ریشه‌یاب ماضی و مضارع از مصدر افعال ناگذر در زبان فارسی

شاپوررضا برنجیان*

عضو هیئت علمی،
گروه پژوهشی زبان‌شناسی رایانه‌ای

دریافت: ۱۳۹۰/۰۲/۲۶ | پذیرش: ۱۳۹۰/۱۲/۱۳

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره ۲۸ | شمارهٔ ۳ | ص ص ۷۸۷- ۸۰۵
بهار ۱۳۹۲

نوع مقاله: پژوهشی

چکیده: در این طرح، نمای کلی از نرم‌افزار ریشه‌یاب ماضی و مضارع از مصدر افعال ناگذر در زبان فارسی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری تهیه و ارائه گردیده است. عملکرد نرم‌افزار به گونه‌ای است که با ارائه مصدر فعل می‌توان به بن ماضی و بن مضارع آن مصدر دسترسی پیدا کرد. علاوه برآن، تعدادی از افعال که بن مضارع آنها برخلاف روال معمول ساخته می‌شوند نیز به‌عنوان استثناء بیان شده‌اند.

کلیدواژه‌ها: ریشه یاب، ریشه سازی فارسی، زبانشناسی رایانه ای، دستور زبان فارسی، فعل ناگذر، بن ماضی، بن مضارع.

*sh_berenjian@yahoo.com

۱. مقدمه

با افزایش روزافزون مدارک و اسناد در شبکه جهانی وب و مراکز اطلاع‌رسانی الکترونیکی و کتابخانه‌های رایانه‌ای و مراکز اسناد دیجیتالی، کندی بازیابی و ذخیره‌یابی در عملکرد کامپیوترها و سرورها مشاهده گردید. بر این اساس، متخصصان علوم مختلف از جمله علوم اطلاع‌رسانی و کامپیوتر جهت کاهش حجم اسناد و مدارک در زمان ذخیره‌سازی و افزایش سرعت در هنگام بازیابی اطلاعات، راهکارهایی را ارائه و پیشنهاد کردند که یکی از این راهکارها به‌کار بردن ریشه‌یاب‌ها در سامانه‌های بازیابی اطلاعات بود. ریشه‌سازی اصطلاحات فرایندی است زبان‌شناختی که می‌کوشد پایه و بن هر واژه موجود در متن را مشخص کند (مهراد و ناصری ۱۳۸۷).

۲. روش پژوهش

ابتدا تمامی افعال زبان فارسی از فرهنگ معین (معین ۱۳۷۱) استخراج گردید که تعداد آنها به ۸۱۰۰ مصدر می‌رسید، سپس افعال ساده از افعال مرکب جدا شد و بعد از آن، افعال پیشوندی استقاقی نیز از افعال ساده مجزا گردید.

اگر چه تعدادی از افعال در این فرهنگ ضبط شده است که چندان کابردی در زبان فارسی امروزی ندارد، ولی به دلیل انجام بررسی دقیق و استخراج قوانین مورد نیاز این طرح، تمامی آنها مورد بررسی و تجزیه و تحلیل قرار گرفت و بن‌های ماضی و مضارع آنها استخراج گردید. سپس، تمام افعال بر اساس شکل مصدر و بن‌های ماضی و مضارع دسته‌بندی شدند و هر یک از افعال در گروه‌های مربوط قرار گرفتند.

از تمامی این مصدرها، تمامی افعال و مصدرهای مرکب و افعال مرده (مانند: یوبیدن، و رساخیدن و ...) و قدیمی حذف گردید و تعداد ۱۴۵۲ فعل باقی ماند.

بررسی‌های انجام‌شده، مصدرهای زبان فارسی را به شرح زیر مشخص نمود:

۱. تعریف مصدر: مصدر کلمه‌ای است که مانند فعل بر وقوع کاری یا حالتی دلالت دارد، بدون زمان و شخصی (صیغه‌های شش‌گانه) و مفرد و جمع. در کتاب *پنج استاد* آمده است مصدر از برای بیان حدوث فعلی است که به فاعلی منسوب باشد و یا شکل نامحدود فعل که در بسیاری از زبان‌ها شکل بنیادی فعل شمرده می‌شود (چکنی ۱۳۸۲) .

۱-۱. انواع مصدر:

۱-۱-۱. مصدر اصلی، مصدری است که از اصل به عنوان مصدر به شمار می‌رود.

۱-۱-۲. مصدر جعلی (صناعی): مصدری است که در اصل مصدر نیست و با افزودن نشانه "یدن" و "نیدن" به آخر اسم پدید می‌آید، مانند طلب+یدن← طلبیدن (مدرسی ۱۳۸۷).

۱-۱-۳. مصدر بریده (مرخم، مخفف)≠(مصدرتام)، مانند گفت، شنود (انوری و احمدی ۱۳۸۵).

۱-۱-۴. مصدر تام: مصدری است که حذفی صورت نگرفته باشد.

۱-۱-۵. مصدر بسیط (مصدر ساده): دارای اجزایی نیست که بتوان آنها را جدا کرد، مانند (آمدن).

۱-۱-۶. مصدر ذووجهین (دوگانه): مصدرهایی که افعال آنها گاهی به صورت لازم و گاهی به صورت متعدی به‌کار می‌روند، مانند (شکستن) (فرشیدور ۱۳۸۲).

۱-۱-۷. مصدر لازم: مصدری است که فعل آن به مفعول بی‌واسطه نیاز نداشته باشد، مانند (آمدن).

۱-۱-۸. مصدر متعدی: مصدری است که فعل آن علاوه بر فاعل به مفعول بی‌واسطه نیز نیازمند باشد، مانند (زدن).

۱-۱-۹. مصدر سماعی: مصدرهایی که در اصل متعدی هستند، مانند (خوردن).

۱-۱-۱۰. مصدر قیاسی: مصدرهایی که در اصل لازم هستند، ولی با افزودن "-اند" یا "-نیدن" به آخر بن مضارع آنها متعدی می‌شوند.

۱-۱-۱۱. مصدر متعدی دو مفعولی: در برخی از مصدرهای متعدی نیز مانند مصدرهای لازم به ماده مضارع (-اندن) و (-انیدن) افزوده می‌شود. در این صورت، فعل جدید علاوه بر مفعول بی‌واسطه به مفعول باواسطه نیز نیاز خواهد داشت و بدون آن دو، معنی جمله ناقص خواهد بود، مانند خورانیدن.

۱-۱-۱۲. فعل کامل≠ناقص: فعلی است که در هنگام صرف همه زمان‌ها وجود دارد.

۱-۱-۱۳. فعل ناقص: فعلی است که در همه زمان‌ها صرف نمی‌شود.

۱-۱-۱۴. مصدر مرکب ≠ ساده: مصدری است که به‌وسیله پیشوند از مصدر ساده ساخته می‌شود.

با توجه به مطالب اشاره شده، مصدرهای ساده و لازم مورد بررسی و تجزیه و تحلیل قرار گرفت. همچنین، شکل‌های غیرمعمول مصدرهای مختلف (مانند پذرفتن مربوط به پذیرفتن) نیز حذف گردید.

در جدول ۱، فهرست مصدرهای موجود در زبان فارسی و چگونگی ارتباط هر کدام آورده شده است.

جدول ۱. فهرست مصدرهای موجود در زبان فارسی و چگونگی ارتباط هر کدام

مصدر	مخالف
اصلی	≠ جعلی
جعلی	≠ اصلی
سماعی (متعدی)	قیاسی (لازم)
مرخم	≠ تام
بریده	≠ تام
مخفف	≠ تام
تام	≠ مرخم، بریده، مخفف
قیاسی	سماعی
کامل	≠ ناقص
ناقص	≠ کامل
لازم	≠ متعدی
متعدی	≠ لازم
ذووجهین، دوگانه	≠ یک وجهی
ساده	≠ مرکب
مرکب	≠ ساده

در مجموعه افعال، تعدادی با شکل‌های مختلف کاربردی و گویش‌های متفاوت مناطق مختلف وجود دارند که آنها نیز مورد بررسی قرار گرفتند، مانند (ستاندن، ستدن، استدن).

تعدادی از افعال، گاهی به صورت کامل و گاهی به صورت ناقص و مخفف و گاهی به صورت سابق خود، ظاهر می‌شوند و بن‌های مضارع مشترک و یکسانی دارند که ما آنها را به دلیل اینکه هنوز در جامعه کاربرد دارند نیز به همان شکل در گروه‌های مربوط آورده‌ایم، مانند گسستن، گسیختن، گسختن، گسلیدن=گسل.

- بر اساس بررسی‌ها، مصدرهای مرکب و مصدرهای پیشوندی از فهرست مصدرها حذف گردید و فقط مصدرهای ساده (بسیط)باقی ماند.
- تعدادی از افعال کم‌بسامد که در زبان فارسی امروزی کاربردی ندارند (مانند ورساخیدن و نحستن) حذف گردید.

۲. سپس، تعیین ریشه‌های ماضی و مضارع مصدرها آغاز گردید و نتایج زیر به‌دست آمد.

۲-۱. تمامی فعل‌های فارسی از بن‌های ماضی و مضارع ساخته شده‌اند.

۲-۱-۱. بن ماضی: بن ماضی از مصدر بدون (-َن) پایانی ساخته می‌شود، مانند (رفتن=رفت).

۲-۱-۲. بن مضارع: از فعل امر دوم شخص مفرد بدون (ب-) آغازی ساخته می‌شود، مانند (رفتن=برو=[رو]).

۳. در زبان فارسی، تعدادی از واژه‌ها و تعدادی از زمان‌های افعال به طور معمول، از شکل‌های امری افعال ساخته می‌شوند. از این رو، از نظر زبان‌شناختی، پیدا کردن وجه امری واژه، نخستین گام در استخراج ریشه به‌حساب می‌آید.

وجه امری مصدرها، با حذف دو (۲) یا سه (۳) کاراکتر از آخر مصدرها به‌دست می‌آیند، مانند

۳-۱. وزاندن=وزان(دن).

۳-۲. ترسیدن=ترس(یدن).

۳-۳. گاهی سه حرف (-ادن) از آخر مصدر حذف می‌شود تا بن مضارع ساخته شود، مانند (ایستادن=ایست ، نهادن=نه).

۳-۴. گاهی نیز ضمن حذف سه حرف آخر از مصدر، حرفی نیز به باقی‌مانده، افزوده می‌شود، مانند حذف (-ختن) و افزودن (-ز): آویختن=آویز.

۳-۵. حذف (-ودن) و افزودن (-آی): ستودن=ستای.

۳-۶. حذف (-شتن) و افزودن (-ر): کاشتن=کار.

۳-۷. حذف (-فتن) و افزودن (-ب): یافتن=یاب.

۳-۷-۱. حذف (-فتن) و افزودن (-ف) مانند: (بافتن=باف) (شکفتن= شکف).

۳-۸. حذف (-ستن) که به پنج شکل دیگر امکان ساختن بن مضارع وجود دارد.

۳-۸-۱. حذف (-ستن) و افزودن (-ی): آراستن=آرای.

۳-۸-۲. حذف (-ستن) بدون افزودن حرفیدن: دانستن=دان.

۳-۸-۳. حذف (-ستن) و افزودن (-ه): کاستن=کاه.

۳-۸-۴. حذف (-ستن) و افزودن (-ند): بستن=بند.

۳-۸-۵. حذف (-ستن) و افزودن (و[ی]): جستن=جوی.

۴. باید توجه داشت که هر یک از این حروف متناسب با مصوت باقی‌مانده، پس از حذف (-ستن) متفاوت است، مانند دو پاراگراف زیر.

۴-۱. اگر مصوت باقی‌مانده (ا)[a] یا (آ)[ā] باشد، پس از آن مصوت، واج (ه)[h] افزوده خواهد شد، مانند (خواستن=خواه).

۴-۲. اگر مصوت باقی‌مانده (اُ)[o] باشد، به مصوت مرکب (او)[ow] تبدیل می‌شود، مانند (رستن=رو) و گاهی با مصوب بلند "و" و صامت "ی" تقابل دارد، مانند (شُست=شوی) (فرشیدور ۱۳۸۳، ۴۲۷).

بدیهی است که هر یک استثناءهای خود را دارند.

چون در زبان فارسی صرف افعال توسط دو بن ماضی و مضارع (که اشکال مختلف دارند) ساخته می‌شوند، بنابراین، به‌نظر می‌رسد باید ذخیره‌سازی به صورت مصدر انجام گیرد.

همان‌گونه که در زبان انگلیسی به صورت مصدر بدون (to) ذخیره می‌شود.

سپس، گروه‌بندی افعال فارسی دوباره مورد بررسی قرار گرفت و پس از آن، استثناءهای هر یک از گروه‌ها، استخراج و مشخص گردید.

در این میان، تعدادی از افعال نیز با عنوان افعال ناقص استخراج گردید و همچنین، به مطالبی در خصوص افعال برخوردیم که به تعدادی از آنها در زیر اشاره می‌شود:

- بعضی از افعال به نظر می‌رسید دارای بن‌های ماضی مشترک هستند (البته بدون توجه به مصوت‌های کوتاه)، مانند رَفتن=رَفت ، رُفتن=رُفت.
- بعضی نیز به نظر می‌رسید دارای بن‌های مضارع مشترک هستند (البته بدون توجه به مصوت‌های کوتاه)، مانند کَندن←کَن ، کردن←کُن.
- تعدادی نیز خود افعال دارای شکل ظاهری یکسان هستند، مانند جَستن← جُستن.
- تعدادی از افعال نیز دارای یک بن مضارع مشترک هستند، مانند (گسلیدن، گسستن، گسیختن، گسختن←گسل).

پس از تهیه استثناءها، جهت تعیین خط‌مشی‌های پژوهشی، جلسات مختلفی با متخصصان و پژوهشگران برجسته انجام گرفت و قواعد و Rule Baseهای تهیه‌شده پس از بحث و بررسی به تصویب رسید و به قسمت برنامه‌نویسی ارائه گردید تا در تهیه الگوریتم مورد استفاده قرار گیرند.

پس از این مرحله، برنامه‌نویس جهت ورود اطلاعات اولیه، نرم‌افزاری تهیه نمود (تصویر ۱).

تصویر ۱. نمایش صفحه ورود اطلاعات نرم‌افزار

از این زمان، ورود اطلاعات آغاز گردید. در قسمت مصدر فعل، تمام مصدرهای استثناء فارسی که بن مضارع آنها با هیچ یک از قواعد هم‌خوانی نداشت، وارد گردید و بن مضارع آن نیز در قسمت "بن مضارع" وارد شد و در برابر هر یک، نوع آنها نیز مشخص شد (تصویر ۲) و تعداد آنها به طور خودکار در صفحه نمایش داده می‌شد.

تصویر ۲. تعیین نوع استثناء در زمان درون‌داد

۳. اشکالات اولیه برنامه ورود اطلاعات

ابتدا این برنامه به اشکالاتی برخورد که توسط متخصص کامپیوتر و برنامه‌نویس رفع گردید. در زیر به تعدادی از آنها اشاره می‌شود:

۱. واژه‌های تکراری مشخص نمی‌شدند و برنامه هشدار[1] نمی‌داد.
۲. در قسمت تغییر اطلاعات، هر دو کاراکتر با عنوان بن فعل و ریشه فعل ثبت شده بودند.
۳. (Refresh) غیرفعال بود.
۴. چنانچه بن تکراری بود بدون توجه به شکل مصدر هشدار می‌داد.
۵. واژه های ذخیره‌شده زیر گروه (-یدن) در آمار ثبت نمی‌شد.

پس از رفع این اشکالات توسط برنامه‌نویس و تغییرات جزیی در صفحه اصلی، تصحیحات آغاز گردید (تصویر ۳).

تصویر ۳. نحوه استفاده از گزینه امکانات

در این صفحه با کلیک بر روی واژه "امکانات" (در قسمت راست بالای صفحه) دو گزینه ظاهر می‌شود: بازیابی و ویرایش اطلاعات F1 و خروج. با کلیک بر روی "خروج" از برنامه خارج می‌شویم، اما با کلیک بر روی "بازیابی و ویرایش اطلاعات" صفحه دیگری مانند تصویر ۴ ظاهر می‌شود.

1. alarm

تصویر ٤. تصحیح اطلاعات افعال و چگونگی افزایش شرح

با کلیک بر روی هر یک از واژه‌ها می‌توان بن فعل و مضارع آن را تغییر داد و در قسمت شرح نیز شرحی در مورد آن فعل وارد نمود. پس از آن، می‌توان بر روی قسمت تغییر اطلاعات و یا حذف رکورد کلیک نمود که در این صورت صفحه‌ای مانند تصویر ۵ ظاهر می‌شود و سؤال "تغییرات ذخیره شود" را نشان می‌دهد.

تصویر ۵. پذیرش و یا رد تغییرات انجام‌شده

درصورت کلیک بر روی"Yes"، سؤال "رکورد حذف شود" ظاهر می‌شود (تصویر ۶).

تصویر ٦. تایید نهایی حذف رکورد

با کلیک مجدد بر روی "Yes" رکورد حذف می‌شود و در صورت کلیک بر روی "حذف رکورد" سؤال "رکورد حذف شود" را نشان می‌دهد (تصویر ۷) با تعیین پاسخ، هر یک صفحه دوباره به حالت اولیه خود باز می گردد.

تصویر ۷. اعلام انجام موفقیت‌آمیز اصلاحات از سوی برنامه

٤. جامعه آماری پژوهش

جامعه آماری این پژوهش را تمامی افعال فارسی موجود در فرهنگ معین تشکیل می‌دهد که توسط چهار نفر (کارشناس ارشد) استخراج گردید. این افعال شامل تمامی افعال مختوم به (-دن) و (-تن) و همچنین، افعال لازم و متعدی و افعال مرکب و ساده و افعال قدیمی و مرده بوده است. همچنین، افعال پیشوندی و اشکال تمامی مصدرهای زبان و نیز شکل‌های مختلف تمامی افعال استخراج گردید، مانند فراشتن، افراشتن، فراختن، فرازیدن، افراختن.

٥. مراحل انجام پژوهش

طرح پژوهشی "ریشه‌یاب ماضی و مضارع از مصدر افعال ناگذر در زبان فارسی" که در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری انجام گرفت در تاریخ ۸۷/۱۰/۱ آغاز شد و در سه مرحله انجام پذیرفت:

۱. مرحله اول
- ۱-۱. تهیه فهرست افعال فارسی از فرهنگ معین
- ۱-۲. دسته‌بندی و تعیین سه کاراکتر ماقبل آخر
- ۱-۳. تعیین ریشه‌های ماضی و مضارع افعال
- ۱-۴. تهیه قوانین و (Rule Base)ها
- ۱-۵. تجزیه و تحلیل و تهیه فهرست و گروه‌بندی افعال
- ۱-۶. تعیین استثناءهای هریک از گروه‌ها
- ۱-۷. استخراج فرمول‌های مورد نظر

۲. مرحله دوم
- ۲-۱. تحلیل سامانه و برآورد نیاز کاربران
- ۲-۲. طراحی پایگاه داده‌ها
- ۲-۳. طراحی سامانه ورود اطلاعات
- ۲-۴. طراحی engine
- ۲-۵. طراحی Rule Base
- ۲-۶. طراحی واژه‌نامه
- ۲-۷. طراحی سامانه جستجو

۳. مرحله سوم

۳-۱. نصب سامانه بر روی رایانه
۳-۲. ورود اطلاعات و استثناءها
۳-۳. نصب Rule Base
۳-۴. آزمایش سامانه
۳-۵. تهیه گزارش نهایی.

٦. آزمون نظام

- آزمون نظام در یک محیط واقعی انجام گرفت و به‌خوبی پاسخ داد.

از آنجایی که الگوریتم خاصی طراحی شد، بنابراین این الگوریتم با سایر الگوریتم‌ها تفاوت فاحش دارد. این نظام قادر است پاسخ‌های سریعی را در کمترین زمان با حجم بالا فراهم آورد (تصویر ۸).

تصویر ۸. صفحه ی اصلی کار با نرم‌افزار

در تصویر ۸ که صفحه جستجو است می‌توان مصدر فعل مورد نظر را وارد نمود و با کلیک بر روی (بازیابی بن ماضی و مضارع) نتیجه جستجو در قسمت پایین صفحه نمایش داده می‌شود.

علاوه بر این، علاقه‌مند بودیم تا نتایج به‌دست آمده را با یک نظام دیگر مقایسه کنیم. برای این کار، با تعدادی از مدعیان تهیه ماشین ریشه‌ساز فارسی تماس گرفتیم، اما هیچ‌یک از آنان قادر به ارائه محصولی ملموس نشدند.

۷. پژوهش‌های انجام‌شده

بازیابی اطلاعات و ریشه‌سازی در زبان فارسی

پژوهش در زمینه بازیابی اطلاعات[1] و زبان فارسی چندان چشمگیر نیست. مهم‌ترین و شناخته‌شده‌ترین مطالعاتی که در دسترس عموم است، مقالاتی است که در دو کارگاه پژوهشی زبان فارسی و رایانه به تاریخ‌های ۱۳۸۳ و ۱۳۸۴ در دانشگاه تهران انجام شده است (بی‌بی‌خان، قاصدی، و پاکدل ۱۳۸۳).

این کارگاه‌ها شامل مجموعه‌هایی از خلاصه مقالات و گاهی نیز اصل آنها در زمینه‌های مختلف به‌کارگیری توانمندی‌های علمی زبان فارسی در قالب اطلاعات الکترونیکی در محیط رایانه است، که شرکت‌کنندگان تجربیات خود را با زبان فارسی و رایانه گزارش نموده‌اند. در مجموع، دانش زبان‌شناختی آنها از زبان فارسی یا ضعیف یا ناکافی بوده است و این مسأله، ارزیابی و بازده ابزارهای آنها را ضعیف می‌سازد.

بیشتر آنها از الگوریتم مشابه مانند پورتر (Porter 1980) جهت ریشه‌سازی و ابزارها استفاده کرده‌اند و به همین دلیل، دارای محسنات و معایب مشابه هستند و در کل، دارای تفاوت‌های جزئی از جمله تفاوت در فهرست‌ها و تعداد قواعد و غیره هستند. در بیشتر این ابزارها، به‌طور وضوح فقدان دانش زبان‌شناختی به چشم می‌خورد.

۸. پسوندهای مصدرساز

در اصل، تمامی مصدرهای فارسی یا به (-دن) و یا به (-تن) ختم می‌شوند. بر اساس بررسی‌هایی که انجام گرفت مشخص شد که جهت تعیین چگونگی ساخت بن مضارع باید بر اساس سه حرف آخر مصدر اقدام نمود که هر یک به شرح زیر تعیین می‌شوند:

۱. سه حرف آخر مصدرهایی که به (-یدن) ختم می‌شوند، با حذف (-یدن) بن مضارع ساخته می‌شود، مانند (خریدن←خر، ترسیدن←ترس).
۲. با حذف (-ستن) از آخر مصدرها بن مضارع ساخته می‌شود، مانند (مانستن←مان) (لازار ۱۳۸۴).
۳. با حذف (-دن) از آخر مصدرها بن مضارع ساخته می‌شود مانند (شنودن←شنو).
۴. با حذف (-ادن) از آخر مصدر، بن مضارع ساخته می‌شود، مانند (نهادن←نه).
۵. با حذف (-ختن) از آخر مصدرها و افزودن (-ز) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند (آویختن←آویز).

1. information retrieval

۶. با حذف (-ودن) از آخر مصدرها و افزودن (-آی) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند (ستودن←ستای).
۷. با حذف (-شتن) از آخر مصدر و افزودن (-ر) به باقی‌مانده مصدر، بن مضارع ساخته می‌شود، مانند (کاشتن←کار).
۸. با حذف (-فتن) از آخر مصدر و افزودن (-ب) به باقی‌مانده مصدر، بن مضارع ساخته می‌شود، مانند (تافتن←تاب).

این قانون دائمی نیست، زیرا گاهی پس از حذف (-فتن) حرف (-ف) به آخر آن افزوده می‌شود، مانند (بافتن←باف).

جدول ۲. پسوندهای مصدرساز و چگونگی شکل‌گیری بن‌های مضارع بر مبنای مشخصه‌های قواعد صورت صوتی حروف پایانی افعال

پسوندهای مصدرساز	به (آ) ختم شود. پس از حذف پسوند مصدرساز اضافه خواهد شد.	به (آ) ختم شود.	به (-ب)، صامت ختم شود.	به (-ُ) ختم شود.	باقی‌مانده به صامت ختم شود.
دن	-	-	-	-	-
یدن	-	-	-	-	و شنیدن←شنو
ادن	(ه) دادن←ده	-	-	-	-
ختن	-	-	ز		-
ودن	-	-	آی	و شنودن←شنو	-
شتن	-	ر	-	-	-
فتن	-	ف/ب	-	-	-
ستن	ی	ه	ند	و[ی]	-

بنابراین، چون در ساخت بن مضارع در زبان فارسی اشکالات گوناگونی در حذف حروف آخر و افزودن حروف جدید پیش می‌آید، بنابراین استثناءها افزایش می‌یابند و در صورت تهیه Rule Baseها به ضرورت باید الگوریتمی تهیه کرد تا ابتدا بتواند واژه را از انتها بررسی و معکوس عمل کند. در مرحله بعدی، ابتدا یک حرف را شناسایی و سپس، حرف دوم را و در صورت مطابقت با Rule Baseها، با حرف سوم تطبیق داده شود. چون پسوندهای جمع‌ساز فارسی هیچ‌کدام بیش از سه حرف نیستند، بنابراین نیاز به بررسی بیشتر از سه حرف آخر کلمات نیست.

پسوندهای مصدرساز فارسی و چگونگی شکل‌گیری بن‌های مضارع بعضی از افعال در جدول ۱ آورده شده است.

لازم به اشاره است که ساخت بن ماضی فقط با حذف (-ن) از آخر مصدر در تمام افعال فارسی انجام می‌گیرد.

۹. معرفی نظام ریشه‌یاب بن ماضی و مضارع از مصدر افعال

این نظام توانایی ساخت بن ماضی و مضارع از مصدرهای فارسی را دارد. این نظام از بخش‌های زیر ساخته شده است:

۱. تشخیص مصدر از بن‌های فعل فارسی
۲. تشخیص پسوندهای مصدرساز فارسی
۳. حذف پسوندهای مصدرساز و ارائه بن ماضی
۴. حذف پسوندها و ارائه بن مضارع
۵. تشخیص استثناهای موجود.

در این نظام تاکنون هشت قاعده تعریف شده است و از مزایای این نظام آن است که کاربر می‌تواند دانش این نظام را براساس نیاز خود به‌روزرسانی کند و سپس به ارزیابی نتایج به‌دست آمده بپردازد.

از آنجایی که از تمامی قواعد به‌صورت عددی پیشینه‌ای ثبت می‌شود، بنابراین کاربر می‌تواند از کار خود گزارش آماری نیز تهیه نماید.

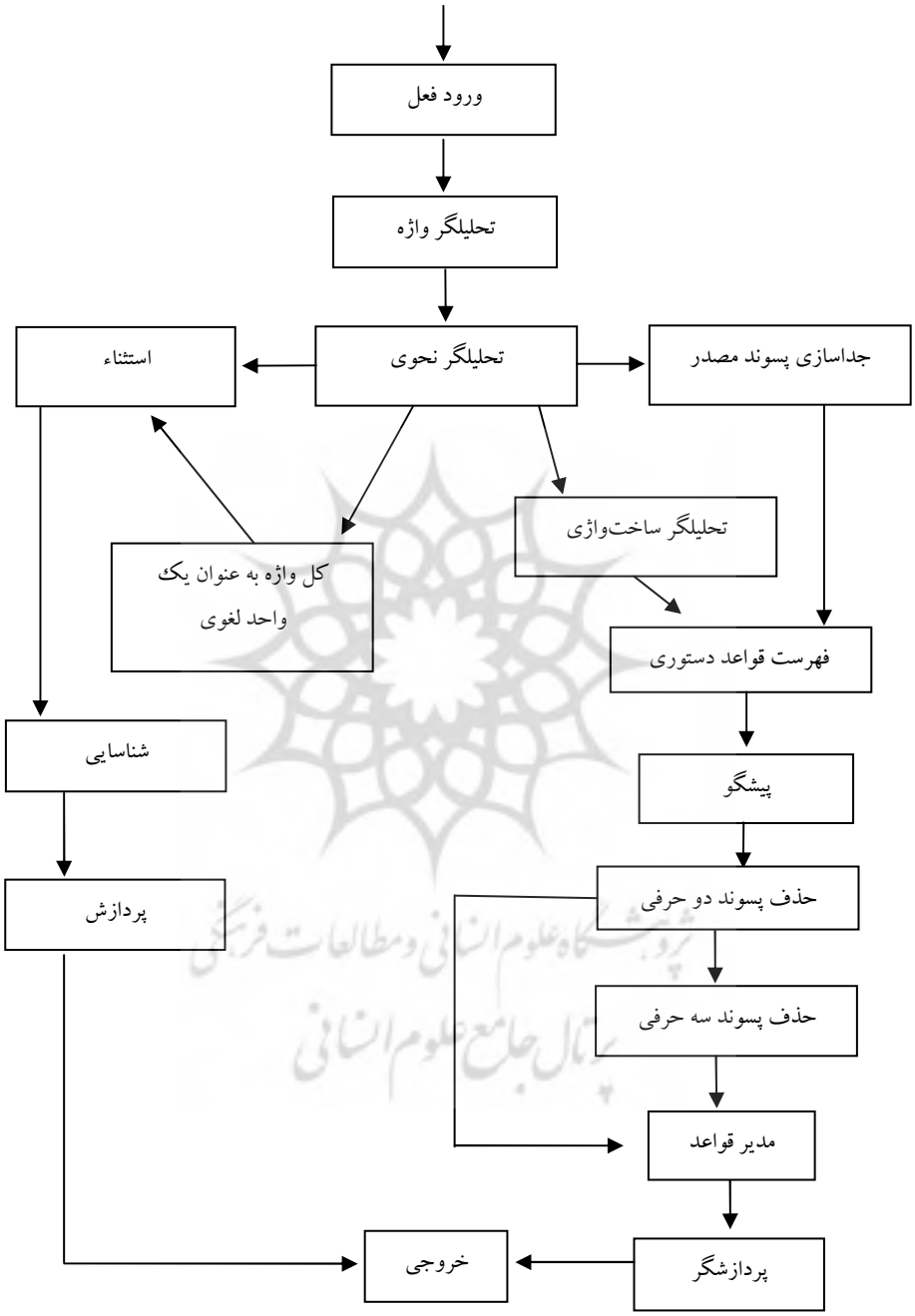
در این نظام، ابتدا کاربر می‌تواند در هر بار یک مصدر را ارائه کند و بن ماضی و مضارع آن را مشاهده نماید.

۱۰. مزایای ساخت نظام ریشه‌یاب ماضی و مضارع از مصدر افعال

از این نظام استفاده‌های زیادی می‌توان برد از آن جمله می‌توان به موارد زیر اشاره نمود:

- از این نظام می‌توان برای کاهش شکل‌های مختلف یک فعل استفاده نمود.
- این نظام در دسته‌بندی خودکار متن در فایل‌های بزرگ کامپیوتری کاربرد دارد.
- از این نرم‌افزار می‌توان در مورد پژوهش در بن و ریشه کلمه‌ها جهت استفاده در نظام‌های اطلاعاتی به منظور فشرده‌سازی داده‌ها استفاده نمود.

در نمودار ۱ قسمت‌های مختلف یک ریشه‌ساز آورده شده است.



نمودار ۱. قسمت‌های مختلف یک ریشه‌یاب بن مضارع از مصدر

۱۱. نحوه عملکرد

هر مصدر فعل پس از ورود، ابتدا توسط پردازشگر، مورد پردازش لغوی، ساخت‌واژی و نحوی قرار می‌گیرد. پردازشگر که از روش‌های زبان پایه و مبتنی بر الگو بهره می‌برد، خود دارای چهار جزء اصلی است:

۱. تحلیلگر نحوی: این تحلیلگر یک تجزیه‌گر چارت عمودی است که مسئول تبدیل مصدر با لیستی از شکل درختی است که تجزیه مصدر را بر عهده دارد.
۲. تحلیلگر ساخت‌واژی: در طول کار، هرگاه سامانه با یک مصدر ناشناخته برخورد کند، این قسمت از تحلیلگر سعی در تقلیل یا تشبیه‌سازی آن مصدر با یکی از اشکال از پیش تعیین‌شده شناخته شده می‌کند.
۳. احتساب کل واژه به عنوان یک واحد لغوی: این قسمت، چنانچه موفق به انجام دو جزء بالا نگردد، کل واژه را به عنوان یک واحد لغوی در نظر می‌گیرد.
۴. جزء چهارم، استثناءهای خاص است که در طول کار، واژه‌های خاص را که منطبق بر واژه‌های از پیش تعیین شده باشند، شناسایی و سعی در ریشه‌یابی آن می‌کند. در غیر این‌صورت، به استثناءهای عام و از شناسایی و پردازش به خروجی می‌روند.

جهت استخراج دانش لغوی فعل، ابتدا فهرستی از قواعد دستور زبان که افعال بر آنها تطابق دارد، تهیه می‌شود. با توجه به این فرض که تمام افعال ورودی از نظر دستوری صحیح (مجاز) هستند، فهرست قواعد دستوری ارائه‌شده به پیشگو این امکان را می‌دهد که نقش(های) صرفی و نحوی افعال ناشناس را تعیین کند.

در این مرحله، فعل جدید با ویژگی‌های استخراج‌شده آن به قواعد دستوری ارسال می‌شود تا در نهایت، مصدر توسط پردازشگر زبان طبیعی به بن مضارع و ماضی تبدیل شود.

در این پردازشگر، مصدرهای فارسی در زبان فارسی به ۸ گروه زیر تقسیم می‌شوند:

۱. مصدرهایی که با حذف (-یدن) از انتهای آنها بن مضارع ساخته می‌شود، مانند دویدن←دو.
۲. مصدرهایی که با حذف (-دن) از انتهای آنها بن مضارع ساخته می‌شود، مانند آوردن←آور.
۳. مصدرهایی که با حذف پسوند (-ستن) از انتهای آنها بن مضارع ساخته می‌شود، مانند مانستن←مان.

۴. مصدرهای که با حذف پسوند (-ادن) از آخر آنها بن مضارع ساخته می‌شود، مانند نهادن←نه.
۵. مصدرهایی که با حذف (-ختن) از انتهای آنها و افزودن (-ز) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند آویختن←آویز.
۶. مصدرهایی که با حذف (-ودن) از انتهای آنها و افزودن (-آی) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند ستودن←ستای.
۷. مصدرهایی که با حذف (-شتن) از آخر آنها و افزودن (-ر) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند کاشتن←کار.
۸. مصدرهایی که با حذف (-فتن) از انتهای آنها و افزودن (-ب) به باقی‌مانده، بن مضارع ساخته می‌شود، مانند تافتن←تاب.

و در تمامی مصدرها جهت ساختن بن ماضی فقط پسوند (-ن) از انتهای مصدر حذف می‌شود و بن ماضی ساخته می‌شود.

دسته دیگری نیز با عنوان استثناءها وجود دارد که نمی‌توان بر اساس ضوابط و پیشنهادات موجه در مورد آنها قواعدی ارائه نمود، مانند کردن←کن و ....

در ضمن، به‌دلیل اینکه پردازشگر تمامی واژه‌های مختوم به (-دن) و (-تن) را به عنوان مصدر فعل تلقی می‌کند، بنابراین سایر واژه‌های مختوم به این پسوندها، به‌عنوان استثناء وارد سامانه شده‌اند، مانند تن، تهمتن و .... .

۱۲. نتیجه‌گیری

در سال‌های اخیر مقالات متعددی در خصوص ریشه‌ساز زبان فارسی در مجلات به چاپ رسیده است که هر یک از آنها همواره بر مبنای مباحث نظری و بر اساس مقالات سایر زبان‌ها نوشته شده است که متأسفانه تمامی آنها بر مبنای مشاهدات علمی این سامانه‌ها نبوده است و مبنای علمی و عملی ندارند. اما این سامانه برای اولین بار به‌طور علمی و عملی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری به‌اجرا در آمده است. آنچه مشاهده می‌شود، بخشی از پژوهش انجام‌گرفته در این مرکز است. از این نظام در شبکه درونی و سایت مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری استفاده می‌شود.

۱۳. منابع

انوری، حسن، و حسن احمدی گیوی. ۱۳۸۵. *دستور زبان فارسی*. تهران: مؤسسه فرهنگی فاطمی.

بی‌بی‌خان، محمود، محمد اسماعیل قاصدی، و میترا پاکدل. ۱۳۸۳. *مجموعه سخنرانی‌ها و گزارش‌ها و چکیده طرح‌ها*. تهران: دانشگاه تهران.

چکنی، ابراهیم. ۱۳۸۲. *فرهنگ دائره‌المعارف زبان و زبان‌ها*. خرم آباد: دانشگاه لرستان، نشر بهنام.

فرشیدور، خسرو. ۱۳۸۲. *دستور مفصل امروز*. تهران: انتشارات سخن.

فرشیدور، خسرو. ۱۳۸۳. *فعل و گروه فعلی و تحول آن در زبان فارسی*. تهران: انتشارات سروش.

قریب، عبدالعظیم، ملک الشعرا بهار، و سایرین. ۱۳۷۱. *دستور زبان فارسی (پنج استاد)*. به کوشش امیرالشرف الکتابی. تهران: انتشارات واژه.

لازار، ژیلر. ۱۹۷۵. *دستور زبان فارسی معاصر*. ترجمه مهستی بحرینی. تهران: انتشارات هرمس. مرکز بین‌المللی گفتگوی تمدنها.

مدرس، فاطمه. ۱۳۸۷. *از واج تا جمله*. تهران: نشر چاپار.

معین، محمد. ۱۳۷۱. *فرهنگ فارسی (۶ جلدی)*. تهران: انتشارات امیرکبیر.

مهراد، جعفر، و مریم ناصری. ۱۳۸۷. *پردازش زبان طبیعی و بازیابی اطلاعات*. شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری؛ تهران: نشر چاپار.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14 (3): 130-137

An Introduction to a Past and Present Stemmer for Persian Intransitive Verbs

Shapourreza Berenjian*
Faculty Member, Computational Linguistics Department, RICEST

Iranian Journal of Information Processing & Management
Iranian Research Institute Iranian For Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.28 | No.3 | pp: 787-805
Spring 2013

Abstract: The vast quantity of scientific information has made linguistic analysis almost impossible without application of linguistic software. Stemming is considered to be one of the essentials of information retrieval systems. This paper introduces a stemmer for Persian intransitive verbs, made by the author. In this software, rule-based algorithms and Bruth Force algorithms and also syntactic and morphological analyzers are used. Persian intransitive verbs are generally divided into 8 groups and the rest listed under exceptions, available to users when necessary. This stemmer is able to perform the followings in the shortest time: 1) Identifying infinitives from Persian verb stems 2) Identifying Persian infinitive-maker suffixes 3) Removing infinitive-maker suffixes and providing past stems 4) Removing suffixes and offering present stems 5) Identifying exceptions and offering their stems. This software can be applied in the areas of information retrieval, language teaching, syntax, morphology and grammar writing.

Keywords: stemmers, Persian stemming, computational linguistics, Persian grammar, intransitive verbs, past stems, present stems

*Corresponding author: sh_berenjian@yahoo.com