تنوع نگارشی در زبان فارسی و تهیه خودکار دادگان املایی از پیکره زبانی مبتنی بر وب

مسعود قیومی* ساغر شریفی† مرضیه صناعتی‡

*دانشگاه آزاد برلین، برلین، آلمان
masood.ghayoomi@fu-berlin.de
†دانشکده زبان‌های خارجی، دانشگاه آزاد اسلامی واحد کرج، البرز، ایران
saghar.sharifi@kiau.ac.ir
‡پژوهشکده زبان‌شناسی، سازمان میراث فرهنگی، صنایع دستی، و گردشگری، تهران، ایران
msanaati20@gmail.com

چکیده

در عصر اطلاعات و ارتباطات، وب جایگاه ویژه‌ای پیدا کرده‌است، چراکه با کاربران بسیار متنوعی در تعامل بوده و می‌توان از آن به‌عنوان یک منبع اطلاعاتی غنی زبانی استفاده کرد. تهیه پیکره مبتنی بر وب می‌تواند برای پردازش‌های زبانی منبع مناسبی باشد. ولی استفاده از این منبع ساده نیست. ازآنجا که کاربران مختلفی موجب خلق متن در وب می‌شوند، برخورد با پدیده تنوع نگارشی اجتناب‌ناپذیر خواهد بود. در این مقاله، به بررسی این پدیده در پیکره زبانی حاصل از وب برای زبان فارسی می‌پردازیم و با معرفی یک الگوریتم تلاش می‌کنیم تنوع نگارشی واژه‌ها را به‌طور خودکار استخراج کرده و براساس آن، دادگانی برای تنوع نگارشی واژه‌های فارسی تهیه کنیم. سپس به طبقه‌بندی تنوع نگارشی می‌پردازیم. این دادگان می‌تواند برای اتخاذ شیوه‌های آموزش زبان فارسی به غیرفارسی‌زبانان، و یا در زبان‌شناسی پیکره‌ای و پردازش زبان طبیعی مورد استفاده قرار گیرد.

کلمات کلیدی

پردازش خودکار زبان فارسی، زبان‌شناسی پیکره‌ای، پیکره مبتنی بر وب، فاصله لونشتاین، تنوع نگارشی، دادگان، طبقه‌بندی

۱ مقدمه

وب در ابتدای پیدایش خود به‌عنوان یکی از منابع مهم اطلاعات مورد استفاده قرار می‌گرفت. وبگاه‌های مختلف خبری یا بایگانی مقالات می‌توانست اطلاعات مفیدی را به کاربران ارائه دهد. در سال‌های اخیر این منبع اطلاعات نقش دیگری را نیز به‌عهده گرفته‌است و آن برقراری ارتباط میان کاربران می‌باشد. وبلاگ‌ها و گپ‌سراها نمونه‌هایی از نسل دوم وب می‌باشد که آنها این امکان را فراهم کرده‌اند تا انسان‌ها بتوانند به‌عنوان کاربر به تعامل با یکدیگر بپردازند و به این وسیله تفکرات، نظرات، و احساسات خود را درمورد مسایل گوناگون با استفاده از ابزار زبانی نوشتاری ابراز کنند. مسلماً داده حاصل از نسل جدید وب که روزانه به‌طور چشمگیری درحال گسترش است می‌تواند به‌عنوان یک منبع غنی‌تر اطلاعات مورد استفاده قرار گیرد. اما باتوجه به متنوع‌بودن کاربران به‌لحاظ تحصیلی، فرهنگی، اجتماعی، گویشی، جنسیتی، و سنی، تنوع سبکی در نوشتار آنها اجتناب‌ناپذیر است. این تنوع ممکن است با سبک نوشتار رسمی و معیار در کتاب، روزنامه، یا مجله متفاوت باشد.

یکی از کاربردهای عمده وب در پردازش زبان طبیعی می‌باشد، چراکه در پردازش زبان طبیعی یکی از منابع مهم مورد نیاز پیکره زبانی[1] است. در پژوهش‌های پیشین انجام‌شده برروی زبان‌های گوناگون، بیشتر از پیکره‌های استاندارد مانند وال‌استریت‌ژورنال[2] برای زبان انگلیسی یا پیکره همشهری[3] برای زبان فارسی استفاده شده‌است. براساس این داده‌های استاندارد، ابزار، روش، و الگوریتم‌های مورد نیاز پردازش زبان تهیه و معرفی شده‌است. پردازش داده وب با استفاده از ابزارهای تهیه‌شده یک چالش بسیار بزرگ و مهم محسوب می‌شود، زیرا داده وب بسیار پرنوفه است و پیش‌پردازش زیادی نیاز دارد تا داده به‌شکل استاندارد درآید و بتوان از ابزار تهیه‌شده برای پردازش آن استفاده کرد. زبان فارسی از این قاعده مستثنی نیست، و حتی با مشکلاتی به‌مراتب بیش از زبان‌های دیگر روبه‌روست. یکی از مشکلات اصلی آن، مربوط به ویژگی‌های خط زبان فارسی است. ازجمله چالش‌های تهیه پیکره و پردازش زبان فارسی، نحوه نگارش چندواژگی واژه‌های مرکب، تفاوت سبک نوشتاری رسمی و محاوره‌ای، همزه و پایه آن، و همچنین واژه‌های خارجی است.

آنچه در این مقاله به آن می‌پردازیم ارائه یک الگوریتم است که بتواند به‌طور خودکار تنوع نگارشی واژه‌ها را از پیکره زبانی استخراج کند و پس از طبقه‌بندی تنوع‌نگارشی به‌صورت یک دادگان سازمان‌دهی شود.

ساختار مقاله حاضر چنین است: در بخش ۲ مقاله، به پیشینه پژوهش‌های انجام‌شده در حوزه معیارسازی خط زبان فارسی می‌پردازیم. در بخش ۳، الگوریتم استفاده‌شده را معرفی می‌کنیم. بخش ۴ ویژگی‌های پیکره‌های زبانی مورد استفاده را بیان می‌کند. بخش ۵، به یافتن واژه‌های مشابه و طبقه‌بندیِ داده می‌پردازد. در بخش ۶، طبقه‌بندی تنوع نگارشی بررسی و تحلیل می‌شود. در بخش ۷، به نتیجه‌گیری پرداخته خواهد شد.

۲ پیشینه پژوهش‌های انجام‌شده در حوزه معیارسازی خط زبان فارسی

فرهنگستان زبان و ادب فارسی [۲] به‌عنوان متولی معیارسازی خط زبان فارسی دستور خط زبان فارسی را تهیه کرده است. اگرچه تلاش شده که با ارائه قواعد، معیارسازی خط زبان فارسی صورت پذیرد، برای هر قاعده فهرستی از استثناها ارائه شده‌است. این امر مبیّن این نکته است که معیارسازی خط زبان فارسی کاری ساده‌ای نیست. قیومی و ممتازی [۳]، قیومی و همکاران [۴] و شمس‌فرد [۹] به بررسی چالش‌های تهیه پیکره و پردازش زبان فارسی پرداخته‌اند.

شمس‌فرد و همکاران [۱۰] مجموعه ابزاری را تهیه کرده‌اند که با استفاده از آن می‌توان متن را با توجه به دستور خط فرهنگستان زبان و ادب

1 Corpus
2 https://catalog.ldc.upenn.edu/LDC2000T43
3 http://ece.ut.ac.ir/dbrg/hamshahri/

Archive of SID
www.SID.ir

فارسی معیارسازی کرد. این ابزار متن‌باز[4] نبوده و به‌صورت رایگان در دسترس نیست. این ابزار سه کار تصحیح و تقطیع واژگانی، تحلیل واژگانی، و برچسب‌گذاری مقوله دستوری را انجام می‌دهد. همچنین سرابی و همکاران [۷] و سراجی و همکاران [۸] مجموعه ابزاری را تهیه کرده‌اند که پس از پیش‌پردازش داده ورودی، کار برچسب‌گذاری مقولات دستوری و همچنین تجزیه جملات در چارچوب دستور وابستگی انجام می‌شود. هزم[5] نیز مجموعه ابزار متن‌باز به زبان پایتون است که مانند دو مطالعه بالا عمل می‌کند. هدف قسمت پیش‌پردازش معیارسازی نسبی متن ورودی است. کارهای انجام‌شده مبین اهمیت جایگاه معیارسازی داده زبان فارسی و پردازش خودکار آن است.
در مقاله حاضر به‌نوعی به‌دنبال معیارسازی متن فارسی هستیم، با این هدف که علاوه‌بر برآوردن نیاز افرادی که در حوزه پردازش زبان فارسی فعالیت می‌کنند، بتوان از آن برای زبان‌شناسی پیکره‌ای و همچنین اتخاذ روش‌هایی در حوزه آموزش زبان فارسی به غیرفارسی‌زبانان و رفع مشکلات در نوشتار این زبان استفاده کرد.

۳ معرفی الگوریتم یافتن واژه‌های مشابه

برای یافتن تنوع نگارشی واژه‌ها در این پژوهش، از الگوریتم فاصله لونشتاین[6] [۵] استفاده کرده و مدل جدیدی را برای گسترش‌دادن این الگوریتم معرفی می‌نماییم. در الگوریتم اصلی فاصله لونشتاین، مدلی از برنامه‌نویسی پویا[7] استفاده شده است که سه حالت جایگزینی، انطباق، یا درج حروف بین واژه‌های یک پیکره را در نظر می‌گیرد و با احتساب هر یک از سه حالت عددی که بیانگر فاصله دو واژه است ارائه می‌گردد. شبه‌کد این الگوریتم در زیر آمده است:

ورودی: دو واژه (W_1, W_2)
محاسبه فاصله لونشتاین:
محاسبه طول واژه‌های W_2 و W_1: N و M
ساخت ماتریس دوبعدی $(N+1 \times M+1)$
پرکردن ردیف صفر ماتریس با نمایه ستون مرتبط با آن
پرکردن ستون صفر ماتریس با نمایه ردیف مرتبط با آن
شروع حلقه
شروع از ردیف یک و ستون یک تا ردیف N و ستون M به صورت ردیف به ردیف
محاسبه امتیاز سلول $[i,j]$ براساس حالات ممکن
درج در واژه اول:

$$[i,j]=[i-1,j]+1$$

درج در واژه دوم:

$$[i,j]=[i,j-1]+1$$

انطباق در دو واژه:

$$[i,j]=[i-1,j-1]$$

جایگزینی یک حرف در دو واژه:

$$[i,j]=[i-1,j-1]+1$$

انتخاب یکی از چهار حالت با کمترین امتیاز
پایان حلقه
خروجی: یک جفت واژه به‌همراه امتیاز محاسبه شده در سلول $[N,M]$ فاصله لونشتاین

در نسخه گسترش‌یافته پیشنهادی ما، این الگوریتم به‌صورت وزن‌دهی‌شده مورد استفاده قرار می‌گیرد. به این معنا که در حالت درج یا جایگزینی همیشه عدد ۱ به مقدار سلول‌های مورد نظر اضافه نمی‌شود. بلکه بسته به اینکه چه حرفی درج‌شده و یا چه حروفی جایگزین یکدیگر شده‌است، وزن عددیِ متفاوتی به سلول مورد نظر اضافه می‌شود. در این مدل، دو عدد ۰/۱ و ۱ را درنظر می‌گیریم؛ به این ترتیب که برای حروف خاصی که در فارسی به‌دلیل تنوع نگارشی درج می‌شود عدد ۰/۱ اضافه شده و برای سایر حروف مانند نسخه اصلی الگوریتم عدد ۱ اضافه می‌شود. همین‌طور برای جایگزینی حروفی که معمولاً در فارسی به دلیل تنوع نگارشی جایگزین هم می‌شود عدد ۰/۱ و برای سایر حروف عدد ۱ اضافه می‌شود. این وزن‌دهی شامل واژه‌های ساده و مرکب می‌باشد.
این وزن‌دهی سبب می‌شود راحت‌تر بتوان واژه‌های دارای تنوع نگارشی را جستجو کرد، چراکه فاصله چنین واژه‌هایی کمتر از ۱ می‌باشد. نمونه‌هایی که برای آنها وزن کمتری لحاظ می‌گردد از قرار زیر است:

- **جایگزینی**

۱: جایگزینی حرف «آ» و «ا»: در پیکره مورد استفاده، واژه‌هایی یافته می‌شود که در آنها به‌جای «آ» از «ا» استفاده شده است، مانند: آرام و ارام؛ الآن و الان.
۲: جایگزینی همزه با معادل کرسی آن سبب تنوع نگارشی می‌شود:
۲-۱: جایگزینی «ا» و «أ»، مانند: راس و رأس.
۲-۲: جایگزینی «أ» و «ئ»، مانند: مسأله و مسئله.
۲-۳: جایگزینی «ا» و «إ»، مانند: انشاءالله و إنشاءالله.
۲-۴: جایگزینی «و» و «ؤ»، مانند: مسوول و مسؤول.
۲-۵: جایگزینی «و» و «ئ»، مانند: مسوول و مسئول.
۲-۶: جایگزینی «ی» و «ئ»، مانند: رییس و رئیس.
۳: جایگزینی حروف در گونه رسمی و گفتاری
۳-۱: جایگزینی شناسه «د» و «ه» در آخر فعل‌هایی مانند: می‌خورد و می‌خوره.
۳-۲: جایگزینی «ا» و «و» در واژه قبل از حروف «م» یا «ن»، مانند: نان و نون.

- **درج**

۱: درج «ء» در واژه، مانند: انشاءالله و انشاالله، سوءاستفاده و سواستفاده
۲: درج «ئ» در واژه، مانند: آئینه و آینه
۳: درج فاصله مجازی، مانند می‌شود و میشود و می‌¬شود

٤ پیکره‌های زبانی استفاده‌شده

در پژوهش حاضر، از پیکره تاک‌بانک[8] استفاده شده‌است. این پیکره از وب تهیه شده و شامل وبلاگ و متون نوشتاری و روزنامه‌ای در وب با حجم بیش از ۵۰۰ میلیون واژه است. قاعدتاً پردازش این حجم داده که از منابع گوناگون جمع‌آوری شده کار بسیار دشوار و پیچیده‌ای است، چراکه با متن معیار فاصله زیادی دارد. با استخراج بسامد اولیه از پیکره، ویژگی چندنگارشی واژه‌ها خود را به‌راحتی نشان می‌دهد، زیرا سبب می‌شود که در فهرست واژگان استخراج‌شده چندین مدخل با بسامد متفاوت در فهرست یافت. برای مطالعه جامع‌تر تنوع نگارشی واژه‌ها، پیکره تاک‌بانک را با پیکره بی‌جن‌خان [۱] و همچنین واژه‌های پیکره وابستگی [۶] ادغام کردیم. لازم به ذکر است این دو پیکره از منابع نوشتاری و روزنامه تهیه شده و استاندارد شده‌اند.

۵ یافتن واژه‌های مشابه و طبقه‌بندی داده

در بخش ۳، مدل پیشنهادی ما برای محاسبه فاصله لونشتاین توصیف شد که از آن برای یافتن واژه‌های مشابه استفاده می‌شود. واژگان پیکره ادغام‌شده که در بخش ۴ ذکر گردید استخراج شده و به‌عنوان داده ورودی به الگوریتم داده می‌شود تا فاصله لونشتاین تمامی واژه‌ها محاسبه شود. در مرحله بعد، جفت واژه‌هایی که فاصله لونشتاین آنها کمتر از ۱

4 Open Source
5 https://github.com/sobhe/hazm/
6 Levenshtein Distance
7 Dynamic Programming
8 http://www.sketchengine.co.uk/documentation/wiki/Corpora/TalkBankPersian

Archive of SID

www.SID.ir

است از خروجی الگوریتم انتخاب می‌گردد. در مرحله آخر، واژه‌های مشابه کنار هم مرتبط می‌شود تا تحلیل داده و طبقه‌بندی آن ساده‌تر گردد.
از آنجا که هدف این مقاله ارائه یک طبقه‌بندی از تنوع نگارشی واژه‌ها به‌همراه بسامد آن برای نمایش مقبولیت نگارش واژه توسط گویشوران فارسی است، داده به‌دست‌آمده به‌صورت دستی ارزیابی کیفی شده است. در گام نخست تحلیل داده، ۵۰۰ واژه که دارای تنوع نگارشی است توسط دو زبان‌شناس تحلیل گردید. در مواردی که اتفاق نظر وجود نداشت با تجزیه و تحلیل موردی، تصمیم‌گیری شده است. نتایج اولیه بررسی این واژه‌ها از سه جنبه بررسی شده است که در قسمت بعدی توضیح داده می‌شود.

٦ تحلیل طبقه‌بندی تنوع نگارشی

با بررسی ۵۰۰ واژه از فهرست واژه‌های مشابه، مواردی که سبب پدیدآمدن تنوع نگارشی در واژه‌های مشابه شده‌است به شرح زیر دسته‌بندی می‌شود.

٦-١ تنوع ناشی از ویژگی‌های خط فارسی

٦-١-١ عدم رعایت مرز درون‌واژه‌ای

در واژه‌های مشتق و غیربسیط (اعم از تصریفی یا اشتقاقی) که دارای پیشوند و پسوندهایی مانند «می»، «ی»، «ای»، «ها» است، استفاده نادرست از فاصله مجازی و یا عدم استفاده از آن، مثال‌هایی مانند «حرفها» و «حرف_s_ها»، «نمیدهد» و «می_s_ گو_s_ ید» را به دست داده است.[9] این مسئله در واژه‌های مرکب و مشتق مرکب مانند «راه_s_اندازی» نیز صادق است. البته جدانویسی می‌تواند به این دلیل باشد که شخص، نمونه مورد نظر را دو واژه فرض کرده است، مانند «ازنو» که به‌صورت « از_s_نو» هم آمده یا «برضد» که به‌صورت «بر_s_ضد» نیز نوشته شده است.

٦-١-٢ همزه

در مواردی، نوشتن همزه بر روی کرسی‌های گوناگون باعث پدیدآمدن صورت‌های متنوع شده است، مانند «سؤالهای» و «سئوالهای». در برخی از واژه‌ها، نوشتن همزه به جای«ی» سبب تفاوت در تلفظ می‌شود، مانند: «امریکایی» و «امریکائی» و یا «سوئی» و «سویی». همچنین، نوشتن یا ننوشتن همزه عامل دیگری در گوناگونیِ واژه‌هاست، مانند: «رأی»، «رای»، و «راءی».

٦-١-٣ نگارش «ا» به جای «آ»

حضور یا عدم حضور «آ» یا مَدّ، در بعضی واژه‌ها سبب تفاوت در تلفظ واژه می‌شود، مانند «امریکا» و «آمریکا» یا «الآن» و «الان». درحالی‌که در «آنها» و «انها» یا «آبان» و «ابان» تفاوت در تلفظ نداریم. این امر ممکن است به چند دلیل باشد: سلیقه و سبک مؤلفان، شتابزدگیِ سازندگان پیکره یا بی‌اطلاعی آنها.

٦-١-٤ حذف

در مواردی، حذف یک حرف از واژه را داریم، مانند حذف الف از فعل «است» در «آنست» (=آن است) و «آن_s_ست» و «آنند» (= آن اند/هستند). حذف در این حالت تغییر معنایی در واژه ایجاد نمی‌کند و در یک دوره‌زمانی، بیشتر مشاهده می‌شود که نویسنده‌ها تابع سرهم‌نویسی بوده‌اند.

٦-٢ تنوع ناشی از تغییرات زبان‌شناسی

گاهی شاهد تنوعات نگارشی هستیم که از تغییرات آوایی حاصل می‌شود. این تغییرات آوایی اغلب به‌صورت قاعده‌مند رخ می‌دهد. در زیر به برخی از مواردی اشاره می‌شود که نتیجه این تنوعات و تغییرات است.

٦-٢-١ سبک متفاوت

گاهی هنگامی که تغییرات آوایی در انواع مقولات دستوری ایجاد می‌شود واژه‌هایی با سبک متفاوت اعم از رسمی، محاوره‌ای عامیانه یا ادبی پدید می‌آید، مانند: «اگه» (اگر)، «داره» (دارد) و «همون» (همان)، «نون» (نان)، «بام» (باهام)، و «برا» (برای).
در فعل‌ها، تغییرات آوایی و در پی آن، تغییر سبکی ممکن است در چند بخش ایجاد شود، مثلاً در ستاک فعل مانند «می_s_گم» (می‌گویم)، یا در شناسه آن مانند: «نمیده / نمی_s_ده» (نمی‌دهد)، و یا در هر دو، مانند:« نمی_s_ره» (نمی‌رود) و «می_s_شه» (می‌شود). همان‌گونه که از نمونه‌ها برمی‌آید این تغییر سبکی عمدتاً از رسمی و معیار به محاوره‌ای یا عامیانه میل می‌کند.
از سوی دیگر، برخی از صورت‌های محاوره‌ای، عامیانه و ادبی درنتیجه تغییر آوایی به‌دست نیامده‌است. واژه‌های «ازش» (از+ش) و «اِوا» (در صورتی که نشانه تعجب باشد) و «اوی» (هنگامی که برای صدا کردن به کار می‌رود) از آن جمله است. به عبارتی، این واژه‌ها معادلِ تغییریافته واژه دیگری نیست.

٦-٢-٢ چندمعنایی و هم‌نویسگی

در مواردی، تغییرات آواییِ باعث ایجاد چندمعنایی یا هم‌آوا هم‌نویسگی در واژه‌ها می‌گردد، مثلاً «بوم» می‌تواند صورت محاوره‌ایِ «بام» یا به معنای «بوم» نقاشی و یا به معنای جغد باشد.
نمونه دیگر «نمیره» است که هم به معنای «نَمیرد» و هم «نِمی‌رود» است. در تمام این موارد ممکن است همزمان با تغییر معنا، شاهد تفاوت در سبک واژه‌های متفاوت باشیم. همچنین این پدیده معنایی در تمام مقولات دستوری رخ می‌دهد و محدود به مقوله خاصی نیست.

٦-٢-٣ ساخت صرفی نحوی

همچنین، در برخی موارد، تغییرات آوایی سبب ایجاد تغییر در ساخت صرفی یا نحوی واژه‌ها می‌گردد. «منو» که به‌صورت «من_s_و» هم آمده در واقع از دو تکواژِ «من» و «را» تشکیل شده است که در آن، «را» دستخوش تغییر آوایی و به واژه‌بستِ «و» (o) تبدیل شده است. «بِم» یا «به_s_ام» نیز که صورتِ تغییریافته «به+م» (=به من) است نیز چنین وضعیتی دارد.

٦-٣ سایر موارد

در پیکره نمونه‌هایی از قبیل «ا_s_ن_s_ها»، «ر_s_ها» و «میکه» (که احتمالاً صورت درست آن «میگه» است) یافت شده‌است که به ماهیت خط فارسی (مبنی بر چسبیده بودن یا جدانویسی) مربوط می‌شود. علت وجود چنین مواردی می‌تواند شتابزدگی، بی‌دقتی، و البته بی‌اطلاعی شخص پدیدآورنده متن باشد. این اشکالات، سبب به‌وجودآمدن موارد بی‌معنیِ بسیار زیاد و متنوعی در پیکره شده‌است.

٧ نتیجه‌گیری

در این مقاله، به بررسی تنوع نگارشی در پیکره مبتنی بر وب پرداخته شد. برای استخراج تنوع نگارشی، تلاش کردیم با وزن‌دهی به حروف خاصی در الگوریتم لونشتاین، به‌طور خودکار واژه‌هایی که تنوع نگارشی داشت را مشخص کنیم. سپس با ارزیابی کیفی اولیه داده، طبقه‌بندی‌ای را باتوجه به دلایل ایجاد تنوع نگارشی تهیه کردیم. این طبقه‌بندی، سه طبقه کلی

9 برای نمایش فاصله مجازی از علامت _s_ استفاده شده‌است.

Archive of SID

www.SID.ir

برای تنوع نگارشی را دارد که می‌تواند ناشی از ویژگی‌های رسم‌الخط زبان فارسی، ویژگی‌های زبان‌شناختی، و یا سایر موارد باشد. نتیجه این پژوهش می‌تواند در آموزش زبان به غیرفارسی‌زبانان، زبان‌شناسی پیکره‌ای، و یا پردازش خودکار زبان فارسی مورد استفاده قرار گیرد.
در ادامه این کار پژوهشی، علاوه‌بر بررسی تعداد بیشتری از واژه‌های مشابه، و سازمان‌دهی آن به‌صورت یک دادگان، به انطباق دستور خط فرهنگستان با واقعیت زبانی و بررسی آماری تنوع نگارشی بین گویشوران خواهیم پرداخت.

منابع

[1] بی‌جن‌خان، محمود، **نقش پیکره زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای**، مجله زبانشناسی، ۱۹ (۲):۴۸ ،۶۷ ،۱۳۸۳.

[2] فرهنگستان زبان و ادب فارسی ، **دستور خط فارسی،** فرهنگستان زبان و ادب فارسی، تهران، ۱۳۸۹.

[3] Ghayoomi, Masood and Saeedeh Momtazi, "**Challenges in developing Persian corpora from on-line resources**", In Proceedings of 2009 IEEE International Conference on Asian Language Processing, Singapore, pp: 108-113, 2009.

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

[5] Levenshtein, Vladimir I., "**Binary codes capable of correcting deletions, insertions, and reversals**", Soviet Physics Doklady, 10 (8): 707–710, 1966.

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

[7] Sarabi, Zahra and Hooman Mahyar and Mojgan Farhoodi, "**ParsiPardaz: Persian language processing toolkit**", In Proceedings of IEEE 3rd International eConference on Computer and Knowledge Engineering, Mashad Ferdowsi University, October 31-November 1, pp. 73-79, 2013.

[8] Seraji, Mojgan and Beáta Megyesi and Joakim Nivre, "**A basic language resource kit for Persian**", In Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey, May 19-21, pp: 2245-2252, 2012.

[9] Shamsfard, Mehrnoush, "**Challenges and open problems in Persian text processing**", In Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November 25-27, Poznań, Poland, pp: 65-69, 2011.

[10] Shamsfard, Mehrnoush and Hoda Sadat Jafari and Mahdi Ilbeygi, "**STeP-1: A Set of fundamental tools for Persian Text Processing**", In Proceedings of the 7th International Conference on Language Resources and Evaluation. Valletta, Malta, May 19-21, pp: 859-865, 2010.

Archive of SID

www.SID.ir