تنوع نگارشی در فارسی و تهیهٔ پیکرهٔ زبانی در قالب فرهنگ

ساغر شریفی (دانشگاه آزاد اسلامی واحد کرج)
مرضیهٔ صناعتی (پژوهشکدهٔ زبان‌شناسی، کتیبه و متون، پژوهشگاه سازمان میراث فرهنگی)
مسعود قیومی (پژوهشکدهٔ زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی)

۱ـ مقدمه

ویژگی‌های خط فارسی سبب بروز مشکلاتی در املای این زبان می‌شود، تا جایی که کمتر کسی‌را می‌توان یافت که درمورد املای فارسی دچار مشکل نشده باشد (مشیری ۱۳۶۶، ص ۹). نایک‌دستی در رسم‌الخط فارسی از جمله عواملی است که منجر به ایجاد تعدد و تنوع در کاربرد نشانه‌های موجود در خط شده‌است. این عدم یک‌دستی نه تنها در متون رسمی، از جمله کتاب‌های درسی و داستان، نامه‌های اداری، روزنامه‌ها و مانند آن، مشاهده می‌شود، بلکه در سایر متون مکتوبی که در آن‌ها گونه‌های گفتاری، غیررسمی و محاوره‌ای به‌کار می‌رود، همچون پیامک‌ها و غیره، نیز رواج فراوان پیدا کرده‌است. پژوهشگران و افرادی که با نگارش سروکار دارند نیز در این زمینه به‌تفصیل سخن گفته و به مواردی ـ عمدتاً صوری و نوشتاری ـ اشاره کرده‌اند. به منظور حل برخی از این مشکلات در حوزهٔ پژوهش بر روی زبان فارسی، وجود پیکرهٔ زبانی[1] متعادلی که دستخوش یکسوشدگی نباشد ضروری است. تهیهٔ این پیکرهٔ زبانی از منابع متفاوت و پردازش خط فارسی سبب بروز مشکلاتی می‌شود که قیومی و ممتازی (Ghayoomi and Momtazi 2009)، قیومی و دیگران (Ghayoomi et al. 2010)، شمس‌فرد

1. linguistics corpus

(Shamsfard 2011) و ستوده و هنرجویان (۱۳۹۱؛ ۱۳۹۳) موارد چالش‌برانگیز آن را به‌تفصیل مطرح کرده‌اند. ازجملهٔ این موارد، مشکلِ تعیین مرز درونی واژه، سبک نگارش، ورود حروف خاص عربی به فارسی، و نگارشِ واژه‌های اروپایی با خط فارسی است.

در این مقاله سعی شده‌است با نگاهی متفاوت، موارد چالش‌برانگیزِ تهیهٔ پیکرهٔ زبانی در قالب فرهنگ تنوع نگارشی واژه‌ها در فارسی به کمک رایانه بررسی شود و طبقه‌بندی دانه‌ریزتری[1] از عواملی که منجر به تنوع نگارشی می‌شود ارائه گردد. برای این کار، از فاصلهٔ لِونشتاین[2] برای یافتن تنوع نگارشی واژه‌ها در پیکره استفاده و سپس هزار واژه از داده‌های استخراج‌شده از سه پیکرهٔ ادغام‌شده (← بخش ۳) بررسی شده‌است.

پس از ساماندهی داده‌های استخراج‌شده از این پژوهش می‌توان فهرستی از دادگان[3] به دست داد و نتایج حاصل از تحلیل این داده‌ها را برای زبان‌آموزان غیرفارسی‌زبان و نیز در پژوهش‌های مربوط به زبان‌شناسی پیکره‌ای[4] استفاده کرد.

۲ـ سابقهٔ تحقیق و بیان مسئله

یکی از مشکلات عمدهٔ خط فارسی نوشتن بسیاری از واژه‌ها به چند طریق است، مانند اتاق / اطاق، خوشبخت / خوش‌بخت و هیأت / هیئت. درمورد دلایل این تنوع در نگارش کلمات فارسی بسیار صحبت شده‌است. مشیری در مقدمهٔ فرهنگ فارسی آوایی ـ املایی، مشکل املا را امری رایج می‌داند. او به اشتباهات فاحش املایی افراد باسواد نیز اشاره می‌کند و یکی از دلایل وجود این مشکلات را تداخلِ واژه‌های زبان‌های دیگر، از قبیل عربی، ذکر می‌کند. همچنین از هم‌آواها (کلماتی که یکسان تلفظ می‌شوند، ولی در معنا و املا متفاوت‌اند) نام می‌برد، مانند غالب / قالب و قضا / غذا (مشیری ۱۳۶۶، ص ۹).

افراد دیگری نیز در این زمینه به نکاتی اشاره کرده‌اند که وجه مشترک آن‌ها به قرار زیر است:

---

1. fine-grained (منظور از دانه‌ریز «دربردارندهٔ اطلاعات جزئی» است.)
2. Levenstein distance
3. database
4. corpus linguistics

نوشتن واکه‌های کوتاه و علامت تشدید در خط فارسی که سبب به‌وجود آمدن خوانش‌های متفاوت می‌شود[۱]، عدم تناظر میان واج‌ها و نویسه‌ها (وجود بیش از یک نویسه برای یک واج، مانند «ت» و «ط» برای واج /t/ و به‌کاربردن یک نویسه برای چند واج، مانند «و» برای نشان دادن واج‌های /u/، /v/ و /o/، ننوشتن علامت مد برروی الف، نوشتن همزه بر روی کرسی‌های گوناگون (مانند «ئ»، «ؤ» و «أ»)، وجود شکل‌های مختلف برای برخی از حروف (مانند «ع»)، وجود حروفی که نوشته می‌شوند، ولی خوانده نمی‌شود (مانند «و» در خواستن)، سرهم‌نویسی، جدانویسی و بی‌فاصله‌نویسی به‌صورت دل‌بخواهی و بدون منطق (مانند می‌شود، میشود و می@شود[۲])، ورود واژه‌های عربی با املای خاص خود (مانند حتی‌الامکان)، فراوانی نقطه‌ها و دندانه‌ها ودرنتیجه شبیه شدن حروف (مانند «ر»، «ز»، «ژ» و «س»، «ش»)، تفاوت میان حروف از نظر اتصال به حروف مجاور خود[۳] (آشوری ۱۳۶۵؛ افتخاری ۱۳۸۴؛ امیرجانلو ۱۳۸۱؛ طیب ۱۳۷۱؛ مرتضایی ۱۳۸۰؛ نثری ۱۳۱۴). این ویژگی در بخش ۵ـ۱ـ۱ـ توضیح داده خواهد شد.

صادقی و زندی‌مقدم (۱۳۹۱) در اشاره به شیوه‌های گوناگون نوشتنِ واژه‌ها نمونه‌های اسمان / ااسمان / آسمان، جدای / جدایی / جدائی و مانند آن‌ها را در واژه‌های مشتق و بسیط و نیز مسئلهٔ سرهم‌نویسی و جدانویسی در واژه‌های مرکّب را با نمونه‌های تجارت‌خانه، چاپخانه، زبانشناسی و زیست‌شناسی مطرح و به تفاوت‌های بنیادیِ زبان عربی و فارسی به‌عنوان یکی از عوامل ایجاد تعدد نگارش اشاره می‌کنند. آن‌ها از نقش فرهنگستان زبان و ادب فارسی در این رابطه نام می‌برند. مراکز دیگری، مانند مرکز نشر دانشگاهی، که با مشکلات تنوع نگارشی دست‌به‌گریبان بوده‌اند نیز با نگارش جزوات و کتابچه‌هایی سعی در حل این گونه مسائل کرده‌اند (سمیعی گیلانی ۱۳۶۶).

همچنین مقالاتی درمورد شیوه‌های ایجاد پیکره‌های زبانی مبتنی بر وب و چگونگی رفع مشکلات املایی در متون نوشته شده‌است که از جملهٔ آن‌ها می‌توان قیومی و دیگران (Ghayoomi et al. 2010) را نام برد.

---

۱. این خوانش‌های متفاوت ممکن است واژه‌های عربیِ غیررایج در فارسی و نیز اصطلاحات گویشی را نیز شامل شود.
۲. @ به معنای فاصله در متن پیکره است.
۳. این ویژگی در بخش ۵ـ۱ـ۱ـ توضیح داده شده‌است.

در خط‌های دیگری غیر از خط فارسی نیز به مشکلات املایی توجه شده‌است. ازجمله، در مقالهٔ ون هالتِرن و اوستیک (van Halteren and Oostdijk 2014)، برای شناسایی واژه‌هایی که در نوشتن آن‌ها از قوانین استاندارد نوشتاری تبعیت نشده و ممکن است در پردازش زبان طبیعی مشکل‌ساز شوند از تخمین خودکار استفاده شده و راه‌حلی نیز برای حل مشکل پیشنهاد شده‌است.

داسیگی و دیاب (Dasiqi and Diab 2011) نیز به تنوع و تشتت در نگارش واژه‌ها در گونهٔ محاوره‌ای عربی (گونهٔ مصری و شامی) و مشکلات ناشی از آن اشاره کرده و برای شناسایی این تنوع‌ها و کاهش مشکلات مربوط، به جست‌وجوی شباهت‌های معنایی بافتی پرداخته‌اند.

مشکلات ناشی از تنوع املایی موضوع پژوهش‌های دیگری نیز بوده‌است. ازجمله، پژوهش عرب‌مقدم و سنچال (Arab-Moghaddam and Senechal 2001) که در آن، به تأثیر چندآوایی و چندنویسگی[1] (نبودِ تناظر میان واج‌ها و نویسه‌ها) پرداخته شده و تفاوت در پردازش مهارت‌های ذکرشده در انگلیسی و فارسی ناشی از آن ذکر شده‌است.

به همین ترتیب، بلوچ (Baluch 2005) نوشته نشدنِ واکه‌های کوتاه در خط فارسی، تعدد نقطه‌ها (که منجر به شباهت حروف می‌شود)، ساخت صرفی زبان فارسی و نیز دوزبان‌گونگی را علاوه بر چندآوایی و چندنویسگی، از عوامل مؤثر در کاهش باسوادی دانسته‌است.

همان‌گونه که مشاهده می‌شود، اغلب مواردِ ذکرشده به ماهیت و ویژگی‌های خط فارسی اشاره دارد. در این مقاله نیز سعی بر آن بوده تا با بررسی عوامل ایجادکنندهٔ تنوع نگارشی در فارسی، تقسیم‌بندی نسبتاً متفاوتی به دست داده شود. بدین منظور، از روش داده‌بنیان، که در بخش بعدی به آن پرداخته می‌شود، استفاده شده‌است.

۳ـ پیکره‌های زبانی مورد استفاده

در روش‌های داده‌بنیان، وجود پیکره از اهمیت زیادی برخودار است. برای انجام این پژوهش، از داده‌های حاصل از ادغام سه پیکرهٔ زبانی استفاده شده‌است:

۱. پیکرهٔ بی‌جن‌خان[2] (بی‌جن‌خان ۱۳۸۳) که با حجم ۲/۵ میلیون واژه به‌صورت رایگان و برخط (آن‌لاین) موجود است، بخشی از پیکرهٔ بزرگ صدمیلیون‌کلمه‌ای است. متون

1. polyphony and polygraphy
2. http://ece.ut.ac.ir/dbrg/bijankhan/

تشکیل‌دهندهٔ این پیکره نسبتاً متنوع و شامل متون نوشتاری رسمی است. در این پیکره برچسب دستوری (واژ-نحوی) و معنایی واژه‌ها مشخص شده‌است.

۲. دادگان درختی وابستگی زبان فارسی (Rasooli et al. 2013) که گروه دادگان[1] آن را تهیه کرده‌است، با حجم بیش از پانصدهزار واژه رایگان و به‌صورت برخط موجود است. این پیکره متعادل و از تنوع متنی مناسبی برخوردار است. در این پیکره، مقوله‌های دستوری و نیز ساخت نحوی جمله براساس دستور وابستگی مشخص شده‌است. حجم قابل توجهی از پیکره را متن روزنامه‌ای تشکیل می‌دهد.

۳. پیکرهٔ تاک‌بانک[2] با بیش از پانصدمیلیون واژه متشکل از متون وبگاه‌ها و وبلاگ‌های فارسی است که گروه پژوهشی اشلوما آرگومان[3] آن را در مؤسسهٔ فنی ایلینوی[4] تهیه کرده‌است. قسمتی از این پیکره از تعامل کاربر و محیط وب حاصل شده و از این جنبه که تنوع سبکی و نگارشی در آن دیده می‌شود و از دادهٔ استانده، مانند دو پیکرهٔ قبلی، فاصله دارد بااهمیت است.

۴ـ روش رایانه‌ای برای یافتن تنوع نگارشی واژه‌ها

همان‌گونه که در مقدمه اشاره شد، هدف این مقاله دست‌یابی به چهارچوبی است برای یافتن موارد تنوع نگارشی واژه‌ها، که ازجمله چالش‌های زبان‌شناسی پیکره‌ای و تهیهٔ پیکره در زبان فارسی است. در همین راستا تلاش شده‌است تا با ارائهٔ روشی رایانه‌ای این موارد مشخص شود و طبقه‌بندی جدیدی از موارد مذکور ارائه گردد. قیومی و همکاران (۱۳۹۴) تلاش کرده‌اند با معرفی یک الگوریتم بتوانند تنوع نگارشی واژه‌ها را به‌طور خودکار از پیکرهٔ ترکیبی که در بخش ۳ ذکر شد استخراج کنند. الگوریتم معرفی‌شدهٔ آن‌ها شکل گسترش‌یافتهٔ الگوریتم فاصلهٔ لِونشتاین (Levenstein 1966) است. با استفاده از فاصلهٔ لِنشتاین می‌توان با درنظرگرفتنِ دو حالتِ جایگزینی و درج، فاصلهٔ دو واژه را محاسبه کرد. به‌طور پیش‌فرض، وزنِ هریک از فاصله‌های محاسبه‌شده ۱ است. در الگوریتم گسترش‌یافتهٔ قیومی و همکاران (همان)، وزن‌دهی تغییر کرده و وزن ۰/۱ برای فاصلهٔ بعضی از حروف خاص که سبب ایجاد تنوع نگارشی شده محاسبه

1. http://www.dadegan.ir/
2. http://www.sketchengine.co.uk/documentation/wiki/Corpora/TalkBankPersian
3. Shlomo Argamon
4. http://web.iit.edu/

می‌گردد. در سایر موارد، وزن ۱ مد نظر قرار می‌گیرد. هدف از وزن‌دهی به موارد خاص، آسان‌سازیِ جست‌وجوی این موارد است تا بتوان دادگانی حاصل از واژه‌های دارای تنوع نگارشیِ تهیه کرد.

۵ـ عوامل ایجادکنندهٔ تنوع نگارشی

در این مقاله، منظور از تنوع نگارشی مجموعه‌واژه‌هایی است که بر یک مصداق دلالت می‌کنند و دارای حداکثر شباهت آوایی هستند. در گروهی از این واژه‌ها، تغییری در تلفظ ایجاد نمی‌شود، مانند باتری / باطری و آباژور / آباژر، حال آنکه در واژه‌های گروه دیگر تلفظ‌های متفاوت دیده می‌شود، ازقبیل دشک / تشک و دهان / دهن. روشن است که طبق تعریف بالا، مقولهٔ دستوری نباید تغییر کند، یعنی اگر دو واژه با شرایط یادشده دارای دو مقولهٔ دستوری متفاوت باشند (مانند آزاد / آزادی و یا ظلم / ظالم)، مصداق تنوع نگارشی محسوب نمی‌شوند.

براساس پیکرهٔ استفاده‌شده در این پژوهش و با بررسی حدود ۱۰۰۰ واژه، عوامل مختلفی که منجر به پیدایش تنوع نگارشی می‌شود شناسایی و دو نوع تقسیم‌بندی، یکی دانه‌ریز و دیگری دانه‌درشت[1]، برای واژه‌ها ارائه شده‌است.

عوامل ایجادکنندهٔ تنوع نگارشی در تقسیم‌بندی دانه‌ریز به شرح زیرند:

۵ـ۱ـ جدانویسی یا سرهم‌نویسی

۵ـ۱ـ۱ـ حروف چسبان

جدانویسی یا سرهم‌نویسی به ویژگی‌های نوشتاریِ خط فارسی مربوط می‌شود و به‌نظر می‌رسد مهم‌ترین مشکل در زمینهٔ تنوع نگارشی، بحث فاصله است، زیرا ماهیت برخی از حروفِ فارسی به گونه‌ای است که هم از سمت راست و هم از سمت چپ به حرف مجاور خود می‌چسبند، مانند «س»، «ص» و «ب». در واژه‌هایی که در جایگاه آغازین یا میانی آن‌ها یکی از این حروف وجود دارد، ممکن است بودن یا نبودن فاصله سبب پیدایش صورت‌های گوناگون نوشتاری شود، مانند می‌شود / می@شود.

۵ـ۱ـ۲ـ حروف غیرچسبان

1. coarse-grain (منظور از دانه‌درشت «دربردارندهٔ اطلاعات کلی» است)

تعدادی دیگر از حروف فقط از سمت راست قابلیت اتصال دارند، مانند «آ»، «ا»، «د»، «ذ»، «ر»، «ز»، «ژ» و «و». این امر سبب به‌وجودآمدنِ صورت‌های متنوع و گوناگون، و گاه مبهم یا بی‌معنا، در املای کلمه می‌شود[۱].

در تمام واژه‌های بررسی‌شده (اعم از بسیط و غیربسیط) که در جایگاه آغازین و میانی دارای حروف «آ»، «ا»، «د»، «ذ»، «ر»، «ز»، «ژ» یا «و» هستند (یعنی حروفی که از سمت چپ به حرف بعدی خود نمی‌چسبند)، ممکن است جدانویسی رخ دهد، مانند ا‌گر / ا@گر و داد / دا@د. به بیان دیگر، در این واژه‌ها، فاصله در میان کلمه بعد از حروف غیرچسبان می‌تواند سبب پدید آمدن صورت‌های نگارشی متعدد شود. همان‌گونه که در این دو نمونه مشاهده می‌شود، حضور فاصله پس از «ا» ضرورتی ندارد.

۵ـ۱ـ۳ـ مرز تکواژ

در واژه‌های غیربسیط (شامل واژه‌های مشتق، مرکب و مشتق‌مرکب)، علاوه بر موارد ذکرشده در بالا، جدانویسی در مرز تکواژ نیز رخ می‌دهد، صرف‌نظر از اینکه تکواژ موردنظر به یکی از حروف «آ»، «ا»، «د»، «ذ»، «ر»، «ز»، «ژ» یا «و» ختم شود. واژه‌های این گروه ممکن است به‌صورت سرهم، جدا از هم (با فاصله) و یا بدون فاصله (با نیم‌فاصله) نوشته شوند، مانند میشود / می@شود / می‌شود و یا لشکرکشی / لشکر@کشی. به‌عبارت دیگر، در این‌گونه واژه‌ها فاصله در میان واژه و یا در مرز تکواژ وجود دارد. در این حالت، پدید آمدن صورت‌های گوناگون نگارشی می‌تواند به‌علت شتاب و بی‌دقتی نویسنده یا حروف‌نگار و یا نبودِ شیوه‌نامه‌ای واحد برای اِعمال صحیح فاصله باشد.

همچنین جدانویسی یا سرهم‌نویسی می‌تواند ناشی از تغییر آوایی و یا براثر حذفِ یک یا چند حرف از واژه رخ دهد (← بخش ۵ـ۴ـ و ۵ـ۵ـ).

۵ـ۲ـ حضور «ا» به جای «آ»

در تعدادی از کلمات، ننوشتن علامت مد برروی حرف الف (یا به عبارتی، نوشتن «ا» به‌جای «آ») سبب پدید آمدن گونه‌های مختلفی در نگارش می‌شود. صورت‌های املاییِ

۱. علاوه بر عوامل یادشده، جدا و سرهم‌نویسی ممکن است در مواردی سلیقه‌ای باشد، به‌طوری‌که می‌توان گفت در گذشته، تمایل به سرهم‌نویسی وجود داشته و اکنون گرایش به جدانویسی رایج‌تر است.

پدیدآمده ممکن است دو واژهٔ معنادار باشند، مانند قرآن / قران. همچنین این امکان وجود دارد که یکی از آن‌ها در زبان فارسی بی‌معنا بوده (مانند آبان و ابان) و یا شبیه به واژه‌ی معنادار دیگری باشد (همچون آتو / اتو) و تنها کار جست‌وجوی رایانه‌ای را با مشکل مواجه سازد.

۵ـ۳ـ همزه

نوشتن همزه برروی کرسی‌های گوناگون (مانند «ئ»، «ؤ»، «أ»، و «إ» در کلماتی مانند مسؤول / مسئول و هیأت / هیئت)، جایگزینی آن با حرف «ی» (مانند آیین به‌جای آئین) و یا نوشتن واژه بدون همزه (چه با کرسی، مانند رای به جای رأی و چه بدون کرسی، مانند سو / سوء) نیز می‌تواند منجر به پیدایش صورت‌های گوناگون برای یک واژه شود.

۵ـ۴ـ عوامل واژی ـ آوایی: جایگزینی

گاه تغییرات آوایی، به‌ویژه در گفتار غیررسمی، در ایجاد تنوع املایی نقش دارند. این تغییرات به گونه‌ای کم‌وبیش قاعده‌مند رخ می‌دهند و در املا به‌صورت جایگزینیِ یک حرف با حرف دیگر بروز می‌کنند که این امر می‌تواند سبب تغییراتی در املایِ واژه شود. تنوع‌های نگارشی از نوع مباحث آوایی را می‌توان در سه زیربخش در نظر گرفت:

۵ـ۴ـ۱ـ تغییر سبک

در مواردی، تغییر آوایی (جایگزینی) را می‌توان به‌صورت درون‌واژه‌ای در عناصر غیرفعلی و فعلی مشاهده کرد که موجب تبدیل صورتی از یک واژه به واژهٔ دیگر می‌شود. این تغییر عمدتاً سبب تغییر در سبک واژه از صورت معیار به گفتاری یا عامیانه (مانند اگر ← اگه و همان ← همون) و به‌ندرت از معیار به ادبی (مانند راه ← رَه) می‌شود[1].

در فعل‌ها (به‌جز فعل‌های ربطی)، به‌نظرمی‌رسد این تغییرات تابع نظم خاصی است، از جمله تغییر در ستاک فعل و در شناسه‌ها. در برخی فعل‌ها، مانند آوردن، خواستن و

۱. گونه‌های گفتاری و عامیانه عمدتاً در گفتار کاربرد دارند، اما برخی نویسندگان برای حفظ سبکِ گفتاری یا به‌کارگیریِ سبکِ ویژهٔ خود، از این گونه‌ها در متون استفاده می‌کنند.

دادن، تغییرات آوایی هم در ستاک حال فعل و هم در شناسهٔ آن رخ می‌دهد[۱]، مانند می‌آورم ← می‌آرم / میارم و می‌شَوید ← میشین. حال آنکه در برخی دیگر، تغییر فقط در شناسه است، مانند بخندند ← بخندن، می‌خورد ← می‌خوره.

لازم به ذکر است تغییر آوایی در فعل‌های پربسامد که متعلق به سطح معیار زبان هستند رخ می‌دهد و فعل‌هایی با سیاق رسمی و ادبی (مثل رهانیدن، روییدن، بوییدن و گشتن / گردیدن) که در محاوره کاربرد چندانی ندارند دستخوش چنین تغییراتی نمی‌شوند.

۵ـ۴ـ۲ـ ایجاد ابهام

تغییر آوایی (جایگزینی) در مواردی سبب پیدایش کلمات هم‌نویسه و چندمعنا و درنتیجه، ابهام در معنای آن‌ها می‌شود، به‌طوری‌که مثلاً صورت زبانیِ بره ممکن است هم به‌معنای «برود» و هم «بَرِّه» (= بچهٔ گوسفند) باشد.

۵ـ۴ـ۳ـ ایجاد صورت غیربسیط

در برخی موارد نیز، تغییر آوایی (جایگزینی) از مرز درون واژه فراتر رفته و برون‌واژه‌ای محسوب می‌شود. به عبارت دیگر، گاه این تغییر مرز دو واژه را می‌شکند و آن‌ها را به یک عنصر ترکیبی تبدیل می‌کند. معمولاً در این حالت با تبدیل یک تکواژ آزاد به تکواژ وابسته و الصاق آن به واژهٔ دیگر روبه‌رو هستیم، مانند مَنو (= من + را / من + وَ) و منه (= من + است).

براساس آنچه ذکر شد، در نمونهٔ آخر، به‌جای اینکه در متن از صورت «من است» استفاده شود، شاهد کاربرد «منه» هستیم که هم‌زمان تغییر آوایی‌ای را در بر گرفته و تکواژی آزاد و مستقل را به یک صورتِ وابسته تبدیل کرده‌است.

۵ـ۵ـ عوامل واژی ـ آوایی: حذف

حذف یک حرف از واژه نیز مانند جایگزینی ممکن است سبب تغییراتی در املا شود:

۵ـ۵ـ۱ـ تغییر سبک

---

۱. این نتیجه‌گیری حاصل بررسی پانصد فعل بسیط فارسی است که در کتاب فعل بسیط فارسی و واژه‌سازی (علاءالدین طباطبایی) فهرست شده‌است.

گاه حذف یک یا چند حرف از واژه سبب تغییر سبک آن می‌شود، مانند شی (به‌جای صورت کاملِ بشوی یا شَوی) و یا تبدیل راه به ره. در مثال اول، تغییر سبک از سطح معیار به گفتاری و در دومی تغییر از سطح معیار به ادبی است.

۵ـ۵ـ۲ـ ایجاد ابهام

در دستهٔ دیگری از واژه‌ها علاوه‌بر حذف، با فرایند چندمعنایی و ابهام‌زایی معنایی نیز مواجهیم. برای مثال، در حذف «و» از شوی و تبدیل آن به شی، ممکن است شی با خوانشی متفاوت، با واژهٔ شیء خلط شود.

۵ـ۵ـ۳ـ ایجاد صورت غیربسیط

در مواردی، حذف یک یا چند حرف از یک واژه یا تکواژ سبب پدید آمدن عنصری وابسته می‌شود که با اتصال به واژه‌ای دیگر، صورتی بزرگ‌تر از یک واژه ایجاد می‌کند. برای مثال، درختا براثر حذف «ه» از تکواژ ـها یا «ن» از ـان و اتصال هریک از این دو تکواژ به واژهٔ درخت ایجاد شده است. به همین ترتیب، آنست در نتیجهٔ حذف «ا» از واژهٔ است و به‌جای صورت کاملِ «آن است» پدید آمده‌است.

۵ـ۶ـ سایر موارد

نویسندگان این مقاله در بررسیِ داده‌ها با مواردی مواجه شده‌اند که یا در زبان فارسی بی‌معنا به نظر می‌رسند، مانند «میک»، «دای» و «ههها»، و یا ممکن است بخشی از یک جمله یا آیهٔ عربی باشند، مانند «لها»، و یا نام تجاریِ کم‌وبیش نامأنوسی در فارسی تلقی شوند، از قبیل «مون» و «مای». بدیهی است که چنینَ صورت‌هایی در فرهنگ جایی ندارند و لذا در اینجا تحت عنوان «سایر موارد» طبقه‌بندی شده‌اند.

۶ـ تحلیل و طبقه‌بندی داده‌ها

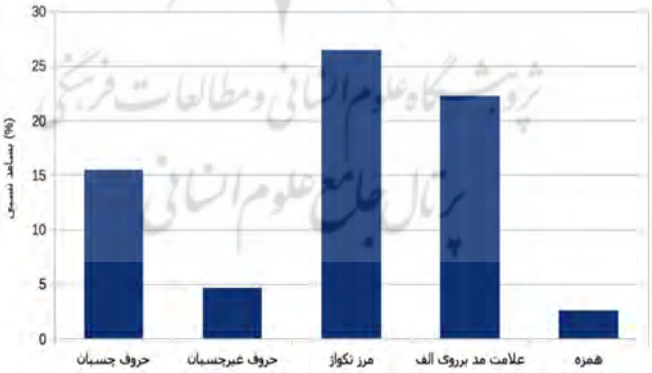
در ادامهٔ مقاله، پس از بررسی هزار واژهٔ دارای تنوع نگارشی که از پیکرهٔ ترکیبی استخراج شده‌اند، به تحلیل و طبقه‌بندی داده‌ها می‌پردازیم. لازم به ذکر است که مجموع بسامد این ۱۰۰۰ واژه به ۳۶,۶۵۲,۸۲۹ رسیده‌است. داده‌های حاصل از الگوریتم گسترش‌یافتهٔ فاصلهٔ لونشتاین از منظر زبان‌شناسی تحلیل کیفی شده‌است. در ادامه، پس از دسته‌بندی ریزدانهٔ تنوع نگارشی، عوامل کلی که سبب چنین تنوعی می‌شود توضیح داده و از نظر آماری بررسی شده‌است.

جدول ۱ نشان می‌دهد که در یک دسته‌بندی دانه‌ریز، تنوع نگارشی در زبان فارسی را می‌توان به دسته‌های زیر تقسیم کرد:

ردیف	عوامل ایجاد تنوع نگارشی		نمونه
۱	جدانویسی و سرهم نویسی	حروف چسبان	نمی@شه
		حروف غیرچسبان	برا@ی
		مرزتکواژ/ واژه	لشکر@کشی
۲	مد		قران
۳	همزه		مسؤول
۴	تغییر آوایی (جایگزینی)	تغییر سبک	اگه
		ابهام	بره
		صورت غیربسیط	منو
۵	حذف	تغییر سبک	بشی
		ابهام	شی
		صورت غیربسیط	آنست
۶	سایر		لها

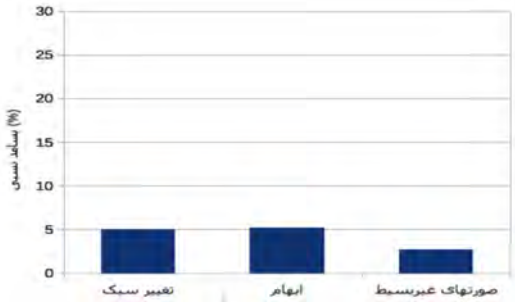
جدول ۱: تقسیم‌بندی دانه‌ریز مقولات تنوع نگارشی

موارد جدانویسی یا سرهم‌نویسی که شامل حروف چسبان در میان واژه، حروف غیرچسبان در میان واژه و فاصله در مرز تکواژ است به همراه بودن یا نبودن علامت مد و مسئلهٔ همزه را می‌توان زیر عنوان «رسم‌الخط» به شرح زیر در نمودار ۱ نشان داد:

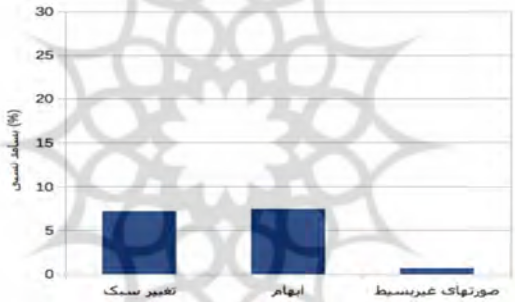


نمودار ۱: دسته‌بندی دانه‌ریز موارد تنوع نگارشی مربوط به رسم‌الخط

بسامد تغییر سبک، ابهام و ایجاد صورت‌های غیربسیط ناشی از تغییرات آوایی (جایگزینی) در نمودار ۲ و موارد مشابه که در نتیجهٔ حذف ایجاد شده‌اند در نمودار ۳ آمده‌اند.



نمودار ۲: دسته‌بندی دانه‌ریز موارد تنوع نگارشی مربوط به تغییرات آوایی (جایگزینی)



نمودار ۳: دسته‌بندی دانه‌ریز موارد تنوع نگارشی مربوط به حذف

بر این اساس، در تقسیم‌بندی دانه‌درشت با ادغام گروه‌های مشابه، مقولات تنوع نگارشی را می‌توان در جدول ۲ به شرح زیر نمایش داد.

مقولهٔ دانه‌درشت	بسامد نسبی (٪)
رسم‌الخط	۷۱/۳۵
واژی ـ آوایی	۲۸/۱۳
سایر موارد	۰/۵۲

جدول ۲: دسته‌بندی دانه‌درشت مقوله‌ها

لازم به ذکر است که هریک از صورت‌ها (تنوع‌های نگارشیِ پدیدآمده) ممکن است در بیش از یک دسته قرار گیرند، مانند «منو» که براثر تغییر آواییِ هم دستخوش تغییر

سبک شده (از رسمی به گفتاری)، هم صورتی است غیربسیط (متشکل از مـن + را یـا من + و) و هم دارای ابهام (مَنو یا مِنو) می‌باشـد. تقریبـاً در تمـام گـروه‌هـای یادشـده صورت‌های دارای ابهام دیده می‌شود. همچنین، براساس آنچه در جدول ۲ آمده اسـت، مشخص شد که بیشترین تنوع نگارشی مربوط به رسم‌الخـط و کمتـرین آن مربـوط بـه تغییرات آوایی است.

مقولهٔ دانه‌درشت	بسامد نسبی (٪)
رسم‌الخط	۷۱/۳۵
واژی ـ آوایی	۲۸/۱۳
سایر موارد	۰/۵۲

جدول ۱: دسته‌بندی دانه‌درشت مقوله‌ها

۷ـ سخن آخر

همان‌گونه که پیش‌تر گفته شد، به دلیل ویژگی‌های خط فارسـی و نیـز براثـر تغییـرات آوایی در سطح گفتار فارسی، مشکل تنوع نگارشی در این زبان بسیار جـدی اسـت. در این زمینه، شاید جدی‌ترین مشکل بحثِ جدانویسی یا سرهم‌نویسی[1] باشـد. در ایـن دو مبحث، در واژه‌های غیربسیط، اگر فاصله در مرز تکواژ رخ دهـد مـی‌تـوان بـا تـدوین شیوه‌نامه‌ای مناسب، مواردِ فاصله، نیم‌فاصله یا سرهم‌نویسـی را تعیـین کـرد. البتـه لازم است در پیکره اطلاعات ساختواژی (مربوط به تکواژها) برای هریک از واژه‌ها گنجانده شده باشد. درصورتی‌که فاصله در مرز تکواژ نباشد، شاید رعایت یک‌دستی و آمـوزش شیوهٔ خط فارسی و ماهیت آن مفید واقع شود.

علاوه بر این‌ها، در شناسایی موارد تنوع نگارشی لازم است که به‌جز عوامل مربوط به رسم‌الخط و شیوهٔ نوشتن، عوامل زبانی دیگر، مانند تغییرات آوایی و حـذف، نیـز در نظر گرفته شود. همچنین در مواردی که تنوع نگارشی ناشی از تغییـرات آوایـی اسـت، می‌توان با دسته‌بندیِ این تغییرات و کشف فرایندهای به‌کاررفته و نیز تهیـهٔ فهرسـتی از واژه‌های دارای تنوعِ نگارشی، این موارد را در مواد آموزشی برای زبان‌آموزان، بـه‌ویـژه در تهیهٔ فرهنگ تنوع‌های نگارشی که هدف و انگیزهٔ اصـلی ایـن پـژوهش بـوده‌اسـت، گنجاند. براساس داده‌های تحلیل‌شده، موارد مربوط به رسم‌الخط مهم‌ترین عامل ایجـاد

۱. در اینجا تنها نمونه‌هایی که در پیکره از فراوانی قابل‌توجهی برخودار بوده‌اند بررسی شده‌اند و مواردی از قبیل نبود تناظر واج‌ها و نویسه‌ها در خط فارسی بررسی نشده‌اند.

تنوع نگارشی در فارسی به‌شمارمی‌رود. این امر نشان‌دهندهٔ اهمیت و نقشِ پیکره در بررسی‌های زبان‌شناسی پیکره‌ای و زبان‌شناسی رایانشی[1] است.

نکتهٔ آخر آنکه این نتیجه با بررسی هزار واژه به‌دست‌آمده‌است. بدیهی است که با بررسی داده‌های بیشتر و تحلیل آن‌ها می‌توان به نتایج جدیدتر و دقیق‌تری دست یافت.

منابع

افتخاری، اسماعیل (۱۳۸۴)، «درآمدی بر خط‌شناسی و نظم‌گریزی نوشتار فارسی»، مجلهٔ دانشکدهٔ علوم‌انسانی دانشگاه اصفهان، شمارهٔ ۱۴: ۵۸، صفحه‌های ۳۳ـ۶۲.

آشوری، داریوش (۱۳۶۵)، «چند پیشنهاد دربارهٔ نگارش و خط فارسی»، مجلهٔ نشر دانش، شمارهٔ ۶: ۳۶، صفحه‌های ۲ـ۸.

بی‌جن‌خان، محمود (۱۳۸۳)، «نقش پیکرهٔ زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای»، مجلهٔ زبان‌شناسی، سال ۱۹، شمارهٔ ۲، صفحه‌های ۴۸ـ۶۷.

ستوده، هاجر و زهره هنرجویان (۱۳۹۱)، «مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تاثیرات آن‌ها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات»، مجلهٔ کتابداری و اطلاع‌رسانی، سال ۱۵، شمارهٔ ۴، صفحه‌های ۵۹ـ۹۲.

ستوده، هاجر و زهره هنرجویان (۱۳۹۳)، «بررسی تنوع الگوهای نگارش فارسی و تأثیر آن بر جامعیت بازیابی اطلاعات (مطالعهٔ موردی: پیکرهٔ همشهری)»، مجلهٔ کتابداری و اطلاع‌رسانی، سال ۱۷، شمارهٔ ۲، صفحه‌های ۳۱ـ۴۹.

سمیعی گیلانی، احمد (۱۳۶۶)، آیین نگارش، چاپ اول، تهران، مرکز نشر دانشگاهی.

صادقی، علی‌اشرف و زهرا زندی‌مقدم (۱۳۹۱)، فرهنگ املایی خط فارسی، ویراست دوم، تهران، فرهنگستان زبان و ادب فارسی.

صدرامیرجانلو، اصغر (۱۳۸۱)، «کاستی‌های خط فارسی و پیامدهای آن در زبان فارسی»، مجلهٔ دانشکدهٔ علوم‌انسانی دانشگاه تهران، تابستان و پاییز، صفحه‌های ۳۹۳ـ۴۰۷.

طباطبایی، علاءالدین (۱۳۷۶)، فعل بسیط فارسی و واژه‌سازی، تهران، مرکز نشر دانشگاهی.

طیب، محمدتقی (۱۳۷۱)، «هم‌نگاری در خط فارسی»، مجلهٔ دانشگاه اصفهان، شمارهٔ ۱۴، صفحه‌های ۱۵ـ۳۸.

قیومی، مسعود، ساغر شریفی و مرضیه صناعتی (۱۳۹۴)، «تنوع نگارشی در زبان و تهیهٔ خودکار دادگان املایی از پیکرهٔ زبانی مبتنی بر وب»، مجموعه‌مقالات نخستین کنفرانس بین‌المللی وب‌پژوهی، تهران، دانشگاه علم و فرهنگ.

1. computational linguistics

مرتضایی، لیلا (۱۳۸۰)، «مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات»، مجلهٔ اطلاع‌رسانی، ۱۷: ۱ و ۲، صفحه‌های ۱۹۳ـ۲۰۰.

مشیری، مهشید (۱۳۶۶)، فرهنگ فارسی آوایی ـ املایی، تهران، کتابسرا.

نثری، موسی (۱۳۱۴)، «با خط کنونی فارسی چه باید کرد؟»، مجلهٔ مهر، ۳: ۲، صفحه‌های ۲۰۰ـ۲۰۵.

Arab-Moghaddam, Narges and Monique Senechal (2001), "Orthographic and phonological processing skills in reading and spelling in Persian / English bilinguals", *International Journal of Behavioral Developement*, 25 (2), pp. 140-147.

Baluch, Bahman (2005), "Persian orthography and its relation to literacy", *Handbook of Orthography and literacy*, Joshi, R. Malatesha and P. G. Aaron eds., Lawrence Erlbaum Associates, London, pp. 365-376.

Bijankhan, M., and J. Sheykhzadegan and M. Bahrani and M. Ghayoomi, "Lessons from building a Persian written corpus: Peykare," *Language Resources and Evaluation*, vol. 45, no.2, pp.143-164, 2011.

Dasigi, Pradeep and Mona Diab (2011), "CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic", Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, November 8-13, pp. 318-326.

Ghayoomi, Masood & Saeedeh Momtazi (2009), "Challenges in developing Persian corpora from on-line resources", Proceedings of 2009 IEEE International Conference on Asian Language Processing, Singapore, pp. 108-113.

Ghayoomi, Masood and Saeedeh Momtazi and Mahmood Bijankhan (2010), "A study of corpus development for Persian", In *International Journal on Asian Language Processing*, 20(1), pp. 17-33.

Levenshtein, Vladimir I. (1966), "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, 10(8), pp. 707-710.

Rasooli, Mohammad Sadegh and Manouchehr Kouhestani and Amirsaeid Moloodi (2013), "Development of a Persian syntactic dependency treebank", Proceedings of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies*, Atlanta, USA. June 9-14, pp. 306-314.

Shamsfard, Mehrnoosh (2011), "Challenges and open problems in Persian text processing", Proceedings of the 5th Language and Technology Conference: *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, pp. 65-69.

van Halteren, Hans and Nelleke Oostdijk (2014), "Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens", *Journal for Language Technology and Computational Linguistics*, 29(2), pp. 97-12.