شناسائی و پوشش واحدهای خارج از واژگان در فارسی غیررسمی

داود حیدرپور
کارشناس ارشد زبان‌شناسی رایانشی، دانشکده علوم و فنون نوین، دانشگاه تهران
d.heidarpour@ut.ac.ir

مصطفی صالحی
استادیار، دانشکده علوم و فنون نوین، دانشگاه تهران
mostafa_salehi@ut.ac.ir

محمود بی‌جن‌خان
استاد، گروه زبانشناسی، دانشکده ادبیات و علوم انسانی، دانشگاه تهران
mbjkhan@ut.ac.ir

هادی ویسی
استادیار، دانشکده علوم و فنون نوین، دانشگاه تهران
h.veisi@ut.ac.ir

وحید رنجبر
دکتری فناوری اطلاعات، دانشکده علوم و فنون نوین، دانشگاه تهران
vranjbar@ut.ac.ir

چکیده

کلمات سیاق فارسی غیررسمی از نظر ساختار به دو دسته رسمی و غیررسمی تقسیم می‌شود. این کلمات و ساخت آنها را می‌توان شناسائی و تحلیل رایانه‌ای کرد، اما کلمات خارج از واژگان بسیاری نیز وجود دارد که تحت تاثیر تغییرات آوایی و یا نگارش شکسته اینترنتی (بخصوص رسانه‌های اجتماعی) به شکل متنوعی نوشته می‌شود و در واژگان رسمی و غیررسمی شناسائی شده، موجود نیست. این واحدها برای ابزارهای تحلیل‌گر رایانه‌ای غیر قابل شناسائی است، در حالی که ممکن است معادل آنها در واژگان موجود باشد. شناسائی و یا نگاشت آنها به معادلشان در واژگان امکان تحلیل را به ابزار رایانه‌ای می‌دهد. ضمن جمع آوری دادگان بیشتر زیرسیاق‌های فارسی غیررسمی به روش پیکره نمونه‌گیری که حدود پنجاه هزار قطعه شده است، در این مقاله با بررسی تغییرات آوایی، خطاهای نگارشی رایج در آنها، قواعدی استخراج شده است که با بکارگیری آنها شناسائی این کلمات برای ابزارهای تحلیل‌گر بهبود یافته است. برای یک ابزار تحلیل‌گر تصریفی این افزایش پوشش منجر به افزایش فراخوانی به میزان ۱٫۲۵٪ شده است.

کلیدواژه‌ها: تحلیل تصریفی فارسی معاصر، تغییر آوایی کلمات غیررسمی، الگوریتم آوایی، فارسی غیررسمی، فارسی معاصر، کلمات خارج از واژگان.

۱. مقدمه

منظور از فارسی غیررسمی متن‌هایی است که در موقعیت‌های غیررسمی مانند گفتار محاوره‌ای و شبکه‌های اجتماعی تولید می‌شوند. تبدیل کلمه‌های رسمی به غیررسمی در زبان فارسی همیشه منجر به تولید واحد واژگانی مستقل نمی‌شود. گاهی این واژه جدید هیچ تفاوتی در تلفظ، نسبت به واژه رسمی ندارد اما در نگارش متفاوت از آن است. گاهی واژه‌های غیررسمی با نگارش‌های متفاوت نوشته می‌شود. استفاده از این تغییرات که معمولا از الگوهای آوایی، جایگزینی نویسه‌های[1] هم‌صدا و سایر متغیرهای تحت تاثیر نگارش اینترنتی (مثل رسانه‌های اجتماعی)، تبعیت می‌کند برای شناسایی و تحلیل تصریفی کلمات فارسی معاصر (رسمی و غیررسمی) لازم است. برای یک ابزار رایانه‌ای با منابع واژگانی محدود رسمی و غیررسمی فارسی، شناسائی و

[1] Character

تحلیل این واحدهای خارج از واژگان امکان‌پذیر نیست. تاکنون در پژوهش‌های فارسی به دلیل تمیز ندادن واحدهای واژگانی مستقل از نگارش‌هایی که کلمات خارج از واژگان تولید می‌کند، پوشش لازم برای شناسائی این گونه نگارش‌ها بدست نیامده است. در ادامه چالش‌های پیش رو در پردازش فارسی رسمی و به ویژه غیررسمی که به طور مستقیم با مسئله واحدهای خارج از واژگان در ارتباط است بیان می‌شود.

۱. ۱. فاصله‌گذاری

فاصله‌گذاری بین تک‌واژها در زبان فارسی معمولا به درستی رعایت نمی‌شود. در موارد بسیاری هیچ چارچوب ثابتی برای استفاده از فاصله، نیم‌فاصله و اتصال وجود ندارد. این مسئله منجر به تولید گونه‌های مختلف یک کلمه چند واحدی[۲] و یا چند قطعه‌ای[۳] می شود. این مو ضوع یکی از علت‌های تنوع و یکد ست نبودن نگارش فارسی در سطح کلمات است. برای مثال می‌توان به کلمه کتاب‌ها که به سه صورت **کتابها**، **کتاب ها** و **کتاب‌ها** نوشته می‌شود، اشاره کرد.

۱. ۲. حروف غیر قطعی

برخی حروف فارسی وجودی غیر قطعی دارند؛ این حروف گاهی با حروف دیگر جایگزین می‌شوند و گاهی می‌توان آنها را حذف کرد. این پدیده به شکلی در فارسی رسمی و به شکلی دیگر در فارسی غیررسمی رخ می‌دهد (جدول ۱).

همزه یکی از حروف الفبای فارسی به حساب می‌آید که به صورت مستقل و متصل در استاندارد یونیکد[۴] (عربی)، دو نویسه مجزا به حساب می‌آید (ء و ئ)، اما در زبان فارسی به شکل مستقل یک حرف محسوب می‌شود. همین‌طور این حرف همراه با واو، الف با فتحه، الف با کسره هم استفاده می‌شود که در استاندارد یونیکد هر کدام یک نویسه مستقل به حساب می‌آید.

حروف هم‌صدا نیز در کلمات غیررسمی گاهی در جای یکدیگر به کار می‌روند. در سیاق فارسی رسمی و غیررسمی این تغییر، خطای املائی تلقی می‌شود اما به دلیل رخداد متناوب آن می‌بایست آن را رفع کرد تا چنین کلماتی به معادلشان در واژگان نگاشت شود.

حرف **ه** در نقش واکه در صورتی که در پایان یک تک‌واژ قرار بگیرد، در بسیاری از موارد، در هنگام اتصال به تک‌واژ دیگر حذف می‌شود.

گاهی به خطا حرف **ه** برای بازنمایی کسره اضافه (در فارسی غیررسمی) در پایان کلمه منتهی به همخوان و یا واکه **ی** استفاده می‌شود. مثل: جنوبه ایران / جنوبِ ایران.

۱. ۳. تغییرات آوایی

تغییرات آوایی عمدتا منجر به تولید واحد واژگانی مستقل از کلمات رسمی می‌شود. در مواردی نیز تغییرات عمده‌ای ایجاد نمی‌کند و صرفا نگارش کلمه رسمی را تغییر می‌دهد. این دو دسته به این صورتند: الف- واحدهای واژگانی مستقل که از تغییر کلمه‌های رسمی به دست می‌آید و می‌بایست در واژگان وجود داشته باشد. مثل: آسمون؛ آسمان، نون؛ نان، هیچکدوم؛ هیچکدام، اصاب؛ اعصاب، اگه؛ اگر. ب – واحدهای غیر واژگانی

---

[2] MUT – Multi Units Token
[3] MTU – Multi Tokens Unit
[4] Unicode

با نگارش متفاوت از رسمی. مثل: عرزش؛ ارزش، عاریایی؛ آریایی، دخدر؛ دختر، انرجی؛ انرژی، امبار؛ انبار، ایسگا؛ ایستگاه، هش؛ هشت. دسته دوم به دلیل تنوعی که در تولید و چرخش از کلمه اصلی دارد معمولا در واژگان موجود نیست.

۱. ۴. کلمات به هم چسبیده

کلمات به هم چسبیده خطای املایی رایجی است که شناسایی و تحلیل کلمات را با مشکل مواجه می‌سازد و در هر دو گونه رسمی و غیررسمی رخ می‌دهد.

جدول ۱ - حروف غیر قطعی فارسی و انواع جایگزینی آنها

	نوع جایگزینی	مثال	
همزه	جایگزینی همزه با حرف اصلی	زمینهٔ (خشونت)	زمینه (خشونت)
		إملا، أعضا، مسؤول	املا، اعضا، مسوول
		املائی	املایی
	حذف همزه مستقل در پایان	انشاء	انشا
	جایگزینی همزه با همزه	مسؤول	مسئول
حروف هم‌صدا	ذ، ز، ض، ظ	گذشته،ضربان، زبان، ظرف	گزشته، ذربان، ضبان، ضرف
	ت، ط	طرف، تماس	ترف، طماس
	س، ص، ث	ساحل، صابون، ثبت	صاحل، سابون، صبت
	غ، ق	قفل، غذا	غفل، قذا
	ح، ه	حوله، هستی	هوله، حستی
	ؤ، ا، أ، إ، ئ، ع، ی	مسؤول، ارزش، هوائی	مسعول / مسیول، عرزش، هواعی
حذف واکه پایانی ه در کلمات رسمی	اسم	به شکل	بشکل
	حرف ربط گروهی	به طوری که	بطوریکه
	شماره / قید	یک به یک	یک بیک
	ضمیر شخصی	به ما	بما
	ضمیر پرسشی	به کجا	بکجا
	ضمیر مبهم	به کسی	بکسی
	جمع با ان	ستاره، آینده، آزاده	ستارگان، آیندگان، آزادگان
حذف واکه پایانی ه در کلمات غیررسمی	صفت + پی‌بست اسنادی ۳	خسته	خستس
	اسم + پی‌بست اسنادی ۳	باقیمانده	باقیماندس
	ضمیر اشاره + پی‌بست اسنادی ۳	این‌همه	این‌همس
	ضمیر مبهم + پی‌بست ضمیری ۶	همه	همشون
	اسم + پی‌بست ضمیری ۵	پنبه	پنبمون
	صفت + پی‌بست ضمیری ۳	برنده	برندش

کلماتی که منتهی به یکی از حروف **و، ر، ز، ژ، د، ذ، ا** با شد با احتمال بی شتری ممکن ا ست به کلمه بعد از خود به شـــکل متصـــل حروف چینی شـــود. علت این امر یک نوع خطای شـــناختی[5] اســـت که در هنگام حروف‌چینی سریع رخ می‌دهد. گرچه این نوع اتصال بخش زیادی از این خطا را پو شش می‌دهد اما شامل همه آن نمی‌شود.

برخی حرف‌های اضافه و ضمایر نیز تحت تاثر این تغییر، گاهی حروف واکه خود را در محل این اتصال از دست می‌دهند. برخی از این تغییرات نماینده دگرگونی‌های درزمانی زبان و گاه تعیین کننده سبک گوینده هستند. مانند ازین؛ از+این، ترا؛ تو+را و موارد مشابه دیگر.

۱. ۵. خطای املایی

خطای املایی یکی از مشـــکلات اجتناب ناپذیر در همه پردازش‌های متنی اســـت. در فارســـی رســـمی و به خصوص فارسی غیررسمی چنین مسئله‌ای می‌تواند به صورت جدی مشکل‌ساز شود. بخش عمده این خطاها بنابر دســـته‌بندی کوکیک (۱۹۹۲) که خطاها را به ســـه دســـته حروف‌چینی، شـــناختی و آوایی تقســـیم کرده ا ست، منـ شا آوایی و شناختی دارد. در این صورت اگر برای سایر مـ شکلات این بخش (چالش‌های پردازش فار سی) چاره‌جویی شود، بخش عمده این خطاها که آوایی و نگار شی ا ست برطرف می شود. اما د سته دیگر خطاها یعنی خطاهای ناشی از حروف‌چینی نیز در بعضی موارد می‌تواند پردازش فارسی را دچار اخلال کند.

۲. پیشینه

آرمین و شمس‌فرد (۱۳۸۹) با استفاده از یک ریشه‌یاب رسمی (شریفلو و شمس‌فرد، ۲۰۰۸) و افزودن برخی تک‌واژهای تصریفی غیررسمی و تغییرات آوایی به آن، یک سیستم ریشه‌یابی برای فارسی غیررسمی طراحی کرده‌اند. روش کار آن به این صورت است که تک‌واژهای لیست شده را شناسایی و حذف می‌کند و تا جایی حذف تک‌واژها را ادامه می‌دهد که به ریشه معادل در واژگان رسمی و یا واژگان غیررسمی جمع‌آوری شده برسد. برای شناسایی ریشه کلمات هم از تعدادی قاعده تبدیل محاوره به رسمی استفاده می‌کند. در نهایت با استفاده از یک مدل زبان دوتایی[6]، کلماتی که در قاعده تصریفشان ابهام وجود دارد، اولویت‌بندی می‌کند. دقت گزارش شده برای این ریشه‌یاب ۹۳ درصد بر روی ۱۰۰ کلمه داده آزمون است.

علینقی‌زاده (۱۳۹۶) نیز با انجام پژوهشی مشابه آرمین و شمس‌فرد (۱۳۸۹) و اصلاح و افزودن قواعد بیشتری به آن موفق شده است دقت آن را ۶ درصد افزایش دهد.

اسدی (۲۰۰۷) و مرادی (۲۰۱۲) با بررسی پیکره‌های گفتاری فارسی، حذف و یا تغییر صداها را بررسی کرده‌اند. این تغییرات که در بیان و در میان سیلاب‌های کلمه اتفاق می‌افتد هنگام نگارش غیررسمی نیز معمولا به حروف چینی منتقل می‌شود.

هدف این مقاله شناسائی الگوهای تغییر آوایی و مدل کردن آنها جهت شناسائی و پوشش بیشتر کلمه‌های غیررسمی است.

---

[5] Cognitive
[6] Bigram

۳. چارچوب پیشنهادی

قواعد تغییر آوایی و سایر تغییرات در حروف کلمات می‌تواند با استفاده از ساختار مبدل حالت متناهی[۷] پیاده شود. چنین ساختاری علاوه بر ایجاد امکان اعمال تغییرات آوایی و هم‌صدا، امکان پیاده‌سازی تغییرات بیانی و تا حدی املائی را نیز فراهم می‌کند. این قواعد صرفا برای شناسائی به کار می‌رود و در مرحله تولید استفاده نمی‌شود.

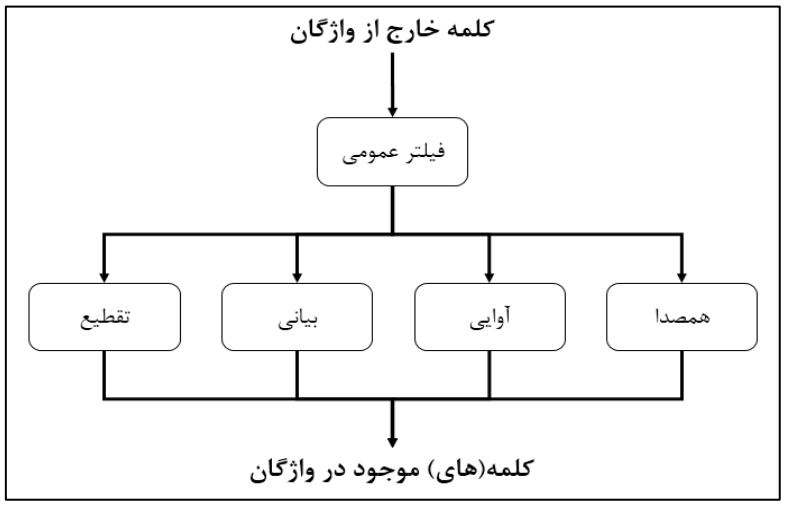
۳. ۱. ابزار پیاده سازی

ابزار فوما[۸] (هولدن، ۲۰۰۹) یک ابزار متن آزاد و رایگان است که صورت‌بندی[۹] لکس[۱۰] و زبان قاعده‌پذیر[۱۱] فناوری زیراکس[۱۲] (بیسلی و کارتونن، ۲۰۰۳) برای ایجاد تغییر در مبدل‌ها و منطق مرتبه اول[۱۳] را پیاده کرده است. همین‌طور ابزار جستجوگر واژه / قاعده را در کنار سایر ابزارهای خود دارد. علاوه بر این کتابخانه‌هایی به زبان‌های سی، جاوا و پایتون جهت استفاده از مبدل‌ها و یا ایجاد تغییر در آنها در سطح برنامه‌نویسی برای ابزاری مستقل را فراهم کرده است.

این ابزار امکان استفاده از چندین مبدل حالت متناهی[۱۴] را در کنار یکدیگر بوجود می‌آورد. بنابراین می‌توان به شکل جدا و یا در دنباله هم تعدادی مبدل استفاده کرد تا هر کدام تغییر خاصی را پوشش دهد.

۳. ۲. ساختمان کلی

از آنجایی که فار سی امروزی (ر سمی و غیرر سمی) زوایای متفاوتی دارد، برای سهولت و مدیریت بر تحلیل کلمات آن می‌باید از چند مبدل متنوع اســـتفاده کرد و تغییرات آوایی را جداگانه در قالب مبدل‌های مجزا استفاده کرد (شکل ۱).



شکل ۱ - ساختار کلی مبدل‌های استفاده شده در این پژوهش

[7] FST – Finite State Transducer
[8] Foma
[9] Formalism
[10] lexc
[11] Regex – Regular Expression
[12] Xerox
[13] First Order Logic
[14] Finite State Transducer - FST

۳. ۲. ۱. مبدل فیلتر عمومی

مرزنماهای فاصله، نیم‌فاصله و اتصال در این مبدل به صورت غیر قطعی[15] می‌توانند جایگزین یکدیگر شوند. دلیل این کار پوشش همه حالت‌های فاصله دادن قطعات کلمه است که هم در فارسی رسمی و هم غیررسمی به صورت متنوع استفاده می‌شود.

نویسه‌های **ا** و **آ** نیز به صورت غیر قطعی تعریف می‌شود. در نگارش کلمات دارای این دو حرف جایگزین شدن آنها با یکدیگر بسیار اتفاق می‌افتد. این جابجایی ممکن است در هر دو گونه رسمی و غیررسمی رخ دهد.

۳. ۲. ۲. مبدل هم‌صدا

در این مبدل حروف هم‌صدا جایگزین یکدیگر می‌شوند. در نگارش فارسی رسمی و غیررسمی به دلیل خطای شناختی و یا به عمد، این حروف می‌تواند جایگزین یکدیگر شوند (حروف هم‌صدا در جدول ۱).

قاعده اضافه[16] برای حرف ه در پایان ساخت اسامی و ساخت‌های غیر فعلی نیز در این مبدل قرار دارد. این ساخت به نوعی خطا تلقی می‌شود.

۳. ۲. ۳. مبدل آوایی

ممکن است برخی کلمات غیررسمی خارج از واژگان باشد. قواعد این قسمت می‌تواند کلمات رسمی واژگان را به معادل غیررسمی‌شان نزدیک / تبدیل کند. برخی کلمات غیررسمی نیز واحد واژگانی نیستند و تنها در نگارش ممکن است متفاوت از معادل رسمی‌شان باشند، سعی شده تا این قواعد آنها را نیز شناسائی کند. این قواعد در جدول ۲ قرار دارد.

۳. ۲. ۴. مبدل بیانی

در این مبدل کلمه‌هایی که دارای حروف تکراری هستند، تبدیل به حالت عادی می‌شود. تفاوتی نمی‌کند این حروف در کجای کلمه باشد و یا چند حرف از کلمه تکرار شود، کلمه اصلی با توجه به قاعده این مبدل به کلمه اصلی بدون تکرار حروف تبدیل می‌شود. مثال: چراااااااا؟ ← چرا، اههههههههه ← اه، نمیییییییییییخوام ← نمی‌خوام.

---

[15] Non-deterministic

[16] کسره اضافه

جدول ۲ - قواعد آوایی

قاعده	مثال
خوا ← خا	خواندن ← خاندن، خواستن ← خاستن
ان ← ون در میان کلمه	تهران ← تهرون
ام ← وم در میانه کلمه	آرام ← آروم
اه ← ا در پایان کلمه	ایستگاه ← ایسگا، کوتاه ← کوتا، باشگاه ← باشگا
آ و ا ← عا و ع در ابتدای کلمه	آریایی ← عاریایی، ارزش ← عرزش
نب ← مب	انبار ← امبار، انبه ← امب
شت ← ش در پایان کلمه	هشت ← هش، گذاشت ← گزاش، اردیبهشت ← اردیبهش
ست ← س در پایان کلمه	دست ← دس
چه ← چ در ابتدا کلمه	چهل ← چل
ند ← ن در پایان کلمه	بلند ← بلن، اسفند ← اسفن
فت ← ف در پایان کلمه	تفت ← تف، جفت ← جف
کر ← ک در پایان کلمه	فکر ← فک
در ← د در پایان کلمه	قدر ← قد
کر ← گر	لشکر ← لشگر
ش ← چ، چ ← ش	هیچ‌کس ← هیشکس
ت ← د، د ← ت	دختر ← دخدر، تشک ← دشک
ژ ← ج، ج ← ژ	انرژی ← انرجی

۳. ۲. ۵. مبدل تقطیع

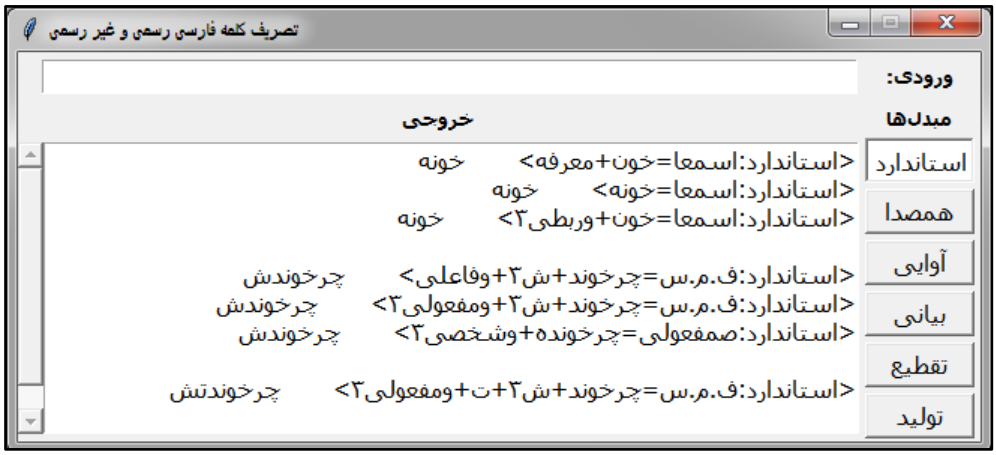
این مبدل کلمه را به دو قطعه می‌شـــکند. این تقطیع کلمه تنها یک بار اتفاق نمی‌افتد بلکه به تعداد حروف کلمه ا صلی، قطعه تولید می شود (این تعداد شامل خود کلمه ا صلی نیز می شود). هر کدام از قطعات تولید شده اگر در واژگان با شد، در خروجی ظاهر می شود. در کلمه‌های تولید شده در خروجی، آن دو کلمه‌ای که مجموع طولشان مساوی طول کلمه اصلی است و به همان ترتیب، کلمه اصلی را می‌سازند، می‌توانند به عنوان جایگزین انتخاب شوند. برای مثال کلمه **ازمردم**، قطعه‌های شکل ۲ را تولید می‌کند. از این شش گروه تنها جفت کلمه(هایی) از بین آنها می‌تواند انتخاب شود که هر دو در واژگان موجود باشد.

ازمردم	ا-زمردم	از-مردم	ازم-ردم	ازمر-دم	ازمرد-م

شکل ۲ - قطعه‌های تولید شده از قطعه ازمردم

۳. ۳. ابزار پیاده‌شده

برای پیاده‌سازی مبدل‌های تحلیل‌گر تصریفی همان‌طور که پیش‌تر ذکر شد از ابزار فوما استفاده شده است. ابزار پیاده‌شده این قسمت یک ابزار برای دسترسی آزمایشی به مبدل‌هاست که دارای یک واسط گرافیکی است و به صورت مستقل و تک به تک می‌توان کلمه را به تحلیل‌گر داد و مبدل مورد نیاز را انتخاب کرد و خروجی آن را دید. تصویر این واسط را در شکل ۳ مشاهده می‌کنید.



شکل ۳ - واسط کاربر (ابزار آزمایشگاهی) مبدل‌ها

مبدل استاندارد در این تصویر یک تحلیل‌گر تصریفی است که می‌تواند کلمات رسمی و غیررسمی فارسی را تحلیل کند. برای ارزیابی این مبدل‌ها آنها را در کنار این تحلیل‌گر قرار داده‌ایم و خروجی این مبدل‌ها به مبدل استاندارد متصل می‌شود تا تحلیل تصریفیِ کلمه خارج از واژگانِ شناسائی شده بدست آید.

۴. ارزیابی

در این ارزیابی از یک تحلیل‌گر تصریفی که قادر است کلمات رسمی و غیررسمی فارسی را تحلیل تصریفی کند استفاده شده است. کلماتی که در تحلیل‌گر شناسائی نشده است و یا به خطا شناسائی شده را به مبدل‌ها داده‌ایم و خروجی آنها را دوباره به تحلیل‌گر تصریفی انتقال دادیم و تغییر در خروجی را اندازه گرفته‌ایم.
از آنجایی که بخشی از کلمات به کار رفته در زبان فارسی معاصر، کلمات رسمی است، علاوه بر نوع خطاها نوع کلمات از نظر رسمی بودن و غیررسمی بودن نیز بررسی شده است.

۴. ۱. مجموعه داده

کلماتی که برای ارزیابی استفاده شده است از جملات زیرسیاق‌های زبان فارسی غیررسمی است؛ زیرسیاق‌های ر سانه‌های اجتماعی، وبلاگ‌ها، زیرنویس فیلم، متن پیاده شده سخنرانی، م صاحبه، اجرای رادیو / تلوزیونی، نامه شخصی، شعر محاوره و نظر کاربران فضای مجازی در پایین مطالب و محصولات وب‌سایت‌های مختلف، برای این هدف جمع‌آوری شده ا ست. دادگان جمع‌آوری شده از این زیر سیاق‌ها از نوع پیکره نمونه‌گیری[۱۷] شده ا ست و نمونه‌های هر سیاقِ زبان فارسیِ غیرر سمی به شکل تصادفی[۱۸] انتخاب شده ا ست (مکانری و

[17] Sampling corpus
[18] Random sampling

هاردی، ۲۰۱۱). سعی شده است تا هم تنوع گونه و سیاق و هم نسبت متون حفظ شود و با این حساب خطای سوءگیری[19] و خطای تصادفی[20] کنترل و به حداقل برسد (بایبر، ۱۹۹۳-الف؛ بایبر، ۱۹۹۳-ب). بنابراین داشتن ویژگی توازن[21] و ویژگی نمایندگی[22] در جمع آوری آن رعایت شده است (مک‌انری و هاردی، ۲۰۱۱). مجموع کل دادگان حدود پنجاه هزار قطعه است و ۳۰ درصد آن تنها برای ارزیابی کنار گذاشته شده است.

مجموع تعداد همه کلماتی که برای ارزیابی به شیوه بالا دوباره از میان داده‌های جمع‌آوری شده (۳۰ درصد ارزیابی) انتخاب شده است ۴,۰۴۰ کلمه است که اگر کلمات یکتای آن را محاسبه کنیم ۱,۷۸۶ کلمه می‌شود.

[19] Bias error
[20] Random error
[21] Balanced
[22] Representative

جدول ۳ - بررسی گونه‌ای خطاها

معیار		رسمی بودن	تحلیل‌گر تصریفی*	هم‌صدا	آوایی	بیانی	تقطیع
از دست رفته (FN)	کمبود واژگان	غیررسمی	۴۰	۴۰	۲۸	۲۸	۲۸
		رسمی	۸۳	۸۳	۸۳	۸۳	۸۳
	نقص در قاعده	غیررسمی	۴	۴	۴	۴	۴
		رسمی	۶	۶	۶	۶	۶
	خطای هم‌صدا و کسره اضافه	غیررسمی	۳	صفر	۰	۰	۰
		رسمی	۰	۰	۰	۰	۰
	به هم چسبیده	غیررسمی	۳	۳	۳	۳	صفر
		رسمی	۱	۱	۱	۱	۱
	تغییر آوایی	غیررسمی	۵	۵	۳	۳	۳
		رسمی	۰	۰	۰	۰	۰
	خطای املایی	غیررسمی	۸	۸	۸	۸	۸
		رسمی	۱۷	۱۷	۱۷	۱۷	۱۷
	ادبی		۳	۳	۳	۳	۳
هشدار نادرست (FP)			۱۳	۱۳	۱۹	۱۹	۳۴
موفقیت (TP)		غیررسمی	۱۵۳۰	۱۵۳۳	۱۵۴۷	۱۵۴۷	۱۵۵۰
		رسمی	۲۱۷۴	۲۱۷۴	۲۱۷۴	۲۱۷۴	۲۱۷۴
از دست رفته (FN)		غیررسمی	۶۰	۵۷	۴۳	۴۳	۴۰
		رسمی	۱۱۲	۱۱۲	۱۱۲	۱۱۲	۱۱۲

* این یک تحلیل‌گر تصریفی با واژگان رسمی و غیررسمی محدود به خود است که در این ارزیابی در کنار مبدل‌های شناسائی کلمات خارج از واژگان استفاده شده است.

۴. ۱. تحلیل نتایج

بررسی گونه‌ای مستقل از متن خطاهای کلمات (تایپ‌های) آزمون در جدول ۳ آمده است. کلمات یکتا کلماتی هستند که صرفا از نظر نگارش کاملا یکسانند و به کاربردهای متنوع احتمالی آنها از نظر صرفی-نحوی توجهی نمی‌شود. در مقابل تایپ‌های یک کلمه، به تمام کلمات یکتای یکسانی گفته می‌شود که از نظر صرفی-نحوی متفاوت از یکدیگر هستند (و هر کدام تحلیل تصریفی متفاوتی تولید می‌کند). همان‌طور که مشاهده می‌شود، در تحلیل‌گر تصریفی از مجموع ۱,۷۸۶ کلمه یکتا، ۳,۷۰۴ تایپ با موفقیت (TP) تولید شده است. ۱۷۲ تایپ را می‌بایست تولید می‌کرده اما از دست داده (FN) است و ۱۳ تایپ را هم به اشتباه (FP) تولید کرده است. سپس مرحله به مرحله (از راست به چپ) مبدل‌های این پژوهش برای شناسائی کلمات خارج از واژگان، برروی ازدست‌رفته‌ها (FN) آزمایش شده و در صورتی که تغییری اتفاق افتاده در جدول مشخص شده است. در این جدول خط زیرین ساده زیر عدد نشانه بهتر شدن شاخص و خط زیرین منحنی زیر عدد نشان دهنده بدتر شدن شاخص نسبت به مبدل قبل‌تر از خود است.

کمتر از ۶۰ درصد تایپ‌ها، رسمی است. بیشتر خطاهای رخ داده در بین همه تایپ‌ها (۷۱,۵ درصد)، به دلیل کمبود واژگان است. رتبه دوم بیشترین ازدست‌رفته‌ها، خطاهای املائی است (۲۵ مورد، ۱۴,۵ درصد همه خطاها)، که هیچ یک از مبدل‌ها قادر به رفع خطا و شناسایی آنها نیستند. خطاهای نقص در قاعده نیز با ۱۰ خطا در جایگاه بعدی است. این خطاها مربوط به تحلیل‌گر تصریفی است و مبدل‌های این پژوهش در بهبود آن نمی‌توانند نقش داشته باشند. خطاهای بعدی خطاهای آوایی است که مبدل آوایی موفق شده از این ۵ خطا ۳ مورد را شناسائی کند، همین‌طور کلمات خارج از واژگان غیررسمی نیز ۱۲ تایپ تولید کرده است که مبدل آوایی توانسته با موفقیت (TP) شناسائی کند (به این معنا که این کلمات صرفا کلمات غیررسمی‌اند و می‌بایست در واژگان غیررسمی تحلیل‌گر وجود می‌داشتند). بیشتر کلمات به هم چسبیده نیز توسط مبدل تقطیع، شناسائی و تقطیع شده است. البته این مبدل در حین شناسائی هشدارهای نادرست (FP) نسبتا زیادی نیز تولید می‌کند که دقت را کاهش می‌دهد. آخرین نوع خطاها هم استفاده از حروف هم‌صدا و استفاده از **ه** به جای کسره اضافه است که همگی آنها توسط مبدل هم‌صدا شناسایی و رفع شده است. ازدست‌رفته‌های ادبی نیز ساختارهای ادبی‌ای است که تحلیل‌گر تصریفی و سایر مبدل‌ها قادر به شناسائی آنها نیست و مجموعا سه خطا تولید کرده است که به مجموع ازدست‌رفته‌های رسمی (۱۱۲) افزوده شده است.

معیارهای ارزیابی اصلی فراخوانی[۲۳]، دقت[۲۴] و معیار اف (۱)[۲۵] تایپ‌های رسمی در ادامه آمده است. مبدل‌های شناسائی این پژوهش تاثیری در نتایج بدست آمده از تحلیل‌گر تصریفی برای تایپ‌های رسمی ندارد (جدول ۴). پایین آمدن معیار دقت نیز به دلیل کلی حساب کردن (بدون در نظر گرفتن رسمی یا غیررسمی بودن تایپ) هشدارهای غلط (FP) است.

جدول ۴ - ارزیابی تایپ‌های رسمی

معیار	تحلیل‌گر تصریفی	هم‌صدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۵,۱	٪۹۵,۱	٪۹۵,۱	٪۹۵,۱	٪۹۵,۱
دقت	٪۹۹,۴۱	٪۹۹,۴۱	٪۹۹,۱۳	٪۹۹,۱۳	٪۹۸,۴۶
معیار اف (۱)	٪۹۷,۲۱	٪۹۷,۲۱	٪۹۷,۰۷	٪۹۷,۰۷	٪۹۶,۷۵

فراخوانی تایپ‌های غیررسمی در مبدل هم‌صدا، آوایی و تقطیع بهبود می‌یابد اما دقت به دلیل افزایش هشدارهای خطا کاهش پیدا می‌کند (جدول ۵).

[23] Recall
[24] Precision
[25] F-measure (1)

جدول ۵ - ارزیابی تایپ‌های غیررسمی

معیار	تحلیل‌گر تصریفی	هم‌صدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۶,۲۳	٪۹۶,۴۲	٪۹۷,۳	٪۹۷,۳	٪۹۷,۴۸
دقت	٪۹۹,۱۶	٪۹۹,۱۶	٪۹۸,۷۹	٪۹۸,۷۹	٪۹۷,۸۵
معیار اف (۱)	٪۹۷,۶۷	٪۹۷,۷۷	٪۹۸,۰۴	٪۹۸,۰۴	٪۹۷,۶۶

معیارهای ارزیابی اصلی فراخوانی، دقت و معیار اف (۱) برای تمامی کلمات مستقل از متن نیز در جدول ۶ آمده است.

جدول ۶ - ارزیابی کل تایپ‌ها

معیار	تحلیل‌گر تصریفی	هم‌صدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۵,۵۶	٪۹۵,۶۴	٪۹۶	٪۹۶	٪۹۶,۰۸
دقت	٪۹۹,۶۵	٪۹۹,۶۵	٪۹۹,۴۹	٪۹۹,۴۹	٪۹۹,۱
معیار اف (۱)	٪۹۷,۵۶	٪۹۷,۶	٪۹۷,۷۱	٪۹۷,۷۱	٪۹۷,۵۷

۵. نتیجه‌گیری

همان‌طور که ملاحظه می‌شود تبدیل واحدهای خارج از واژگان به واحدهای درون واژگان در فراخوانی تایپ‌های رسمی تغییری ایجاد نمی‌کند. اما بیش از یک درصد (فراخوانی) به شناسائی تایپ‌های غیررسمی و در نتیجه کل تایپ‌ها کمک می‌کند. تغییر ایجاد نکردن در فراخوانی تایپ‌های رسمی به این دلیل است که این تغییرات آوایی به فارسی غیررسمی تعلق دارد و در فارسی رسمی کمتر رخ می‌دهد.

از طرف دیگر شناسائی تغییرات آوائی گرچه مثبت‌های صحیح را افزایش می‌دهد اما هیچگاه تغییری در مثبت‌های غلط (FP) نمی‌تواند ایجاد کند. بنابراین انتظار بهبود این شاخص وجود ندارد و در بهترین حالت تنها می‌توان انتظار داشت شاخص تغییر نکند. همان‌طور که در جداول ارزیابی مشاهده می‌شود در مواردی شاخص دقت تغییر نکرده و در بعضی موارد شاخص بدتر شده است (مبدل آوائی و مخصوصا مبدل تقطیع مثبت‌های غلط زیادی به رابطه دقت افزوده‌اند که باعث کاهش شاخص در این مبدل‌ها شده است).

جدا بودن هر مبدل و منحصر بودن محدوده قواعد آوایی هر یک می‌تواند استفاده از آنها را ساده سازد. بسته به نوع متن و بافتی که استفاده می‌شود و مبتنی بر سیاق می‌توان از مبدل متناسب بهره برد. به طور مثال برای متن‌های رسمی نیازی به استفاده از این مبدل‌ها نیست، در حالی که برای متن‌های غیررسمی استفاده از این مبدل‌ها می‌تواند مفید باشد. برای متن‌هایی که در آن از فاصله درست استفاده نشده و فاصله بین کلمات به درستی رعایت نشده است استفاده از مبدل تقطیع می‌تواند مفید باشد.

تغییرات آوایی بین تک‌واژهای ترکیبی و سیلاب‌های کلمات نیز اتفاق می‌افتد. چنین قواعدی در این مقاله گنجانده نشده است اما برای پوشش بیشتر کلمات خارج از واژگان غیررسمی چنین قواعدی می‌تواند مفید باشد.

۶. فهرست منابع

آرمین، نادیه و شمس‌فرد، مهرنوش. ۱۳۸۹. «تبدیل متن محاوره‌ای فارسی به رسمی به کمک ان-گرام‌ها» در شانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی شریف.

علینقی‌زاده، رقیه، دین‌محمدی، غلامرضا، هراتی‌زاده، سامان. ۱۳۹۶. «تبدیل متن محاوره فارسی به متن رسمی فارسی در سطح تصریف و واژگان. » پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته زبان‌شناسی رایانشی در دانشکده علوم و فنون نوین، دانشگاه تهران.

Assadi, Sh.. (2007). “Sound Deletion in Colloquial Persian.”

Biber, D. (1993a). Using register-diversified corpora for general language studies. *Computational Linguistics*, *19*(2), 219–241.

Biber, D. (1993b). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Hulden, Mans. *Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology*. The University of Arizona, 2009.

Hulden, Mans. “Foma: A Finite-State Compiler and Library.” In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, 29–32. Association for Computational Linguistics, 2009.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Kenneth R Beesley, and Lauri Karttunen. Finite State Morphology (Xerox). CSLI Studies in Studies in Computational Linguistics. Stanford, Calif, 2003.

Kukich, Karen. “Spelling Correction for the Telecommunications Network for the Deaf.” *Communications of the ACM* 35, no. 5 (1992): 80–90.

Moradi, H. (2012). “Sound Deletion in Colloquial.”

Sharifloo, Amir Azim, and Mehrnoush Shamsfard. “A Bottom Up Approach to Persian Stemming.” In *IJCNLP*, 583–88, 2008.

Covering Out-of-Vocabulary Words of Informal Persian

Davood Heidarpour
Master of Computational Linguistics, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

Mostafa Salehi
Assistant Professor, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

Mahmoud Bijankhan
Professor, Department of Linguistics, Faculty of Literature and Humanities, University of Tehran, Tehran, Iran

Hadi Veisi
Assistant Professor, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

Vahid Ranjbar
PhD, Information Technology, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

1. Abstract

In addition to formal and informal words, Contemporary Persian also consists of some words which due to Phonetic changes have different written form from both their formal and informal equivalents. These variations are not fixed and usually produce words which are not having unified forms. So these out-of-vocabulary units cannot be analyzed with a word analyzer.

These units usually have equivalents in the vocabulary of the analyzer and there is a need for a unification method to map them to the right words if one wants to analyze them by the analyzer.

This research conduct a phonetic algorithm for these mappings by watching and recording the phonetic alterations that informal words are going through. These algorithms are some rules for converting a word to its phonetic form regardless of its letters, in form of some FSTs and then trying to match and pair them to their formal and informal equivalent.

This research improved an Inflectional analyzer of contemporary Persian recall more than one percent.

Keywords: Persian inflectional analyzing, informal Persian, contemporary Persian, phonetic algorithm, phonetic alteration, Out-of-vocabulary words.