شماره مستند: ۱۹۰/۲۵۳۷/۱/۹

جمهوری اسلامی ایران

دبیرخانه شورای عالی اطلاع رسانی

تدوین نواقص دستورالعمل املایی مصوب فرهنگستان به منظور ایجاد خطایاب املایی صرفی و نحوی زبان فارسی

نسخه ۱.۰

دانشگاه علم و صنعت ایران

فروردین ۸۸

فهرست مطالب

۱ پیشگفتار

با توجه به سرعت رشد فن‌آوری اطلاعات در کشورهای مترقی جهان، هر لحظه‌ای که از دست می‌رود برای صنعت نرم‌افزاری ایران حکم ماه‌ها عقب‌افتادگی را دارد. به همین دلیل باید تلاش نمود تا موانع رشد این صنعت به سرعت بر طرف گردند. یکی از این موانع، عدم وجود دستور خطِ جامعی متناسب با نیازهای سامانه‌های پردازش متن‌های رقمی است.

مهم‌ترین اقدام رسمی انجام شده برای حفظ چهره‌ی خط فارسی توسط فرهنگستان زبان و ادب فارسی از سال ۱۳۷۲ انجام گرفت که تلاشی تحسین برانگیز در این زمینه بود. اما هدف اصلی نگارش از آن، یکسان‌سازی چهره‌ی کلمات برای رایانه نبود. در نتیجه در بسیاری از موارد دست کاربران برای انتخاب شکل کلمات باز گذاشته شده بود که نتیجه‌ی آن چیزی جز پیدایش ابهام در تشخیص کلمات توسط رایانه نیست. این دستور خط، مبنای اصلی بررسی‌های انجام گرفته در گزارش حاضر است. محتوای این گزارش نتیجه‌ی بررسی مورد به مورد دستور خط فارسی و مشکلات آن در سامانه‌های رایانه‌ای می‌باشد.

وجود ارتباط متقابل میان زبان‌دانان و فن‌آوران حوزه‌ی رایانه یکی از نیازهای اصلی جامعه‌ی اطلاعاتی امروز و رشد صنعت پردازش الکترونیک در میهن عزیزمان می‌باشد. اگر در خط فارسی تغییری برای همگام شدن با این کاروان روی ندهد، میزان عقب ماندگی ما در فن‌آوری اطلاعات از سایر کشورها غیر قابل جبران خواهد شد. یعنی زمانی فراخواهد رسید که نرم‌افزارهای خارجی، خلاصه‌ی متون خارجی را در کسری از ثانیه استخراج می‌نمایند و روزنامه‌ها و مقالات با سرعت زیادی تولید می‌شوند در حالی که ما در ایران حتی کار غلط‌یابی املاء را نیز با خطای بالا و به کُندی انجام می‌دهیم. زمانی که موتورهای جستجوی خارجی می‌توانند در زمان جستجوی یک لغت، مترادف‌ها، ریشه‌ها و سایر مشتقات آن را نیز هم‌زمان جستجو کنند تا کاربران به سادگی به اطلاعات مورد نیاز خود دست پیدا کنند، ما تازه در حال رفع مشکل کدِ نویسه‌ی "ی" و یا مشکل اجزای کلمات خود هستیم و یک جستجوی ساده را نیز نمی‌توانیم در وب‌گاه‌های رسمی کشورِ خویش با موفقیت به انجام برسانیم.

امید است که این گزارش بتواند اهمیت موضوع یکسان‌سازی طرز نگارش کلمات و نیز موارد ابهام‌زای موجود در دستور خط فارسی را به حد مطلوب برای خوانندگان گرامی مشخص نماید.

۲ فصل اول- دستور خط فارسی و رایانه

۲.۱ مقدمه

با گسترش استفاده از رایانه‌ها در تولید متون فارسی، چالش‌های جدیدی در تعامل کاربران با این سامانه‌ها به وجود آمده است که در این میان نحوه‌ی نگارش کلمات یکی از عمده‌ترین مسائل این حوزه می‌باشد. اگر کلمات فارسی با صورت یکسانی نوشته نشوند، به مرور زمان خط فارسی چهره‌ی خود را از دست می‌دهد، سرعت درک معنای واژگان برای مخاطبان کاهش می‌یابد، امکان اصلاح صورت‌های غلط کمتر می‌شود و در نهایت متنی که توسط کاربران نوشته می‌شود به هیچ وجه قابل تحلیل توسط سامانه‌های رایانه‌ای نخواهد بود. ماهیت رایانه‌ها به آن‌ها این امکان را نمی‌دهد که به سادگیِ کاربران انسانی، به معنای واژگان پی‌ببرند. برای رایانه، واژه‌ی "حقیقتجو" متفاوت از واژه‌ی "حقیقت‌جو" و نیز "حقیقت جو" است.

مهم‌ترین اقدام رسمی انجام شده برای جلوگیری از زوال خط فارسی توسط فرهنگستان زبان و ادب فارسی از سال ۱۳۷۲ انجام گرفت که تلاشی قابل تقدیر در این زمینه بود. این دستور خط در متنی ۶۰ صفحه‌ای تدوین شده است که سرفصل‌های آن تعیین کننده نحوه‌ی نگارش واژگان، ترکیب آن‌ها، نگارش واژگان مرکب، استعمال همزه و نیز لغات دو املائی می‌باشند. این دستور خط، مبنای اصلی بررسی‌های انجام گرفته در گزارش حاضر است. محتوای این گزارش نتیجه‌ی بررسی مورد به مورد دستور خط فارسی و مشکلات آن در سامانه‌های رایانه‌ای می‌باشد.

بخش‌های بعدی این فصل عبارت هستند از مروری بر ویژگی‌های زبان فارسی در پردازش رایانه‌ای، تاثیر مهارت کاربران در نوشتار رایانه‌ای، زبان فارسی از دید رایانه و در نهایت، دسته‌بندی عمده‌ترین مشکلات این حوزه.

۲.۲ ویژگی‌های خط فارسی در پردازش رایانه‌ای

خط فارسی نیز مانند سایر خط‌های دنیا ویژگی‌هایی دارد که پردازش آن را توسط رایانه مشکل می‌نمایند. این ویژگی‌های ذاتی از آنجا ناشی می‌گردند که در زبان‌های طبیعی بسیاری از واژگان از ترکیب با واژگان دیگر ساخته می‌شوند. اگر در نگارش کلماتی که از این ترکیبات به وجود آمده‌اند قوانین مشخصی رعایت نشود، لغات حاصل ممکن است معنایی متفاوت از آنچه در آغاز مورد انتظار بوده است پیدا کنند.

از دیگر سو، تفاوت‌هایی نیز میان فارسی و سایر زبان‌های طبیعی دنیا وجود دارند. در واقع نحوه‌ی ساخت کلمات و اتصال واژها به هم در فارسی، منجر به پدید آمدن دسته‌ی دیگری از مشکلاتِ مربوط به پردازش متون در این زبان می‌گردد. از این ویژگی به عنوان ویژگی اختصاصی زبان فارسی نام خواهیم برد و آن را در بخش بعدی مورد بررسی قرار خواهیم داد.

۲.۲.۱ ویژگی‌های عمومی

همان‌طور که پیش از این عنوان شد، پردازش لغوی تمام زبان‌های طبیعی برای رایانه‌ها امری دشوار می‌باشد. اولین دلیل آن این است که ترکیب کردن کلمات با هم، منجر به تشکیل واژگانی می‌شود که ممکن است در اثر بی‌دقتی کاربران از دید رایانه به دو یا چند شکل مختلف خوانده شوند. مثلاً در جایی که منظور نویسنده "سیب‌زمینی" است، اگر در اثر بی‌دقتی لغت "سیب زمینی" نوشته شود، رایانه قادر به تشخیص لغت اصلی نخواهد بود.

دومین دلیل پیچیدگی پردازش لغوی زبان‌های طبیعی، ترکیب کردن لغت‌ها با هم و تولید لغت‌هایی است که حاوی اطلاعاتی در مورد مالکیت، جمع یا مفرد بودن و ... هستند (مانند "کتاب‌هایشان"). این لغات جدید در لغت‌نامه وجود ندارند اما معنای آن‌ها همان معنایی است که در لغت اولیه وجود داشته است. از دید رایانه تنها در صورتی دو لغت با هم یکی هستند که به یک شکل نوشته شده باشند.

سومین دلیل مشکل بودن تفسیر لغات از دید رایانه، آن است که برخی از قوانین تولید کلمه در زبان‌های طبیعی، می‌توانند لغاتی به وجود آورند که در لغت‌نامه‌ها وجود ندارند مانند "بازنگریسته". از دیگر سو اگر رایانه بتواند تمام قوانین ساخت لغات را در خود جای دهد، در آن صورت لغاتی که زبان پتانسیل تولید آن را دارد اما گویش‌ورانِ زبان تاکنون آن‌ها را استعمال نکرده‌اند نیز در زمره‌ی لغات مورد تائید رایانه قرار خواهند گرفت.

بنابراین پردازش کلماتِ یک متن به خودی خود رایانه را دچار مشکلاتی در تشخیص لغات می‌کند. این‌ها همگی در صورتی هستند که اشتباهات املائی کاربران در نظر گرفته نشود و نیز تمام کاربران اصول نسبتاً یکسانی را در نگارش خود به کار برند.

۲.۲.۲ ویژگی‌های اختصاصی زبان فارسی

در این بخش ویژگی‌های زبان فارسی که کار پردازش رایانه‌ای را بر روی آن مشکل می‌نماید مورد بررسی قرار خواهد گرفت. لازم به ذکر است که مشکلاتی که این ویژگی‌ها ایجاد می‌کنند مربوط به ماهیت زبان فارسی هستند و لزوماً با تغییر خط فارسی حل نمی‌شوند.

ساخت کلمات در زبان فارسی از قواعدی پیروی می‌کند که متفاوت با زبان نسبتاً فراگیر انگلیسی است. در زبان انگلیسی (به عنوان نمونه)، اضافه شدن پسوندهای متعدد به یک لغت استعمال بسیار محدودی دارد اما ما در فارسی بسیاری از ترکیبات خود را از طریق اتصال تعدادی وند یا واژه‌ی مستقل به واژهای دیگر می‌سازیم مانند "شیرینی‌خوران"، "خداشناس"، "حقیقت‌جو"، "آب‌سردکن" و .... در واقع ما در فارسی، با مشکل جدی فاصله‌گذاری رو به رو هستیم. ساختارهای تولید واژگان ما در بسیاری از موارد از ترکیب چند واژه‌ی مستقل و با معنی، در کنار یکدیگر به وجود آمده‌اند در حالی که اگر میان این اجزاء فاصله درج شود، تبدیل به دو لغت و اگر نیم‌فاصله درج شود تبدیل به یک لغت خواهند شد. این مسئله در انگلیسی بسیار بسیار کمتر رخ می‌دهد.

در زبان فارسی، ما ضمائر مفعولی و نشانه‌های جمع را به لغات متصل می‌کنیم. این مسئله نیز منجر به پیچیدگی تفکیک لغات برای رایانه می‌شود زیرا حروف ضمائر مفعولی و نشانه‌های جمع با تعدادی از وندها و نیز واژگان فارسی مشابهت دارند در نتیجه رایانه نمی‌تواند به سادگی در مورد بخش اصلی لغت قضاوت کند. به عنوان نمونه، "ان" در "درختان" نشانه‌ی جمع است در حالی که "ان" در "خوران" نشانه‌ی صفت فاعلی می‌باشد.

در زبان فارسی قواعد تبدیل بن مضارع به بن ماضی و بالعکس به صورت قانون‌مند وجود ندارد، در نتیجه رایانه برای آنکه بتواند لغات حاصل از بن مضارع یا بن ماضی را تشخیص دهد (مثلاً برای آنکه بداند "ان" در لغت "خوران" نشانه‌ی جمع است یا نشانه‌ی صفت فاعلی)، هیچ راه مشخصی ندارد. در بسیاری از کاربردهای پردازشی زبان، مثلاً در نرم‌افزار موتورهای جستجو (مانند Google)، رایانه‌ها باید تمام لغات متن را استخراج نمایند. در این برنامه‌ها، نشانه‌های جمع یا ضمایر مفعولی باید از لغت حذف شوند زیرا دارای اهمیت نمی‌باشند. از دیگر سو داشتن کارایی بالا (سرعت بالا) یکی از ملزومات این نرم‌افزارها است. تشخیص اینکه یک نشانه (مانند "ان") در یک لغت، یک نشانه‌ی قابل حذف است (مانند نشانه‌ی جمع) یا خیر (مانند نشانه‌ی صفت فاعلی)، نیازمند جستجو در لغت‌نامه‌ای است که باید حاوی بن‌های مضارع و ماضیِ تمام افعال باشد. همچنین رایانه باید حالات ممکن را بررسی نماید تا مطمئن شود که یک لغت را به درستی از نشانه‌های اضافه‌اش جدا نموده است. حال اگر یک نشانه‌ی جمع در کنار نشانه‌ی دیگری مانند فعل "استن" که به صورت اختصاری نوشته شده است یا ضمایر مفعولی قرار گیرد (مانند "درختانند"، "شیرینی‌خورانشان" یا "درختانمان")، تشخیص آن، دو برابر بیشتر زمان خواهد برد. این کار تاثیر بدی در کارایی موتورهای جستجو خواهد داشت و در مواردی برای آن‌ها قابل حل نیست. در زبان انگلیسی چنین مشکلاتی به مراتب کمتر هستند.

در زبان فارسی، اِعراب با وجود آنکه تلفظ می‌شوند، اکثراً نوشته نمی‌شوند. اگر لغتی با اِعراب نوشته شود از دید رایانه با حالت بی‌اعراب آن متفاوت خواهد بود. در نتیجه لازم است که همواره برای کلمات با اعراب ترفند جداگانه‌ای در نظر گرفته شود.

از دیگر سو، اکثر فارسی‌نویسان از گذاشتن برخی از حرکت‌های الزامی مانند تشدیدهای لازم و تنوین خودداری می‌کنند. نتیجه‌ی این امر، متفاوت شدن صورت ظاهری لغت از دید رایانه، با آنچه در لغت‌نامه وجود دارد خواهد بود.

به دلیل آموزش‌های متفاوتی که از طریق نظام آموزش کشور به فارسی‌نویسان داده شده است، حفظ یکپارچگی خط فارسی توسط نویسنگان مختلف مشکل است. در زبان‌های دیگر این اتفاق (تغییر دستور خط) بسیار به ندرت رخ می‌دهد در نتیجه تمام مردم در سنین مختلف از طرز صحیح نگارش لغات و دستور خط خود آگاهی دارند. البته آگاهی داشتن دلیل بر رعایت کردن اصول نیست اما وجود توافق کلی در نگارش لغات برای نگارنده‌ها امری است که منجر به تصحیح متون رسمی و یک‌دستی خط می‌شود. تغییرات خط فارسی در نظام آموزش کشور در هر دوره مانند تحولی برای کاربران قدیم این خط به حساب آمده است به نحوی که شکل این خط زمانی از جدانویسی به پیوسته‌نویسی و بعد مجدداً جدانویسی و در نهایت روشی تلفیقی تغییر شکل داده است.

تفاوت‌هایی که در زبان محاوره فارسی و زبان نوشتار فارسی وجود دارد عامل دیگری برای ناهم‌خوانی خط فارسی است. امروزه بسیاری از کاربرانِ وب‌نوشت‌های فارسی زبان محاوره را مستقیماً مکتوب می‌نمایند. یعنی به جای نوشتن "او، من و تو را آورده است"، نگارش‌های "اون منو تو رو اُورده"، "اون من وُ تو رُ اُورده" و "اون من و تو رو اورده" جایگزین می‌شوند. این تفاوت میان زبان نوشتار و گفتار در سایر زبان‌های طبیعی تاثیر کمتری دارد و حداکثر به تغییر ساختار جمله ختم می‌گردد در حالی که در فارسی هم ساختار اجزای جمله تغییر می‌کند و هم تلفظ متفاوت کلمات، املای آن‌ها را تغییر می‌دهد.

۲.۳ نویسندگان متون رایانه‌ای

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.

۲.۴ دستور خط فارسی مصوب فرهنگستان

دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی، تنها مرجع رسمی برای خط فارسی است که آخرین نسخه‌ی آن سال ۱۳۸۴ ارائه گردید. این مرجع کوچک با ملاحظه‌ی تنوع کاربران آن و به زبانی قابل درک توسط تمامی آن‌ها (و به دور از اصطلاحات تخصص خاص زبان‌شناسان) نوشته شده است و برای اکثر کاربران قابل درک و فهم و نیز کاربردی است و از این جهت گامی مؤثر در یکسان‌سازی چهره‌ی خط فارسی به حساب می‌آید.

این دستورالعمل شامل بندهایی در زمینه‌ی نشانه‌های خط فارسی، املای بعضی از واژگان خاص، پیشوندها و پسوندها، و نیز قانون نگارش ترکیبات است. در انتهای این جزوه لیستی از لغات دو املائی فارسی نیز ارائه شده است.

سر فصل‌های انتخاب شده برای دستور خط فارسی نسبتاً کامل هستند و مسائل اساسی نوشتار فارسی را شامل می‌گردند. سادگی مطالعه و درک آن و مثال‌های ذکر شده برای هر بند از قوانین نیز قابل تقدیرند.

اما این جزوه برای کاربرد در سامانه‌های رایانه‌ای طراحی نشده است. به این معنی که مخاطب این جزوه نویسنگان عادی متون فارسی می‌باشند و با وجود تلاشی که برای ساماندهی خط برای نویسندگان متون رقمی (رایانه‌ای) شده است، نتایج قابل قبولی به دست نیامده‌اند. مهم‌ترین چالش در حوزه متون رقمی، مسئله‌ی فاصله‌گذاری و پس از آن مسئله‌ی اعراب است. تا زمانی که ابهام این دو بخش به طور دقیق برای کاربران رایانه حل نشود، متون تولیدی توسط این کاربران برای مراحل بعدی پردازشی در رایانه مناسب نخواهند بود و در نتیجه تنها کاربردشان آن است که توسط کاربر دیگری خوانده شوند. حال اگر بخواهیم در این متون جستجو کنیم یا میزان پیچیدگی آن را محاسبه نماییم یا آن را به طور

خودکار خلاصه‌سازی کنیم و یا غلط‌های املائی آن را به کمک رایانه رفع کنیم، با مشکلات جدی از حوزه‌ی پردازش رایانه‌ای مواجه می‌شویم.

تا زمانی که دو کاربر ورزیده و دانش‌آموخته‌ی زبان فارسی وجود داشته باشند که املای یک کلمه را به طور متفاوتی بنویسند، سامانه‌های رایانه‌ای با مشکلات پردازشی رو به رو خواهند بود. تعدد لغاتی که این چندگونگی در آن‌ها وجود دارد، ملاک تعیین اهمیت مشکلات دستور خط فارسی در یکسان‌سازی چهره‌ی خط برای کاربران رایانه است.

در مورد وجود قواعد مشخصی که دست کاربران را برای نگارش آزادانه‌ی لغات می‌بندد، در بسیاری از محافل ادبی بحث‌های فراوان شده است همان‌طور که نگرش‌های جدانویسی کامل و پیوسته‌نویسی کامل در برهه‌های زمانی متفاوت، رد یا تائید شدند. اما نباید از یاد برد که پیشرفت کشور در حوزه‌ی انفورماتیک، در گرو اهمیت دادن به مسئله‌ی یکسان‌سازی چهره‌ی خط فارسی است. عدم تحول در این حوزه و عدم تمایل سازمان‌های رسمی و غیر رسمی به تقبّل پروژه‌هایی که خاص زبان فارسی باشند، ناشی از اطلاع آن‌ها از به‌هم‌ریختگی قواعد دستور خط فارسی بوده است. در واقع اگر خط ما تا این حد ابهام‌زا نبود، ما نیز کمی پس از دهه‌ی ۱۹۸۰ که اولین غلط‌یاب املائی زبان انگلیسی به بازار عرضه گردید می‌توانستیم این سامانه را در ایران تولید نماییم.

۲.۴.۱ مشکلات عمده‌ی دستور خط فارسی

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 اما همان‌طور که در بخش‌های قبل کاربران و مشکلات دسته‌بندی شدند، بدون تحولات بنیادین در زبان فارسی مشکلات عدم هم‌خوانی همواره وجود خواهند داشت.

هدف ما از بررسی دستور خط فارسی و بیان چالش‌ها آن است که بتوان برای فارسی دستور خطی داشت که اگر کاربران حرفه‌ای آن‌ها را به کار بَرند، میزان ابهامات تولیدی برای تحلیل رایانه‌ای متون به حداقل برسد. این مسئله برای حفظ حیات الکترونیکی زبان فارسی امری واقعاً حیاتی است.

۲.۴.۲ رویکردهای موجود و اشکالات آن‌ها

پیش از آغاز این بخش، قسمت‌هایی از سطور آغازین مستندِ دستور خط فارسی مرور می‌شوند:

> "در باب دستور خطّ فارسی، همواره اختلاف سلیقه و مشرب وجود داشته‌است؛ بعضی طرف‌دار باز گذاشتن دست نویسنده در انتخاب شیوهٔ نگارش بوده و حداکثر جواز و رخصت را تجویز می‌کرده‌اند و بعضی دیگر، برعکس، گرایش به وضع قوانینی عام و قطعی و تخلّف‌ناپذیر داشته و آرزو می‌کرده‌اند که در عالم خط و کتابت نیز قوانینی شبیه قوانین حاکم بر علائم ریاضیات حاکم باشد. از جهتی دیگر، برخی از اهل فن معایب و مشکلات موجود را در خطّ فارسی تا آن اندازه فراوان و جدّی دانسته‌اند که رفع آن‌ها را جز با افزودن و درکار آوردن حروف و علائم جدید میسّر نمی‌شمرده‌اند، و گروهی دیگر کمترین تحوّل و تبدّلی را در خطّ فعلی نپذیرفته و آن را به زیان زبان می‌دانسته‌اند".

در مورد استراتژی‌های اصلی فرهنگستان زبان و ادب فارسی در باب تغییرات اساسی در زبان و یا افزودن نشانه‌های جدید به آن، اساتید بزرگ فارسی، صاحب نظران تامّ‌الاختیار می‌باشند و در این باب گزارش حاضر نیز سخنی به میان نخواهد آورد. اما در زمینه‌ی فاصله‌گذاری میان کلمات که یکی از بزرگ‌ترین چالش‌های کنونی متون فارسی است، می‌توان دلایل گروه‌های موافق با پیوسته‌نویسی و جدانویسی را به شکل زیر دسته‌بندی نمود:

- پیوسته‌نویسی کامل:
  - تایپ فاصله برای کاربرانی که می‌خواهند لغتی مانند "میشود" را به صورت "می شود" بنویسند باعث خواهد شد که "می" در مواردی در انتهای خط اول باقی بماند و "شود" به ابتدای سطر بعد منتقل گردد. این کار از خوانایی نوشته می‌کاهد.
  - اگر برای جبران انتقال "می" به سطر بعد از نیم‌فاصله استفاده شود مستلزم آشنایی کاربران با این کلید و موقعیت آن در صفحه کلید است. زدن این کلید تنها برای کاربرانی ساده است که با گذشت زمان به آن عادت کرده‌اند.
  - اگر کلمات پیوسته نوشته شوند و لغت‌نامه‌ها نیز بر اساس اصول پیوسته‌نویسی تدوین شده باشند، جستجوی لغات در لغت‌نامه از دید رایانه نیز ساده است.

- جدانویسی کامل:
  - چشم انسان‌ها لغاتِ پُرحرف را به درستی نمی‌خواند. مثلاً اگر بخواهیم لغت "عافیت‌طلب" را به شکل "عافیتطلب" بنویسیم، خواندن آن برای هر خواننده‌ای مشکل است و نیاز به مکثِ اضافه دارد. در نتیجه جدانویسی منجر به سادگی خواندن لغت‌ها می‌گردد.
  - اگر تمام کلمات جدا نوشته شوند ابهامی در معنای کلام باقی نمی‌ماند. هر بخش از کلمه به صورت مجزا نوشته می‌شود. در حالی که در پیوسته‌نویسی، مشخص نیست که تا کجا باید واژها را به هم متصل نمود و اجزای کلمه دقیقاً کدام هستند.
  - جدانویسی در صورتی که همراه با استفاده از نیم‌فاصله باشد، فرایند تفکیک لغات را ساده می‌کند. زبان‌های طبیعی که اجزای متصل کمتری دارند در پردازش رایانه‌ای ساده‌تر هستند. بر عکس هر چه بخش‌های بیشتری از لغت به هم متصل شوند، تشخیص اجزاء برای رایانه پیچیده‌تر خواهد شد. در نتیجه تصحیح املاء، تشخیص نقش دستوری و ... برای آن مشکل‌تر است.
- راه حل بینابینی:
  - نه کاملاً پیوسته و نه کاملاً جدا. به این ترتیب ظاهر کلمات حفظ می‌شوند و حداقلی از اصول اولیه نیز در این میان رعایت می‌گردد.
  - دستورالعمل‌های مجزا برای جدانویسی و پیوسته‌نویسی ارائه می‌شود و سایر موارد مطابق با سلیقه‌ی نویسنده خواهد بود.

۲.۴.۳ نکاتی که نباید فراموش نمود

با وجود رویکردهای مختلف، نکاتی وجود دارند که در جمع‌بندی نهایی نباید فراموش نمود. این نکات عبارتند از:

- اگر حروف لغت زیاد شوند، پیوسته‌نویسی لغت آن را از شکل قابل خواندن خارج می‌نماید. خواندن لغاتی که بیشتر از ۷ الی ۸ حرف در بخش اصلی خود دارند، برای اکثراً خوانندگان سخت و نیازمند به مکث است.
- چشم خوانندگان به الگوی کلمات بیش از حروف آن‌ها عادت دارد. در واقع بسیاری از کلمات قبل از آنکه در ذهن انسان حرف به حرف پردازش شوند، به کمک شکلشان در ذهن تشخیص داده می‌شوند. مثلاً لغات "خدافاحظ" با آنکه غلط نوشته شده در لحظه‌ی اول کلمه‌ی "خداحافظ" را تداعی می‌نماید. در حالی که اگر حرف به حرف پردازش می‌شد، این فرایند

بیشتر طول می کشید. با توجه به این واقعیت، هیچ دستورالعملی در میان مردم فراگیر نخواهد شد اگر بخواهد شکل بسیاری از کلمات را تغییر دهد.

- مشکل لغات واحد مرکبی که با فاصله از هم جدا می‌شوند (مانند "خوش‌خیالی")، زمانی که به انتهای خط می‌رسند و در نتیجه‌ی آن یک بخش از لغت در انتهای سطر اول و بخش دیگر در ابتدای سطر دوم قرار می‌گیرد، به کمک نیم‌فاصله حل می‌شود. اما باید برای استفاده از نیم‌فاصله قواعد دقیق و ثابتی تعیین نمود.
- درست است که استعمال نیم‌فاصله برای کاربران مبتدی سخت است اما اگر بپذیریم که یا باید خط فارسی را از اساس تغییر دهیم یا باید از نیم‌فاصله بیشتر استفاده کنیم، احتمالاً ارزش تمرین را خواهد داشت. اگر استفاده از این نویسه در زمان تدریس نرم‌افزارهای ویرایشگر متن، به کاربران آموزش داده شود، مثلاً در سرفصل‌های دروس پرطرفدار مهارت‌های ICDL، یک فصل با عنوان اصول اولیه نگارش گنجانده شود، برخی از این مشکلات برطرف خواهند شد.
- جدانویسی کلمات اگر بی‌رویه استفاده شود، منجر به اِشکال در خواندن کلمات خواهد شد. مثلاً اگر "بدبختانه" به صورت "بدبخت‌انه" نوشته شود، خواندن آن دشوار و به چشم بیننده ناآشنا خواهد شد. باید دقت داشت که برخی از اجزای کلمه مانند بسیاری از پسوندها، مدت‌هاست که در واژگان زبان رخنه کرده‌اند و جایگاه محکمی یافته‌اند. واژگان تولیدی از این راه، با وجود آنکه در ذات خود مرکب هستند اما بدون پسوند خود کاربردی متفاوت خواهند یافت و از دید خوانندگان هم پیچیده به نظر خواهند رسید.
- با وجود آنکه نباید شکل واژگان زبان را با یک قاعده زیر و رو نمود اما بسیاری از واژگان برای چشم‌های متفاوت، متفاوت به نظر می‌رسند. به عنوان مثال بسیاری از کاربران به لغت "میشود" عادت کرده‌اند در حالی که گروه عمده‌ی دیگری چنین نیستند. برای یکسان‌سازی چهره‌ی خط فارسی، در نهایت تغییر برخی از روش‌های نگارش امری ناگذیر خواهد بود.
- بزرگترین مشکلِ راه حل‌های بینابینی، ابهام آن‌ها در استفاده از قوانین است. زمانی که انتخاب شکل نگارش کلمه رسماً به سلیقه‌ی نویسنده سپرده شود، مشکلات پردازشی زبان آغاز خواهند شد.
- وجود ارتباط متقابل میان زبان‌دانان و فن‌آوران حوزه‌ی رایانه یکی از نیازهای اصلی جامعه‌ی اطلاعاتی امروز و رشد صنعت پردازش الکترونیک در میهن عزیزمان می‌باشد. اگر در خط فارسی تغییری برای همگام شدن با این کاروان روی ندهد، میزان عقب ماندگی ما در فن‌آوری اطلاعات از سایر کشورها غیر قابل جبران خواهد شد. یعنی زمانی فراخواهد رسید که نرم‌افزارهای خارجی، خلاصه‌ی متون خارجی را در کسری از ثانیه استخراج می‌نمایند و روزنامه‌ها و مقالات با سرعت زیادی تولید می‌شوند در حالی که ما در ایران حتی کار غلط‌یابی

املاء را نیز با خطای بالا و به کُندی انجام می‌دهیم. زمانی که موتورهای جستجوی خارجی می‌توانند در زمان جستجوی یک لغت، مترادف‌ها، ریشه‌ها و سایر مشتقات آن را نیز هم‌زمان جستجو کنند تا کاربران به سادگی به اطلاعات مورد نیاز خود دست پیدا کنند، ما تازه در حال رفع مشکل کدِ نویسه‌ی "ی" و یا مشکل اجزای کلمات خود هستیم و یک جستجوی ساده را نیز نمی‌توانیم در وب‌گاه‌های رسمی کشورِ خویش با موفقیت به انجام برسانیم.

۲.۵ نمونه‌هایی از برنامه‌های نیازمند به یکسان‌سازی خط فارسی

نیازی که امروز در قالب این گزارش دیده می‌شود نیازی است که قابل اقماض نیست و رشد صنعت پردازش متون در ایران به آن وابسته است. برای روشن شدن عمق تاثیر خط فارسی در سامانه‌های رایانه‌ای که متن را تحلیل می‌کنند، نمونه‌هایی از این برنامه‌ها را مثال می‌زنیم:

- غلط‌یاب فارسی.
- غلط‌یاب ویرایشی و دستوری.
- موتور جستجوی فارسی.
- بازشناسی حروف فارسی (OCR): نرم‌افزاری که متون Scan شده را به متون رقمی تبدیل می‌کند.
- خلاصه‌ساز فارسی.
- نرم‌افزاری که کلمات کلیدی متن را استخراج کند.
- هر عملیات جستجو در متن فارسی.
- هر عملیاتی که نیازمند به محاسبه میزان شباهت دو متن به یکدیگر باشد.
- هر عملیاتی که پیچیدگی متن را محاسبه نماید.
- نرم‌افزار های فیلتر کردن متون خاص.

۳ فصل دوم- واژگان خاص

۳.۱ مقدمه

در این بخش به مرور بخش‌هایی از دستور خط فارسی می‌پردازیم که تحت عنوان "املای بعضی از واژه‌ها، پیشوندها و پسوندها" آورده شده‌اند. لازم به یادآوری است که این گزارش فقط مواردی را ذکر خواهد نمود که عملیات پردازش رایانه را با اختلال مواجه می‌نمایند.

۳.۲ کلمات خاص

جدول شماره ۱ نشان‌دهنده مواردی است که رایانه را دچار ابهام می‌نمایند. در این جدول ابتدا واژه‌ی مورد بحث (مانند "بی") ذکر گردیده و به دنبال آن قاعده‌ی ابهام‌زا آورده شده. در نهایت نیز مشکل رایانه با این قاعده عنوان شده است. در ستون سمت راست در موارد معدودی، نکته یا پیشنهاد حل مشکل نیز ارائه شده است.

جدول ۱- مواردی از قواعد دستور خط فارسی که رایانه را دچار ابهام می‌نمایند.

لغت	قاعده‌ی جاری	مشکل
ابن	حذف یا حفظ همزه اگر در میان دو اسم	اشکال: اسامی علم برای رایانه شناخته نشده نیستند. راه حل: اگر "بن" یا "ابن" با نیم‌فاصله از دو لغت کنار خود

	علم قرار گیرد. مثال: حسین‌بن‌راضی، حسین‌ابن‌علی	جدا شوند مشکلی رخ نخواهد داد. با وجود آنکه "حسین‌بن‌علی" با "حسین‌ابن‌علی" متفاوت است، اگر با نیم فاصله کنار هم نوشته شده باشد رایانه می‌تواند "بن" را در مواردی با "ابن" و بالعکس جایگزین نماید. نکته: بهتر خواهد بود اگر دستور صریحی برای استفاده از فاصله یا نیم‌فاصله برای استفاده از "بن" و "ابن" وجود داشته باشد.
به	هرگاه صفت بسازد پیوسته نوشته می‌شود.	اشکال: اگر صفتِ حاصل در لغت‌نامه وجود نداشته باشد، رایانه قادر به تائید صحت نگارش لغت نیست. راه حل: صفت‌هایی که با "ب" ساخته می‌شوند باید حتماً در لغت‌نامه وجود داشته باشند. تمام کاربران نیز باید از صفت بودن کلمه‌ی حاصل از این راه اطمینان داشته باشند. نکته: در صورتی که "به" باید جدا از کلمه‌ی بعد از خود نوشته شود، نیازمند دستور صریحی برای استفاده از فاصله یا نیم‌فاصله برای آن هستیم.
به	هرگاه پیش از کلمات عربی ظاهر گردد، پیوسته نوشته می‌شود.	اشکال: لغت‌نامه‌ی جامعی برای لغات عربی به کار رفته در متون فارسی وجود ندارد در نتیجه رایانه راه مشخصی در برخورد با این لغات نخواهد داشت. از دید رایانه این لغات می‌توانند صورت‌هایی از لغات فارسی با غلط‌های املائی تلقی شوند.
بی	در صورتی که کلمه بسیط‌گونه باشد، بی به شکل پیوسته نوشته می‌شود.	اشکال: رایانه از بسیط‌گونه بودن لغات بی‌خبر است مگر آنکه این لغات در لغت‌نامه وجود داشته باشند. اشکال: اگر شُبه‌ای در مورد اینکه کاربران زبان، یک لغت ساخته شده با "بی" را بسیط‌گونه ندانند وجود دارد (مانند بی‌راه)، در آن صورت ممکن است بسیاری از آنان لغت‌های واحد را به دو صورت متفاوت بنویسند. نکته: اگر باید "بی" را جدا از کلمه‌ی بعد از خود نوشت، نیازمند دستور صریحی برای استفاده از فاصله یا نیم‌فاصله برای آن هستیم.
هم	هم اگر کلمه	اشکال: رایانه از بسیط‌گونه بودن لغات بی‌خبر است مگر

	بسیط‌گونه بسازد، به شکل پیوسته نوشته می‌شود.	آنکه این لغات در لغت‌نامه وجود داشته باشند. نکته: در صورتی که "هم" باید جدا از کلمه‌ی بعد از خود نوشته شود، نیازمند دستور صریحی برای استفاده از فاصله یا نیم‌فاصله برای آن هستیم.
هم	اگر جزء دوم لغت، تک‌هجایی باشد، هم به صورت پیوسته نوشته می‌شود.	اشکال: رایانه از تک‌هجایی یا چندهجایی بودن واژگان خبر ندارد. در هیچ مرجع فارسی نیز لغت‌نامه‌ی حرکت‌دار (با علائم صوتی) تدوین نشده است. بنابراین رایانه روشی برای تشخیص تعداد هجاهای کلمات ندارد. تنها راه حل باقی مانده، آن است که صورت پیوسته‌ی این واژگان در لغت‌نامه وجود داشته باشد. اشکال: اگر کاربری لغتی را که با "هم" شورع می‌شد به اشتباه یا به خاطر از یاد بردن قانون تک‌هجایی، به صورت جدا نوشت، تشخیص اینکه این لغت همان لغت اول است برای رایانه ناممکن خواهد بود.
هم	اگر جزء دوم با مصوت "آ" شروع گردد پیوسته نوشته می‌شود. اگر همزه قبل از "آ" ظاهر شود، جدا نوشته می‌شود.	اشکال: برای کاربران اجرای این دو بند می‌تواند ساده نباشد و منجر به جدانویسی برخی از کلمات بند اول گردد. در نتیجه نگارش دو کاربر مختلف ممکن است برای این لغات یکسان نباشد. هر مسئله‌ای که منجر به بروز ابهام در نوشتن لغات گردد رایانه را از مسیر تشخیص درست کلمه منحرف می‌نماید.
هم	هم بر سر کلماتی که با "م" یا "الف" آغاز می‌شوند جدا نوشته می‌شود.	اشکال: این بند نیز بر ابهام جدانویسی و پیوسته‌نویسی "هم" می‌افزاید. نکته: لغت "هم" از جمله مواردی است که نیاز جدی به تدوین قوانین مشخص‌تر و ساده‌تری برای نگارش دارد. تدوین این قوانین برای اهداف یادشده در فصل اول بسیار ضروری می‌باشند.
می/ نمی	جدانویسی "می" و "نمی" از کلمه‌ی بعد از خود.	اشکال: اگر ذکر نشود که این جدانویسی با نیم‌فاصله است یا فاصله، "می" در موارد زیادی می‌تواند "مِی" تلقی شود و "نمی" به صورت "نَمی" (اندکی رطوبت). در نتیجه تفسیر جمله کاملاً متفاوت خواهد شد.

ها	قانون: موارد جدانویسی "ها".	اشکال: همان‌طور که موارد جدانویسی "ها" در دستور خط فارسی گویای آن هستند، تقریباً تمام این موارد ابهاماتی بسیار پیچیده را برای رایانه به دنبال خواهند داشت. حتی خود کاربران نیز به دلیل زیاد بودن این بندها، ترجیح می‌دهند "ها" را جدا بنویسند. به همین دلیل پیشنهاد می‌شود که جدانویسی "ها" به عنوان یکی از قدم‌های بسیار موثر در کاهش ابهام کلمات، به صورت ساده و صریح اعلام گردد. تشخیص جمع بودن یک لغت از روی مشاهده‌ی نشانه‌ی جمع "ها" برای رایانه، امکان پردازش‌های پیشرفته‌تری مانند تطابق دادن فعل با نهاد را در آینده فراهم می‌آورد. همچنین تصحیح املای لغات را بسیار سریع و ساده می‌کند. نکته: استفاده از نیم‌فاصله در نگارش "ها" باید به صورت صریح عنوان گردد.
"ام، ای، است" و ضمائر مفعولی	در مواردی که باید جدا نوشته شوند.	اشکال: فاصله‌گذاری میان اجزای کلمه‌ای که به آن "ام" اضافه شده است حتماً باید مشخص شود که با فاصله است یا نیم‌فاصله. زیرا در صورت استفاده از فاصله ایجاد ابهام می‌نماید. برای فاصله‌گذاری فعل کمکی ماضی نقلی (ام، ای، است و ...) نیز باید قاعده‌ای تنظیم گردد.
کسره اضافه	نشانه‌ی کسره‌ی اضافه در کلمات مختوم به "ه" غیر ملفوظ به شکل "ۀ" نوشته می‌شود.	اشکال: استفاده از "ۀ" یا "ی" هر دو به یک اندازه متداول است به ویژه که در زمان آموزش این نشانه در نظام آموزشی کشور، هر دو در دوره‌های متوالی به دانش‌آموزان تعلیم گردیده است و بی‌نظمی فراوانی در این مورد وجود دارد.
همزه	قوانین همزه	به خاطرسپاری قوانین همزه برای اکثر کاربران مسئله‌ای پیچیده است اگرچه این امر دلیل بر عدم لزوم به این قوانین نیست. تنها نکته‌ی موجود در مورد این قوانین آن است که چه خوب می‌شد اگر روزی تعداد همزه‌های استفاده شده در زبان فارسی، به دو یا سه نوع همزه محدود می‌شد.

		اشکال: اگر کاربری کلاً به جای "أ"، "إ"، "ؤ"، "ء" و "ئ" از "ا"، "و" و "ی" استفاده کند (مثلاً به دلیل سختیِ به خاطرسپاری مکان حروف همزه در صفحه کلید)، آنگاه رایانه قادر به تصحیح خطا و حدس زدن آنکه این لغت دارای همزه بوده است نمی‌باشد. رفع این مشکل از رفع حالتی که کاربر حالت اشتباهی از همزه را به کار برده است پیچیده‌تر می‌باشد.
ترکیبات عربی		نکته: در نگارش ترکیبات چند کلمه‌ای عربی، باید استفاده از فاصله یا نیم‌فاصله مشخص شود.
تنوین	نگارش تنوین	اشکال: همان‌طور که عنوان گردید، نوشتن تنوین الزامی است اما بسیاری از کاربران زبان به دلیل آنکه نگارش آن نیازمند به فشردن یک کلید اضافه است که معمولاً جای آن را نمی‌دانند، تنها به تایپ کردن حرف "ا" قناعت می‌کنند. این مسئله باعث می‌شود که رایانه نتواند در موارد زیادی، حضور تنوین را تشخیص دهد و در نتیجه متن حاصل از تایپِ دو کاربر، مثل هم تعبیر نمی‌شود.
تشدید	اگر عدم حضور تشدید منجر به بروز ابهام گردد، نوشتن آن الزامی است.	اشکال: بسیاری از کاربران در زمان تایپ، توجهی به ابهامِ به وجود آمده در معنای کلمه نخواهند نمود. از دیگر سو، وادار نمودن آن‌ها به تایپ نویسه‌ی تشدید مناسب نیست زیرا اکثراً جای آن را به خاطر نخواهند سپرد. البته خوشبختانه تشدید خیلی کم در متون ظاهر می‌شود.

جدول فوق شامل قوانینی بود که می‌توانستند برای رایانه ابهام ایجاد نمایند. لازم به ذکر است که در تمام مواردی که جدانویسی مطرح می‌شود، باید حتماً عنوان گردد که این عمل با استفاده از فاصله است یا نیم‌فاصله.

تمام موارد فوق مربوط به تک‌کلمه بودند و منطقی خواهد بود اگر فاصله‌گذاری میان اجزای آن‌ها با نیم‌فاصله انجام گردد. صراحت دستور استفاده از فاصله یا نیم‌فاصله به این دلیل اهمیت دارد که اگر میان دو کلمه فاصله ظاهر شود، آن دو کلمه می‌توانند در معنای مستقل خود تفسیر شوند و معنایی که از کنار هم قرار گرفتن آن‌ها حاصل می‌شود در این صورت از میان برود. مثلاً اگر "می‌خورد" به صورت "می خورد" نوشته شود، معنای جمله‌ی زیر کاملاً متفاوت خواهد شد:

- "او می خورد" به این معنی که او قبلاً مِی (شراب) خورده است.
- "او می‌خورد." به این معنی که او چیزی را می‌خورد یا هم اکنون در حال خوردن چیزی است.

۴ فصل سوم- ترکیب‌ها

۴.۱ مقدمه

آنچه تاکنون در مورد ابهامات قوانین نگارش کلمات عنوان گردید، در مورد لغات خاص بود در حالی که در این فصل قوانین مربوط به ترکیبات مرور خواهند شد.

در این گزارش هیچ پیشنهاد قاطعی برای جدانویسی کامل یا پیوسته‌نویسی کامل داده نمی‌شود و تشخیص این مورد بر عهده‌ی اساتید حاضر در فرهنگستان زبان و ادب فارسی خواهد بود. آنچه اینجا به شدت مورد تأکید است، تدوین ***قوانین بدون ابهام*** در زمینه‌ی ترکیبات می‌باشد. قوانینی که چه دستور بر پیوسته‌نویسی دهند و چه جدانویسی، در حوزه‌ی خود ابهام‌زا نباشند.

نویسندگان گزارش اذعان دارند که در مواردی مشخص نمودن دستور دقیق فاصله‌گذاری بین لغت به سادگی انجام نخواهد شد. اما حتی در همان موارد نیز اگر بتوان حکمی قطعی صادر کرد که پردازش متن را ساده‌تر نماید کمک موثری به سامانه‌های رایانه‌ای شده است که اثرات آن خیلی زود در صحنه‌ی فن‌آوری اطلاعات در کشور ظاهر خواهد شد.

۴.۲ پیوسته‌نویسی

جدول شماره ۲ نشان‌دهنده مواردی است که در مورد پیوسته‌نویسی، رایانه را دچار ابهام می‌نمایند.

جدول ۲- مواردی از قواعد دستور خط فارسی که رایانه را در تشخیص ترکیبات دچار ابهام می‌نمایند.

قانون	مشکل
مرکّب‌هایی که بسیط‌گونه هستند مانند: آبرو، الفبا، آبشار، نیشکر، رختخواب، یکشنبه، پنجشنبه، سیصد، هفتصد، یکتا، بیستگانی	اشکال: رایانه هیچ درکی از بسیط‌گونه بودن یک لغت ندارد. بنابراین اگر این لغات به همین شکل در لغت‌نامه وجود نداشته باشند، رایانه نیز قادر به تشخیص خطا نخواهد بود. راه حل: تمام لغات بسیط‌گونه‌ی مرکب، باید حتماً در لغت‌نامه وجود داشته باشند.
اگر جزء دوم لغت با "ا" شروع شود و تک‌هجایی باشد.	اشکال: رایانه از تعداد هجاهای لغات بی‌اطلاع است. در نتیجه به هیچ وجه نمی‌تواند صحت لغت حاصل را تائید نماید مگر آنکه عین لغت به صورت پیوسته در لغت‌نامه وجود داشته باشد. اشکال: اگر کاربران به اشتباه لغت را جدا نوشته باشند، اگر فاصله‌ی کامل میان اجزای آن قرار داده باشند، رایانه هیچ راهی برای اصلاح آن نخواهد داشت. اگر نیم‌فاصله گذاشته باشند، رایانه قادر به تطبیق دادن این لغت با لغت درون لغت‌نامه نخواهد بود گرچه می‌تواند صحت آن را با نگارش جاری، تائید نماید.
اگر جزء دوم با "ا" شروع شود و چند هجایی باشد، دست نویسنده باز است.	اشکال: در این صورت، رایانه تشخیص نخواهد داد که "دل‌آویز" همان "دلاویز" است که کاربری به صورت دوم نوشته است. در نتیجه دچار ابهام جدی می‌شود. این قاعده نیازمند به بررسی مجدد و تصمیم‌گیری قطعی در مورد فاصله‌گذاری آن می‌باشد.
هرگاه در اجزای یک کلمه‌ی مرکب، کاهش واجی روی داده باشد.	این نمونه نیز مانند لغات بسیط‌گونه، نیازمند به وجود لغت در لغت‌نامه است. با این تفاوت که به دلیل استعمال لغت جدید، کابران کمتری آن را به اشتباه جدا می‌نویسند یا حتی "مرکب" می‌دانند.

مرکبی که دست‌کم یک جزء آن کاربرد ندارد.	خوشبختانه در این مورد نیز مانند مورد قبل اکثر کابران از پیوسته‌نویسی استفاده می‌کنند. اما توجه به این نکته ضروری است که چنین لغاتی حتماً باید در لغت‌نامه وجود داشته باشند.
مرکب‌هایی که جدانوشتن آن ابهام به وجود آورد.	همان‌طور که عنوان شد، اگر تمام کاربران به این مسئله توجه کنند مشکلی پیش نخواهد آمد اما اگر نویسنده‌ای اشتباهاً در این مورد جدانویسی را انتخاب کند، رایانه قادر به تشخیص حالت پیوسته نیست.
کلمات مرکبی که جزء دوم آن‌ها تک‌هجایی باشد و جنبه‌ی سازمانی داشته باشند.	رایانه از مسئله‌ی مفهوم لغات بی‌خبر است. در نتیجه نمی‌تواند تشخیص دهد که یک لغت جنبه‌ی سازمانی دارد یا خیر. اگر این لغت به همین شکل در لغت‌نامه باشد، آن را تائید خواهد نمود. اما اگر کاربران زبان، روی‌کرد واحدی را در نگارش این لغات پیش نگیرند (که اکثراً نمی‌گیرند)، رایانه نمی‌تواند یکسان بودن دو لغت از این مجموعه را که یکی پیوسته و یکی جدا نوشته شده است تشخیص دهد.

۴.۳ جدانویسی

جدول شماره ۳ نشان‌دهنده مواردی است که در مورد جدانویسی، رایانه را دچار ابهام می‌نمایند.

جدول ۳- مواردی از قواعد دستور خط فارسی که رایانه را در تشخیص ترکیبات دچار ابهام می‌نمایند.

قانون	مشکل
لغات پیشوندی همواره جدا نوشته می‌شود.	موارد استثناءای که برای "هم" و "به" و "بی" وجود دارد حتماً باید رفع ابهام شده باشند. نکته: ابزار فاصله‌گذاری این لغات (فاصله یا نیم‌فاصله) باید حتماً مشخص شده باشد.
ترکیب‌های اضافی	اشکال: در این مورد که ترکیبات اضافی توسط اکثر کاربران جدا نوشته می‌شوند مشکلی وجود ندارد. مشکل زمانی روی می‌دهد که باید از فاصله و نیم‌فاصله میان

	اجزای ترکیب استفاده نمود. از دید رایانه، "آب میوه" با "آب‌میوه" متفاوت است.
جزء دوم با الف آغاز شود.	همان‌طور که در یکی از بندهای پیوسته‌نویسی قانونی مشابه وجود داشت و دست کاربر در نگارش پیوسته و جدای ترکیباتی که جزء دوم آن‌ها با الف آغاز می‌شود، آزاد گذاشته شده بود، این مورد نیز بسیار ابهام‌زا است. کاربران باید به طور قاطع بدانند که لغتی مانند "دلاویز" یا "دل‌انگیز" باید جدا نوشته شوند یا پیوسته. پیشنهاد: پیشنهاد این است که در مورد این لغات، حالت جدا در نظر گرفته شود و بر استفاده از نیم فاصله نیز تأکید گردد. البته نظر اساتید فرهنگستان در این زمینه حرف نهایی خواهد بود.
حرف پایانی جزء اول با حرف آغازین جزء دوم هم‌مخرج یا مشابه باشد.	اشکال: رایانه از هم‌مخرج بودن حروف بی‌اطلاع است. اگر کاربران بخواهند این کلمات را به هر روشی که مایلند بنویسند، رایانه نه قادر به تصحیح و نه قادر به بازشناسی لغات مشابه است. نکته: اگر فرض بر آن باشد که اکثر ترکیبات جدا نوشته می‌شوند، این مشکل خود به خود مرتفع می‌گردد.
مرکب‌های اتباعی	اشکال: بسیاری از مرکب‌های اتباعی در لغت‌نامه وجود ندارند و توسط نویسندگان زبان نوشته می‌شوند. در این صورت این قاعده باید به گونه‌ای جامع و بی‌ابهام باشد که هر کاربری که تصمیم به نگارش یک نمونه مرکب اتباعی گرفت، آن را مشابه با سایر کاربران بنویسد. نکته: استفاده از فاصله و نیم‌فاصله باید در این قانون مشخص شده باشد.
مرکب‌هایی که یک جزء آن کلمات دخیل هستند.	اشکال: رایانه لغات دخیل را نمی‌شناسد. در نتیجه دخیل بودن یا نبودن لغات باید برای تمام کاربران زبان محرز باشد تا همواره نگارش آن‌ها از چنین کلماتی یکسان باشد. اشکال: تمام مرکب‌های اتباعی به ترکیبات دو کلمه‌ای

	محدود نمی‌شوند. لغاتی مانند "شِرّ و وِر"، "آش و لاش" و ... نیز مرکب اتباعی هستند که درون خود از میانوند استفاده می‌نمایند. در این موارد وجود دستوری قاطع برای استفاده از نیم‌فاصله یا فاصله، امری حیاتی است زیرا به کار بردن فاصله، این مرکب‌های اتباعی را دچار ابهام در معنا می‌کند و منجر به آن می‌شود که جزء دوم (اتباع) در جمله مانند یک غلط املائی به نظر برسد. نکته: اگر فرض شود که پیش‌فرض نگارش ترکیبات، جدانویسی است، این مورد نیز خود به خود مرتفع خواهد شد و تنها مسئله‌ی ابزار فاصله‌گذاری باقی خواهد ماند.
عبارات عربی چند جزئی	استفاده از فاصله و نیم‌فاصله در این مورد باید مشخص شده باشد. اشکال: کابران فارسی زبان اکثراً املای لغات عربی را به یک شکل نمی‌نویسند. به ویژه که عبارات عربی می‌توانند در عین آنکه یک مفهوم را می‌رسانند با الفبای متفاوتی (ناشی از شرایط صرفی متفاوت) نوشته شوند.
نگارش عدد یک (پیوسته و جدا)	اشکال: اگر "یک" توسط کاربران مختلف هم پیوسته و هم جدا نوشته شود، از دید رایانه دو لغت را تولید می‌نماید. ابهام میان این موارد باید بر طرف گردد زیرا در غیر این صورت تمام پردازش‌های بعدی روی این لغات دچار اختلال خواهند شد.
کلمه با پیوسته‌نویسی نامأنوس شود.	اشکال: از دید خیلی از کاربران، لغات "حقیقتجو" و ... نامأنوس نیستند در نتیجه بسیاری از آن‌ها این لغات را به یک حالت نمی‌نویسند. نکته: اگر پیش‌فرض نگارش کلمات، جدانویسی باشد این مورد به سادگی مرتفع می‌گردد.
هر یک از اجزای کلمه‌ی مرکب، چند حرف	اشکال: از دید رایانه، "پای‌برهنه" با "پابرهنه" متفاوت

مختوم داشته باشد.	است و کاربرانی که یکی از این دو صورت را استفاده می‌کنند باید بپذیرند که ممکن است نتیجه‌ی پردازش متن آن‌ها تا حدی نادرست باشد.
یک جزء آن اسم خاص باشد	اشکال: همان‌طور که گفته شد، رایانه از خاص بودن اسامی اطلاع ندارد مگر آنکه در لغت‌نامه برایش تعریف کرده باشند. در غیر این صورت اگر کاربران توافقی برای نگارش کلمات مرکب با اسم خاص نداشته باشند، رایانه نیز قادر به تصحیح یا کشف یکسان بودن دو لغت نیست.
جزء آغازی یا پایانی آن بسامد زیاد داشته باشد.	اشکال: رایانه از بسامد کلمات اطلاع چندانی ندارد. در نتیجه نمی‌تواند تشخیص دهد که "نیکبخت" همان "نیک‌بخت" بوده است. نکته: اگر پیش‌فرض نگارش کلمات، جدانویسی باشد این مورد به سادگی مرتفع می‌گردد.
هرگاه با پیوسته‌نویسی اجزای ترکیب معلوم نشوند	اشکال: رایانه از اینکه در وضعیت خاصی ممکن است اجزای ترکیب معلوم نشوند بی‌خبر است. در نتیجه نمی‌تواند تشخیص دهد که "پاکنام" حالت غلط "پاک‌نام" است یا لغتی است که غلط املائی دارد.

۴.۴ ترکیبات اضافی

ترکیبات اضافی مواردی هستند که کاربران به دلیل آنکه کسره‌ی میانی ترکیب را تلفظ می‌کنند، دو لغت را از هم جدا می‌نویسند. اما مشکل جایی ظاهر می‌شود که باید میان عناصر ترکیب فاصله‌ای درج نمود. اگر از "فاصله" استفاده شود (که در بسیاری از موارد چنین است)، باید قوانینی در مورد استفاده از نیم‌فاصله در مواردی مانند "سیب‌زمینی" وجود داشته باشد. این که به چه دلیل "سیب‌زمینی" به شکل "سیب زمینی" نوشته نمی‌شود باید برای تمام کاربران زبان کاملاً و بدون هیچ تردیدی مشخص شده باشد.

در بررسی‌های ساده‌ی اولیه چنین به نظر می‌آید که زمانی که کسره‌ی اضافه در ترکیبات اضافی تلفظ نمی‌شود (یعنی دو لغت به صورت متداول در کنار یکدیگر به کار رفته‌اند)، نیم‌فاصله به کار

می‌رود. اگر این حدس صحیح است باید در دستور خط گنجانده شود تا کاربران این‌گونه لغات را به طرق مختلف ننویسند و چهره‌ی خط همه جا یکسان باشد.

۴.۴.۱ ترکیبات اضافی چند جزئی

یکی از مسائل پیچیده در فاصله‌گذاری ترکیبات، مربوط به حالت‌هایی است که یک ترکیب چند جزئی، باید نوشته شود. این گونه ترکیبات اگر مانند یک لغت در نظر گرفته نشوند، معنایی متفاوت به وجود می‌آورند. مثلاً اگر "آب‌سردکن" به این صورت نوشته نشده باشد، خواننده باید چند لحظه درنگ کند تا تشخیص دهد که منظور از "آب سرد کن"، یک جمله‌ی امری نیست و یک دستگاه به نام آب‌سردکن است. ذهن انسان قادر به تشخیص این مسئله است زیرا به معنای جمله و نیز قواعد ساخت جمله آگاهی دارد. در نتیجه می‌تواند بگوید "آب‌سردکن" را می‌توان با "ها" جمع بست. اما رایانه به هیچ روی نمی‌تواند تشخیص دهد که "آب سرد کن‌ها" یک عبارت درست است زیرا "ها" به فعل اضافه نمی‌شود. این در شرایطی است که بداند "کن" فعل است که خود نیازمند پردازش پیچیده‌ی دیگری است.

اگر کاربری در موتور جستجوی گوگل، لغت "آب‌سردکن" را جستجو کند، قادر به پیدا کردن صفحاتی که در آن‌ها لغت "آب سرد کن" نوشته شده است نمی‌باشد و بالعکس. این مسئله در مورد تعداد زیادی از لغات مرکب فارسی رخ می‌دهد.

برای یکسان‌سازی نگارش چنین لغاتی سه رویکرد عمده وجود دارد. در رویکرد اول در تمام موارد نیم‌فاصله استفاده می‌شود تا مرز لغات از یکدیگر مشخص باشد. اما در این حال، عباراتی مانند "پررفت‌وآمد" باید به چه فرمی نوشته شوند؟ آیا استفاده از نیم‌فاصله در این سطح جایز است؟

در رویکرد دوم در تمام موارد از فاصله استفاده می‌شود. اما با این تدبیر، رایانه هیچ روشی برای تعیین درستی "آب سرد کن ها" نخواهد داشت. این روش گرچه ساده است اما پردازش‌های بعدی رایانه‌ای را به شدت کاهش می‌دهد.

در رویکرد سوم در مواردی که میان کلماتِ یک ترکیب صدا (کسره اضافه) وجود ندارد، آن‌ها را با نیم‌فاصله و در غیر این صورت با فاصله می‌نویسیم.

رویکرد سوم کاملاً نیازمند قوانین دسته‌بندی شده برای بررسی دقیق‌تر انواع ترکیبات است و احتمالاً منجر به تدوین یک جدول برای بیان شرایط فاصله‌گذاری ترکیبات چند کلمه‌ای (بیش از دو کلمه) می‌گردد.

۴.۴.۲ مثال‌ها و پیشنهادها

در جدول زیر تعدادی از قوانین تولید کلمه که منجر به تولید ترکیبات چند جزئی می‌شوند ذکر گردیده‌اند. در ستون اول از این جدول قانون تولید کلمه عنوان شده است، در ستون دوم مثال ذکر گردیده و در ستون سوم نکته یا پیشنهادی مربوط به این قانون آورده شده است.

جدول ۴- برخی از قوانین تولید ترکیبات چند جزئی

قانون	مثال	پیشنهاد
اسم/ضمیر + بن مضارع ⟵ صفت فاعلی و نیز: اسم/ضمیر + بن مضارع + ی ⟵ صفت	دل‌گشا، زودرنج، خویشتن‌دار، خاطره‌نویس	در تمام ترکیباتی که در آن‌ها بن‌مضارع به کار رفته است، جزء قبل از بن‌مضارع باید با نیم‌فاصله در کنار بن‌مضارع قرار گیرد. در هیچ حالتی نیز پیوسته نوشته نشوند.
گروه وصفی ⟵ صفت	آب‌سردکن	اگر بدون کسره اضافه به هم متصل هستند باید با نیم‌فاصله نوشته شوند.
صفت و موصوف مبهم + بن مضارع ⟵ صفت فاعلی	همه‌چیزدان، هیچ‌چیزندان	اگر بدون کسره اضافه به هم متصل هستند باید با نیم‌فاصله نوشته شوند. در هیچ حالتی پیوسته نوشته نشوند.
صفت مفعولی + "شده" ⟵ صفت مفعولی	گرفته شده، خوانده شده	چون در ساخت افعال مجهول از این قاعده بسیار استفاده می‌شود و رسم نیست که اجزای گروه فعلی را با نیم‌فاصله کنار هم قرار دهند، بهتر است همه‌جا این مورد با فاصله‌ی تمام باشد.
عدد + و + عدد ⟵ صفت شمارشی	بیست و یک	چون رسم نیست ترکیبات عددی با نیم‌فاصله نوشته شوند، بهتر است که این مورد همه‌جا با فاصله‌ی تمام نوشته شود.

صفت + ترکیب عطفی ← صفت	پررفت‌وآمد، با آب و رنگ، با شرم‌وحیا	کاملاً مبهم.
حرف اضافه + متمم + صفت مفعولی ← صفت	به‌هم‌خورده، به‌هم‌ریخته	کاملاً مبهم.
حرف اضافه + ضمیر + صفت ← صفت	از ما بهتران، از خود راضی، از خود بی‌خود	کاملاً مبهم.
ترکیب عطفی (اسم) + صفت ← صفت	دست‌ودل‌باز، دست و پا چلفتی	کاملاً مبهم.
ترکیب عطفی (اسم) + بن مضارع ← صفت	دست‌وپاگیر	کاملاً مبهم.
مرکب اتباعی که در آن‌ها تغییر در واج دوم و استفاده از میانوند وجود دارد.	مارچ‌ومورچ	اگر اجزای مرکب اتباعی با فاصله نوشته شوند، به دلیل بی‌معنی بودن جزء دوم، توسط رایانه به عنوان یک غلط املائی تلقی خواهند شد. در نتیجه به هیچ وجه قابل بازیابی و تبدیل شدن به شکل اصلی آن نیستند. به همین دلیل بهتر خواهد بود اگر مرکب‌های اتباعی همواره با نیم‌فاصله از هم جدا شوند.
مرکب‌های اتباعی که بعد از حرف دوم یک "و" اضافه می‌شود.	هارت‌وهورت	مشابه مورد قبل (مرکب‌های اتباعی)
صفت منفی + و + صفت منفی ← صفت	بی‌بو و بی‌خاصیت	کاملاً مبهم.
فعل امر + و + فعل نهی ← صفت	کجداروم‌ریز، بخور و نمیر	کاملاً مبهم.
اسم + و + بن فعلِ هم‌معنی ← اسم مصدر	مرگ‌ومیر	کاملاً مبهم.
بن ماضی + و + بن مضارع ← اسم مصدر	گفت‌وگو، رفت‌وروب، جست‌وجو	کاملاً مبهم.