفاوت نیز اعمال شود که این تفاوت در طول می‌تواند توسط حذف و یا درج حروف ایجاد شده باشد.

برای مثال فاصله‌ی لونشتاین میان دو کلمه‌ی «کامپیوتر» و «کامپچیوت» معادل 2 است (1 برای حذف «چ» و یک برای درج «ر»).

فاصله‌ی لونشتاین با استفاده از فرمول (3) و (4) محاسبه می‌گردد.

$$f_i(0,0) = 0 \qquad (3)$$

$$f_l(i,j) = min\ [(f_l(i-1,j)+1, f_l(i,j-1)+1, f_l(i-1,j-1)+d(q_i,l_j))]$$
$$\boldsymbol{if}\ (q_i = l_j), d(q_i,l_j) = 0\ \boldsymbol{else}\ d(q_i,l_j) = 1 \quad (4)$$

برای همه‌ی کلمات یک تابع $f_l(0,0)$ محاسبه و معادل صفر می‌گردد. سپس یک تابع $f_l(i,j)$ برای تمامی حروف به صورت مکرر محاسبه می‌شود. طبق فرمول (4)، هر درج، حذف و یا جابجایی یک امتیاز به تابع می‌افزاید [33, 34]. همانند روش همینگ معیار تشابه در این روش نیز با کسر فاصله از یک طبق فرمول (5) محاسبه می‌گردد.

$$Similarity = 1 - \frac{f(|A|-1,|B|-1)}{max\ (|A|,|B|)} \quad (5)$$

3-3-5. روش دامرو-لونشتاین

دامرو با مطالعه و بررسی پیکره‌های متنی بسیار، گونه‌ای دیگر از خطاهای املایی به عنوان جابجایی دو حرف مجاور در یک کلمه را نیز مشاهده نمود [3]. بنابراین با افزودن این نوع خطا به روش لونشتاین، روش دامرو-لونشتاین از چهار نوع خطا، طبق فرمول (6) و (7) پشتیبانی می‌کند. معیار مشابهت در این روش نیز همانند روش لونشتاین از فرمول (5) به دست می‌آید.

$$f_{dl}(0,0) = 0 \quad (6)$$

$$\boldsymbol{if}\ [(i>1)\ and\ (j>1) and\ (q_i = l_{j-1}) and\ (q_{i-1} = l_j)]$$
$$f_{dl}(i,j) = min\ [(f_{dl}(i,j), f_{dl}(i-2,j-2)+d(q_i,l_j)]$$
$$\boldsymbol{else} \quad (7)$$
$$f_{dl}(i,j) = min\ [(f_{dl}(i-1,j)+1, f_{dl}(i,j-1)+1, f_{dl}(i-1,j-1)+d(q_i,l_j))]$$
$$(d(q_i,l_j) = 0\ \ \boldsymbol{if}\ (q_i = l_j)\ \boldsymbol{else}\ d(q_i,l_j) = 1$$

5-3-4. روش وگنر-فیشر

روش وگنر-فیشر، یک روش تغییر یافته از روش لونشتاین است که در آن، بر خلاف روش لونشتاین که هزینه‌ی مشابهی برای حذف، درج و تعویض که معادل 1 بود در نظر می‌گرفت، هزینه‌های متفاوتی برای هرگونه از خطاها در نظر گرفته می‌شود [35].

این روش و همچنین روش لونشتاین و دامرو-لونشتاین، دارای پیچیدگی زمانی و فضایی $O(mn)$ هستند که m طول کلمه اول و n طول کلمه دوم است. روش وگنر-فیشر توسط فرمول‌های (8) و (9) محاسبه می‌شود که همانند محاسبه لونشتاین است با این تفاوت که در لونشتاین کلیه‌ی هزینه‌ها معادل 1 در نظر گرفته شده بود.

$$f_{wf}(0,0) = 0 \tag{8}$$

$$f_{wf}(i,j) = \min\left[\left(f_{wf}(i-1,j) + d(q_i,\varepsilon)\right., f_{wf}(i,j-1) + d(\varepsilon,l_j),\right.$$
$$\left. f_{wf}(i-1,j-1) + d(q_i,l_j)\right)] \tag{9}$$

که $d(q_i,\varepsilon)$ هزینه‌ی حذف، $d(\varepsilon,l_j)$ هزینه‌ی درج و $d(q_i,\varepsilon)$ هزینه‌ی تعویض دو حرف است.

5-3-5. روش هیرشبرگ

روش هیرشبرگ، تغییر یافته‌ی روش وگنر-فیشر است که پیچیدگی زمانی آن همان $O(mn)$ است اما فضای مورد نیاز را از مرتبه‌ی درجه 2 به مرتبه‌ی خطی کاهش داده است [36]. این روش تمرکزی بر افزایش دقت مشابهت‌یابی ندارد و تنها تمرکز بر کاهش پیچیدگی فضایی روش دارد.

5-3-6. روش اوکونن

روش اوکونن [37] تغییر دیگری از روش وگنر-فیشر است که در آن پیچیدگی زمانی و فضایی به $O(dm)$ کاهش یافته است که در آن d فاصله‌ی میان دو کلمه‌ی مورد نظر و m طول کمینه‌ی این دو کلمه است. روش اوکونن تقریبا سریع‌ترین روش برای محاسبه مشابهت میان کلمات است [38].

5-3-7. روش جرو-وینکلر

روش جرو-وینکلر [39] یک استفاده از روش فاصله‌ی جرو است که پیش‌تر برای محاسبه میزان اتصال رکوردها استفاده می‌شد. فاصله‌ی جرو بین دو کلمه A و B با استفاده از فرمول (10) محاسبه می‌شود.

$$f_j = \frac{1}{3}\left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m-t}{m}\right) \quad (10)$$

در این فرمول، m تعداد حروف مشترک و t تعداد جابجایی دو حرف مجاور است. معیار جرو مقداری نرمال شده را نتیجه می‌دهد که در آن صفر نمایانگر هیچ‌گونه شباهت و یک نمایانگر شباهت کامل است.

وینکلر [40] روش جرو را در با تعداد حروف مشترک اول کلمه ترکیب کرد و روش جدیدی که در فرمول (11) انعکاس یافته را ارائه کرد.

$$f_{wj} = f_j + \left(l\rho(1 - f_j)\right) \quad (11)$$

در این روش، l طول زیررشته مشترک در ابتدای دو کلمه و حداکثر 4 است، ρ یک معیار مقیاس اهمیت برای زیررشته‌ی ابتدایی است. ρ به طور پیش‌فرض معادل 0.1 در نظر گرفته می‌شود. این معیار نیز نرمال است و صفر نشانگر عدم تشابه و یک نشانگر تشابه کامل است.

5-3-8. محاسبه فاصله میان حروف

تعویض حروف متاثر از مکان قرارگیری حروف بر روی صفحه‌کلید است. به عنوان مثال دو حرف مجاور هم بر روی صفحه‌کلید، بیشتر امکان دارد که به اشتباه به جای یکدیگر تایپ شوند تا دو حرفی که فاصله بیشتری نسبت به هم بر روی صفحه‌کلید دارند [18]. این فاصله فاصله‌ی میان حروف نامیده می‌شود.

برای روشن‌تر شدن بحث، یک شمای کلی از طرح‌بندی صفحه‌کلید استاندارد انگلیسی در شکل 6 و یک طرح‌بندی صفحه‌کلید استاندارد فارسی در شکل 7 نشان داده شده است.

x/y	0	1	2	3	4	5	6	7	8	9	10	11
0	q	w	E	r	t	y	u	i	o	p	[	]
1	a	s	d	f	g	h	j	k	l	;	‘	
2	z	x	c	v	b	n	m	,	.	/		

شکل 6. شمای کلی طرح‌بندی صفحه‌کلید استاندارد انگلیسی

x/y	0	1	2	3	4	5	6	7	8	9	10	11	12
0	پ												
1		ض	ص	ث	ق	ف	غ	ع	ه	خ	ح	ج	چ
2		ش	س	ی	ب	ل	ا/آ	ت	ن	م	ک	گ	
3		ظ	ط	ز/ژ	ر	ذ	د	ئ	و	.	/		

شکل 7. شمای کلی طرح‌بندی صفحه‌کلید استاندارد فارسی

فاصله‌ی میان حروف برای دو حرف c_1 و c_2 که c_2 در مکان (x_2, y_2) و c_1 در مکان (x_1, y_1) از صفحه‌کلید قرار گرفته‌اند با استفاده از فاصله اقلیدسی [41] از فرمول (14) محاسبه می‌شود.

$$d_e(c_2, c_1) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

(x_2, y_2) *is the position of* c_2 *in keyboard layout*

$$(x_1, y_1) \text{ is the position of } c_1 \text{ in keyboard layout} \qquad (14)$$

به عنوان مثال فاصله‌ی میان حرف *a* که در مکان *(0, 1)* و *p* که در مکان *(9, 0)* قرار گرفته‌اند معادل رابطه زیر خواهد بود:

$$d_e(p, a) = \sqrt{(9-0)^2 + (0-1)^2} = 9.05538$$

و نیز فاصله‌ی میان حرف «چ» که در مکان *(12, 1)* و «ظ» که در مکان *(1, 3)* قرار گرفته‌اند معادل ربطه زیر خواهد بود:

$$d_e(\text{ظ}, \text{چ}) = \sqrt{(12-1)^2 + (1-3)^2} = 11.18033$$

این معیارها می‌توانند به عنوان هزینه‌ی تعویض حروف در روش‌های مشابهت‌یابی ذکر شده به مار روند. نکته‌ای که در اینجا می‌باید به آن توجه گردد این است که زبان فارسی شامل حروفی است که با فشردن کلید *Shift* حاصل می‌شوند و می‌باید این نکته نیز در این فاصله لحاظ شود.

فاصله میان دو حرف می‌تواند بر اساس فرمول (15) نرمال شود. این گونه که حاصل فرمول (14) به بیشترین فاصله‌ی موجود بر روی طرح‌بندی صفحه‌کلید تقسیم می‌گردد.

$$d_s(c_2, c_1) = \frac{d_e(c_2, c_1)}{MaxDistance} * Cost(err_{st})$$

$$Cost(err_{st}) = 0.54 \qquad (15)$$

$MaxDistance =$ *Maximum of Euclidean Character Distance*

6. نتیجه‌گیری

با توجه به پیشرفت و همه‌گیر شدن کاربرد کامپیوترها، نیاز به غلط‌یاب‌های املایی در زمینه پردازش زبان که نوشتار رایانه‌ای و دیگر کاربردهای مرتبط را به شدت تسهیل می‌کند از اهمیت ویژه‌ای برخوردار است. با توجه به چالش‌های خاص زبان فارسی طراحی و ایجاد خطایاب‌های املایی صرفی فارسی نیازمند دقت نظری ویژه هستند.

در این مستند چالش‌های زبان فارسی در خصوص غلط‌یابی املایی را بررسی کردیم که به طور خلاصه شامل: 1) اشکالات رسم‌الخط و دستور نگارش، 2) وجود حروف هم‌شکل بسیار، 3) قوانین ریخت‌شناسی بسیار پیچیده، 4) وجود حروف هم‌آوای بسیار، و 5) توزیع متفاوت انواع غلط‌های تایپی، بودند. در این میان به لزوم توجه به قوانین ساخت و صرف افعال، قوانین ترکیب وندها و قوانین فاصله‌گذاری اشاره کردیم و چالش‌های مرتبط را بر شمردیم.

به ساختار داده مناسب برای خطایابی املایی و ویژگی‌ها، قابلیت‌ها و خصوصیات ویژه‌ی مطلوب اشاره کردیم.

نحوه تولید پیشنهادات و نکات ویژه که در این امر می‌باید مورد توجه قرار گیرند را از نظر گذراندیم و سپس اهمیت مرتب‌سازی پیشنهادات را مورد بررسی قرار دادیم. روش‌های مرسوم برای مرتب‌سازی و مشابهت‌یابی میان پیشنهادات را به طور مفصل بررسی کردیم و نکاتی که در استفاده از آن‌ها برای زبان فارسی می‌باید در نظر داشت را ذکر نمودیم.

امید است که این مستند راه‌گشا و هدایت‌کننده مناسبی برای علاقه‌مندان به طراحی و اجرای غلط‌یاب‌های املایی صرفی زبان فارسی باشد.

7. مراجع

[1] C. M. Eastman, and D. S. McLean, "On the need for parsing ill-formed input," *Comput. Linguist.*, vol. 7, no. 4, pp. 257-257, 1981.

[2] C. Young, C. Eastman, and R. Oakman, "An analysis of ill-formed input in natural language queries to document retrieval systems," *Inf. Process. Manage.*, vol. 27, no. 6, pp. 615-622, 1991.

[3] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171-176, 1964.

[4] E. Galli, and H. Yamada, "Experimental studies in computer-assisted correction of unorthographic text," *IEEE Transactions on Engineering Writing and Speech*, vol. 11, no. 2, pp. 75-84, 1968.

[5] A. Hanson, E. Riseman, and E. Fisher, "Context in word recognition," *Pattern Recognition*, vol. 8, no. 1, pp. 35-45, 1976.

[6] J. Pollock, and A. Zamora, "Collection and Characterization of Spelling Errors in Scientific and Scholarly Text," *Journal of the American Society for Information Science*, vol. 34, no. 1, pp. 51-58, 1983.

[7] C. STERLING, "Spelling errors in context," *British journal of psychology(1953)*, vol. 74, no. 3, pp. 353-364, 1983.

[8] R. Mitton, "Spelling checkers,spelling correctors and the misspellings of poor spellers," *Inf. Process. Manage.*, vol. 23, no. 5, pp. 495-505, 1987.

[9] M. Kyongho, H. William, Wilson *et al.*, "Typographical and orthographical spelling error correction," in Proceedings of 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 2000, pp. 1781-1785.

[10] J. J. Pollock, and A. Zamora, "Automatic spelling correction in scientific and scholarly text," *Commun. ACM*, vol. 27, no. 4, pp. 358-368, 1984.

[11] R. Mitton, "Spellchecking by computer," *Journal of Simplified Spelling Society*, vol. 20, no. 1, pp. 4-11, 1996.

[12] E. Yannakoudakis, and D. Fawthrop, "The Rules of Spelling Errors," *Information Processing and Management,* vol. 19, no. 2, pp. 87-99, 1983.

[13] E. Yannakoudakis, and D. Fawthrop, "An intelligent spelling error detector'," *Information Processing and Management,* vol. 19, no. 2, pp. 101-108, 1983.

[14] E. S. Ristad, and P. N. Yianilos, "Learning String-Edit Distance," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 20, no. 5, pp. 522-532, 1998.

[15] J. L. Peterson, "Computer programs for detecting and correcting spelling errors," *Commun. ACM,* vol. 23, no. 12, pp. 676-687, 1980.

[16] R. Mitton, *Ordering the suggestion of spellcheckers: isolated-word correction,* Technical Report BBKCS-06-07, 2006.

[17] J. Peterson, *Computer Programs for Spelling Correction*: Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1980.

[18] K. Min, and W. H. Wilson, "Syntactic recovery and spelling correction of ill-formed sentences," in Proceedings of 3rd Conference of the Australasian Cognitive Science, 1995.

[19] O. Kashefi, M. Sharifi, and B. Minaei, "A Novel String Distance Metric for Persian Text," *Ieee Transaction on Audio, Speech and Language Processing*, 2009.

[20] M. Pilotti, and M. Chodorow, "Error detection/correction in collaborative writing," *Springer Science andBusiness Media*, 2007.

[21] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys,* vol. 24, pp. 378- 439, 1992.

[22] R. Garfinkel, E. Fernandez, and R. Gopal, "Design of an interactive spell checker: optimizing the list of offered words," *Decision Support Systems,* vol. 35, pp. 385-397, 2003.

[23] V. J. Hodge, and J. Austin, "A comparison of a novel neural spell checker and standard spell checking algorithms," *Pattern Recognition,* vol. 35, pp. 2571-2580, 2002.

[24] M. D. Kemighan, Kenneth W. Church, and W. A. Gale, "A Spelling Correction Program Based on a Noisy Channel Model," in 13th International Conference on Computational Linguistics, 1990.

[25] J. R. Ullman, “A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words,” *Computer Journal*, vol. 20, no. 2, pp. 141-147, 1977.

[26] T. Naseem, and S. Hussain, “A novel approach for ranking spelling error corrections for Urdu,” *Language Resources & Evaluation*, vol. 41, pp. 117-128, 2007.

[27] د. نوگورانی و م. صبوریان، "طراحی و پیاده‌سازی یک خطایاب فارسی"، دومین همایش فارسی و رایانه، تهران، 1384.

[28] B. QasemiZadeh, A. Ilkhani, and A. Ganjeii, “Adaptive Language Independent Spell Checking using Intelligent Traverse on a Tree,” in IEEE Conference on Cybernetics and Intelligent Systems, 2006.

[29] A. Mokhtaripour, and S. Jahanpour, “Introduction to a New Farsi Stemmer,” in CIKM, Arlington, Virginia, USA, 2006, pp. 826-827.

[30] C. Comeau, and W. J. Wilbur, “Non-Word Identification or Spell Checking Without a Dictionary,” *Journal Of The American Society For Information Science and Technology*, vol. 55, no. 2, pp. 169-177, 2004.

[31] M. S. Rasooli, and B. Minaei-Bidgoli, “A new approach for Persian spellchecking,” in 2nd Data Mining Conference, Tehran, Iran, 2008.

[32] R. W. Hamming, “Error detecting and error correcting codes,” *Bell System Tech. J.*, vol. 29, no. 2, pp. 147-160, 1950.

[33] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." p. 707.

[34] V. J. Hodge, and J. Austin, “A Comparison of Standard Spell Checking Algorithms and a Novel Binary Neural Approach,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 15, no. 5, pp. 1073-1081, 2003.

[35] R. A. Wagner, and M. J. Fischer, “The String-to-String Correction Problem,” *J. ACM*, vol. 21, no. 1, pp. 168-173, 1974.

[36] D. S. Hirschberg, “A linear space algorithm for computing maximal common subsequences,” *Communications of the ACM*, vol. 18, no. 6, pp. 341-344, 1975.

[37] E. Ukkonen, “Algorithms for approximate string matching,” *Inf. Control*, vol. 64, no. 1-3, pp. 100-118, 1985.

[38] W. I. Chang, and J. Lampe, "Theoretical and Empirical Comparisons of Approximate String Matching Algorithms," in Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching, 1992.

[39] M. A. Jaro, "Advances in record linking methodology as applied to the 1985 census of Tampa Florida," *Journal of the American Statistical Society*, vol. 84, no. 406, pp. 414-420, 1989.

[40] W. Winkler, and Y. Thibaudeau, "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census," *Research Report RR91/09, US Bureau of the Census*, 1991.

[41] T. Heath, "The Thirteen Books of Euclid's Elements, Vol. 1," *New York: Doner*, 1956.