



دانشکده علوم و فنون نوین

گروه بین رشته‌ای فناوری

تحلیل‌گر تصریفی فارسی معاصر

نام دانشجو:

داود حیدرپور

استادان راهنما:

دکتر مصطفی صالحی

دکتر محمود بی‌جن‌خان

استناد مشاور:

دکتر هادی ویسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته زبان‌شناسی رایانشی

بهمن ۱۳۹۶

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ



دانشکده علوم و فنون نوین
گروه بین رشته‌ای فناوری

تحلیل‌گر تصریفی فارسی معاصر

نام دانشجو:

داود حیدرپور

استادان راهنما:

دکتر مصطفی صالحی

دکتر محمود بی‌جن‌خان

استاد مشاور:

دکتر هادی ویسی

پایان‌نامه برای دریافت درجه کارشناسی ارشد

در رشته زبانشناسی رایانشی



دانشکده علوم و فنون نوین

با سمه تعالی

گواهی دفاع از پایان‌نامه کارشناسی ارشد

هیأت داوران، پایان‌نامه کارشناسی ارشد ناپیوسته آقای داود حیدرپور شماره دانشجویی ۸۳۰۵۹۳۰۲۳ در رشته زبان‌شناسی رایانشی با عنوان "تحلیل گر تصویری فارسی معاصر" را در تاریخ: ۱۳۹۶/۱۱/۳۰

با نمره نهایی: به عدد به حروف

سیت	۲۰	عازم
ارزیابی نموده‌اند.		و با درجه:

و با درجه:

ردیف	مشخصات هیئت داوران	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا مؤسسه	امضاء
۱	استاد راهنمای اول	دکتر مصطفی صالحی	استادیار	دانشکده علوم و فنون نوین دانشگاه تهران	
۲	استاد راهنمای دوم	دکتر محمود بی جن خان	استاد	دانشکده ادبیات و علوم انسانی دانشگاه تهران	
۳	استاد مشاور	دکتر هادی ویسی	استادیار	دانشکده علوم و فنون نوین دانشگاه تهران	
۴	استاد داور مدعو	دکتر محمد بحرانی	استادیار	دانشگاه صنعتی شریف	
۵	استاد داور داخلی و نماینده تحصیلات تکمیلی	دکتر مرتضی ابراهیمی	استادیار	دانشکده علوم و فنون نوین دانشگاه تهران	



تعهد نامه اصالت اثر

اینجانب ... دانش آموخته مقطع کارشناسی ارشد در رشته ... که در تاریخ ... از پایان نامه یا رساله خود تحت عنوان: " " با کسب درجه کارشناسی ارشد دفاع نموده ام، شرعاً و قانوناً متعهد می شوم :

۱ - مطالب مندرج در این پایان نامه یا رساله حاصل تحقیق و پژوهش اینجانب بوده و در مواردی که از دستاوردهای

علمی و پژوهشی دیگران اعم از پایان نامه، کتاب، مقاله و غیره استفاده نموده ام، رعایت کامل امانت را نموده، مطابق مقررات، ارجاع و در فهرست منابع و مأخذ اقدام به ذکر آنها نموده ام.

۲ - تمامی یا بخشی از این پایان نامه یا رساله قبلاً برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین تر یا بالاتر) در سایر دانشگاهها و مؤسسات آموزش عالی ارائه نشده است.

۳ - مقالات مستخرج از این پایان نامه یا رساله کاملاً حاصل کار اینجانب بوده و از هر گونه جعل داده و یا تغییر اطلاعات پرهیز نموده ام.

۴ - از ارسال همزمان و یا تکراری مقالات مستخرج از این پایان نامه یا رساله (با بیش از ۳۰ درصد همپوشانی) به نشریات و یا کنگره های گوناگون خودداری نموده و می نمایم.

۵ - کلیه حقوق مادی و معنوی حاصل از این پایان نامه یا رساله متعلق به دانشگاه تهران بوده و متعهد می شوم هر گونه بهره مندی و یا نشر دستاوردهای حاصل از این تحقیق اعم از چاپ کتاب، مقاله، ثبت اختراع و غیره (چه در زمان دانشجویی و یا بعد از فراغت از تحصیل) با کسب اجازه از تیم استادان راهنمای و مشاور و حوزه پژوهشی دانشکده باشد.

در صورت اثبات تخلف (در هر زمان) مدرک تحصیلی صادر شده توسط دانشگاه تهران از درجه اعتبار ساقط و اینجانب هیچگونه ادعایی نخواهم داشت.

امضا و نام و نام خانوادگی دانشجو:

چکیده

تحلیل تصریفی یکی از پردازش‌های مهم در چرخه پردازش‌های زبانی است. علی‌رغم گسترش روزافزون گونه غیر رسمی فارسی به دلیل گسترش فضای مجازی، ابزارهای پردازش این گونه زبانی به میزان لازم توسعه داده نشده است. پژوهش‌ها و ابزارهای توسعه داده شده برای پردازش متن فارسی بیشتر بر روی فارسی رسمی متمرکز است. این گونه زبانی علاوه بر منابع زبانی و قواعد مختص به خود شامل گونه زبانی رسمی و منابع و قواعد تصریفی آن نیز می‌شود. این دو گونه رسمی و غیر رسمی فارسی را می‌توان فارسی معاصر نامید. گونه غیر رسمی عموماً در پیامرسان‌های تلفن همراه، شبکه‌های اجتماعی و وبلاگ‌ها و محیط‌های شبیه به آنها استفاده می‌شود. گونه رسمی نیز در روزنامه‌ها، کتاب‌ها، قوانین، احکام و غیره به کار می‌رود.

این پژوهش موفق شده است با پوشش فارسی رسمی و غیر رسمی اولین ابزار تحلیل تصریفی فارسی معاصر (رسمی و غیر رسمی) را برای همه اقسام کلمه توسعه دهد. از لحاظ نظری سعی شده است همه ساختارهای تصریفی کلمات فارسی غیر رسمی پوشش داده شود. همچنین منابع واژگان، قواعد تصریفی و نگارشی گونه غیر رسمی علاوه بر دادگان ارزیابی و پیکره زیرسیاق‌های این گونه زبانی فراهم شده است.

فراخوانی و درستی تحلیل گر بر روی ۱۷۸۶ کلمه یکتا که از جملات چهار هزار کلمه‌ای دادگان ارزیابی به دست آمده است، به ترتیب ۹۵,۵۶٪ و ۹۹,۶۵٪ است. با استفاده از مبدل‌های ثانویه برای پوشش تغییرات آوایی این اعداد از نیم تا یک درصد بهبود یافته است.

کلمات کلیدی: تحلیل تصریفی، تحلیل گر تصریفی، پردازش متن، زبانشناسی رایانشی، زبان فارسی غیر رسمی، زبان فارسی رسمی، زبان فارسی معاصر

فهرست مطالب

۱	۱- فصل اول
۲	۱-۱- تعریف موضوع
۲	۱-۲- اهمیت موضوع و کاربردهای آن
۳	۱-۳- دستاوردهای پایان نامه
۴	۱-۴- ساختار پایان نامه
۶	۲- فصل دوم
۷	۲-۱- مقدمه
۷	۲-۲- تجزیه ساخت واژی یا تحلیل تصریفی؟
۷	۲-۲-۱- روش های آماری
۸	۲-۲-۲- روش های مکافه ای
۸	۲-۲-۳- روش های قاعده بنیاد
۱۶	۲-۳- کارهای انجام شده برای فارسی
۱۶	۲-۳-۱- فارسی رسمی
۱۷	۲-۳-۲- فارسی غیر رسمی
۱۹	۲-۴- جمع بندی
۲۰	۳- فصل سوم
۲۱	۳-۱- مقدمه
۲۱	۳-۲- چالش های پردازش فارسی
۲۱	۳-۲-۱- فراهم کردن واژگان
۲۲	۳-۲-۲- قواعد تصریفی
۲۲	۳-۲-۳- ساختار جامع و مانع
۲۳	۳-۲-۴- فاصله گذاری
۲۳	۳-۵- حروف غیر قطعی
۲۵	۳-۶- کلمات به هم چسبیده
۲۶	۳-۷- خطای املایی
۲۶	۳-۱- پیکره فارسی معاصر
۲۸	۳-۱-۱- چارچوب پردازش فارسی
۲۸	۳-۱-۱-۱- کلمه فارسی قابل پذیرش برای تحلیل گر
۲۹	۳-۱-۱-۲- واژگان
۳۰	۳-۱-۱-۳- برچسب تولیدی تحلیل گر برای کلمات
۳۱	۳-۱-۱-۴- قواعد و ساختمان تحلیل گر تصریفی
۳۴	۳-۱-۵- ابزار پیاده سازی
۳۴	۳-۲- نوآوری
۳۶	۳-۳- راهنمای فصل های دیگر روش پیشنهادی

۳۶	۴-۴- جمع بندی
۳۷	۴- فصل چهارم
۳۸	۴-۱- مقدمه
۳۹	۴-۲- اجزای ساختمان فعل
۴۳	۴-۳- ساختمان فعل ساده
۴۴	۴-۳-۱- امری
۴۷	۴-۲-۳- مضارع
۵۰	۴-۳-۳- ماضی
۵۱	۴-۳-۴- فعل‌های چند قطعه‌ای ساده
۵۳	۴-۳-۵- سایر فعل‌ها
۵۴	۴-۴- موارد ویژه
۵۴	۴-۴-۱- فعل‌های ناقص
۵۶	۴-۴-۲- حرف واسطه
۵۷	۴-۴-۵- قواعد نگارشی
۵۸	۴-۵-۱- تغییرات واژه‌بست محاوره هم در انتهای برخی افعال
۵۸	۴-۵-۲- حرفی میانجی در آغاز
۵۸	۴-۵-۳- فاصله دادن ه از حرف بعدی
۵۸	۴-۶- جمع بندی
۵۹	۵- فصل پنجم
۶۰	۵-۱- مقدمه
۶۱	۵-۲- اجزای ساختمان کلمه غیر فعلی
۶۱	۵-۲-۱- جمع
۶۱	۵-۲-۲- نکره / موصولی
۶۲	۵-۲-۳- واژه‌بست شخصی
۶۳	۵-۲-۴- معرفه
۶۴	۵-۲-۵- اضافه
۶۴	۵-۲-۶- واژه‌بست ربطی
۶۵	۵-۲-۷- وندها و واژه‌بست‌های محاوره
۶۶	۵-۲-۸- واژه‌بست ربطی ویژه
۶۷	۵-۲-۹- تکواز متنه به واکه
۶۷	۵-۳- ساختمان‌های غیر فعلی
۶۸	۵-۳-۱- اسم
۶۹	۵-۳-۲- صفت
۷۰	۵-۳-۳- قید
۷۰	۵-۳-۴- صفت یا ضمیر اشاره
۷۲	۵-۳-۵- صفت یا ضمیر مبهوم
۷۳	۵-۳-۶- صفت یا ضمیر پرسشی
۷۴	۵-۳-۷- صفت یا ضمیر تعجبی

۷۴	- ضمیر شخصی ۸-۳-۵
۷۵	- ضمیر مشترک ۹-۳-۵
۷۵	- حرف اضافه ۱۰-۳-۵
۷۵	- شماره ۱۱-۳-۵
۷۶	- حرف ربط ۱۲-۳-۵
۷۶	- مصدر ۱۳-۳-۵
۷۶	- صفت مفعولی ۱۴-۳-۵
۷۷	- شبہ جمله ۱۵-۳-۵
۷۷	- شاخص ۱۶-۳-۵
۷۷	- قواعد نگارشی ۴-۵
۷۸	- جمع‌بندی ۵-۵
۷۹	۶- فصل ششم
۸۰	- مقدمه ۶
۸۰	- مبدل استاندارد ۲-۶
۸۱	- مبدل همسدا ۳-۶
۸۲	- مبدل آوایی ۴-۶
۸۳	- مبدل بیانی ۵-۶
۸۴	- مبدل تقطیع ۶-۶
۸۵	- مبدل تولید ۷-۶
۸۶	- ابزار واسط کاربر ۸-۶
۸۶	- جمع‌بندی ۹-۶
۸۸	۷- فصل هفتم
۸۹	- مقدمه ۱-۷
۸۹	- معیارهای ارزیابی ۲-۷
۹۰	- ارزیابی قواعد بدست آمده از کلمات ۳-۷
۹۳	- جمع‌بندی ۴-۷
۹۴	۸- فصل هشتم
۹۵	- مقدمه ۱-۸
۹۵	- بهبود تحلیل‌گر تصریفی ۲-۸
۹۶	- شناسایی دقیق تنها قاعده سازنده کلمه ۳-۸
۹۷	- استفاده از اطلاعات تصریفی برای شناسایی سازه‌های نحوی ۴-۸
۹۹	- پژوهش‌های زبانی ۵-۸
۹۹	- ابزارهایی که می‌توان مبتنی بر تحلیل‌گر تصریفی ساخت ۶-۸
۱۰۱	مراجع
۱۰۴	برچسب‌های تحلیل‌گر
۱۰۵	قطعات به کاررفته در قاعده‌ها

۱۰۶	واژه‌نامه فارسی به انگلیسی
۱۰۸	واژه‌نامه انگلیسی به فارسی
۱۱۰	۱- پیوست یک
۱۱۱	۱-۱- مقدمه
۱۱۱	۱-۲- فعل‌های تک قطعه‌ای و چند قطعه‌ای
۱۱۲	۱-۳- تفاوت فعل‌های رسمی و غیر رسمی
۱۱۵	۱-۴- فعل‌های ساده
۱۲۹	۱-۵- فعل پیشوندی
۱۳۵	۱-۶- قواعد نگارشی
۱۳۷	۲- پیوست دو
۱۳۸	۲-۱- مقدمه
۱۳۸	۲-۲- موارد عمومی گره‌های بخش غیر فعلی
۱۴۰	۲-۳- اسم
۱۴۳	۲-۴- سایر قسم کلمه‌ها
۱۴۴	۲-۵- قواعد نگارشی
۱۴۵	۳- پیوست سه
۱۴۶	۳-۱- مقدمه
۱۴۶	۳-۲- مبدل استاندارد
۱۴۸	۳-۳- سایر مبدل‌های شناسایی
۱۵۲	۳-۴- مبدل تولید

فهرست شکل‌ها

۸	شکل ۱-۲ بخشی از قواعد الگوریتم ریشه‌یاب پورتر
۱۰	شکل ۲-۲ اتوماتای شناسایی کلمه «باران»
۱۱	شکل ۳-۲ شبکه کد الگوریتم اتوماتای حالت متناهی
۱۲	شکل ۴-۲ مبدل حالت متناهی برای فعل ساده گذشته فارسی
۱۲	شکل ۵-۲ لایه زیرساخت و روساختی که مبدل به هم نگاشت می‌کند
۱۲	شکل ۶-۲ مبدل حالات متنهای برای وند جمع چند اسم
۱۳	شکل ۷-۲ تعدادی از قواعد نگارشی فارسی
۱۳	شکل ۸-۲ سه لایه تحلیل تصویری
۳۳	شکل ۹-۳ ساختار کلی تحلیل‌گر و مبدل‌ها
۳۴	شکل ۱۰-۳ ابزار واسط کاربر و مبدل‌ها

پیوست

۸۱	شکل ۱-۶ همه حالت‌های تولید شده در مبدل
۸۴	شکل ۲-۶ قطعه‌های تولید شده از از مردم
۸۶	شکل ۳-۶ واسط گرافیکی کاربر برای استفاده از مبدل‌ها
۱۱۱	شکل ۱-۱ گره‌های بکار رفته در نمودارهای مفهومی
۱۱۲	شکل ۲-۱ نحوه بارگذاری و جداسازی فعل‌های تک و چند قطعه‌ای
۱۱۳	شکل ۳-۱ تمایز فعل رسمی و غیر رسمی
۱۱۴	شکل ۴-۱ تمایز فعل رسمی و غیر رسمی
۱۱۶	شکل ۵-۱ ساختار لایه‌های سازنده فعل رسمی ساده
۱۱۷	شکل ۶-۱ ساختار لایه‌های سازنده فعل ساده غیر رسمی
۱۱۸	شکل ۷-۱ لایه‌های زیرساخت فعل وجهی
۱۱۸	شکل ۸-۱ کد ساخت فعل وجهی
۱۱۸	شکل ۹-۱ لایه‌های زیرساخت اسناد
۱۱۹	شکل ۱۰-۱ کد ساخت اسناد
۱۲۰	شکل ۱۱-۱ جداسازی فعل‌های لازم و متعددی
۱۲۱	شکل ۱۲-۱ کد شناسه‌های متفاوت مضارع
۱۲۲	شکل ۱۳-۱ کد محدود کردن فعل امری به ساخت مفرد
۱۲۳	شکل ۱۴-۱ کد پیاده سازی فعل خوا
۱۲۳	شکل ۱۵-۱ کد ساخت فعل‌های امری بدون پیشوند ب
۱۲۴	شکل ۱۶-۱ کد ساخت فعل مضارع ساده
۱۲۵	شکل ۱۷-۱ نشانه گذاری فرا متنی برای بن‌های غیر رسمی دارای محدودیت
۱۲۶	شکل ۱۸-۱ کد بستن مسیر ساخت‌های غیر مجاز برای برخی بن‌های غیر رسمی
۱۲۷	شکل ۱۹-۱ کد محدود سازی برخی بن‌های رسمی
۱۲۷	شکل ۲۰-۱ کد پیاده‌سازی ت واسط

شکل ۲۱-۱ کد پیاده‌سازی فعل‌های ناقص	۱۲۹
شکل ۲۲-۱ لایه‌های زیرساخت فعل پیشوندی غیر رسمی	۱۳۰
شکل ۲۳-۱ کد اتصال پیشوند به بن فعل پیشوندی	۱۳۱
شکل ۲۴-۱ کد محدود کردن ساخت بعضی بن‌های غیر رسمی پیشوندی	۱۳۲
شکل ۲۵-۱ کد محدود کردن ساخت بعضی بن‌های رسمی پیشوندی	۱۳۲
شکل ۲۶-۱ کد ساخت استثنایات امری پیشوندی	۱۳۳
شکل ۲۷-۱ کد ساخت فعل ناقص باز-ذار	۱۳۴
شکل ۲۸-۱ کد ساخت فعل ناقص باز-گذاشت	۱۳۴
شکل ۲۹-۱ فعل‌های پیشوندی ناقصی که مثل فعل‌های کامل تعریف می‌شوند	۱۳۵
شکل ۳۰-۱ قاعده بازنویسی برای حذف ساختهای غیرمجاز فعل‌های ناقص پیشوندی	۱۳۵
شکل ۳۱-۱ قاعده تبدیل واژه‌بست هم به واژه	۱۳۵
شکل ۳۲-۱ قواعد افزودن ی میانجی برای پیشوند التزامی و نفی	۱۳۶
شکل ۳۳-۱ قاعده فاصله دادن حرف ه از حرف بعدی	۱۳۶
شکل ۳۴-۱ قواعد تبدیل نشانه‌های مرزناما	۱۳۶
شکل ۱-۲ نحوه علامت گذاری حرف-واکه پایانی واژگان	۱۳۸
شکل ۲-۲ گره پایانی غیر فعل	۱۳۹
شکل ۳-۲ گره‌ای واکه آگاه که در تناسب با حرف پایانی تکواز قبلی تکواز جدید را به آن اضافه می‌کند	۱۳۹
شکل ۴-۲ تغییر نشانه فرامتنی به دلیل تغییر واکه پایانی	۱۴۰
شکل ۵-۲ ساختار تمام گره‌های ساختمان اسم	۱۴۲
شکل ۶-۲ ایجاد محدودیت در مسیر بای ساختهایی که محدودیت بیشتری دارند	۱۴۳
شکل ۷-۲ قاعده نگارشی افروden تکواز نفی به مصدر و صفت‌های مفعولی	۱۴۴
شکل ۸-۲ قواعد بازنویسی حذف یا تغییر واکه ه در هنگام اتصال به تکوازی دیگر	۱۴۴
شکل ۱-۳ نحوه بارگذاری فایلهای لکس	۱۴۶
شکل ۲-۳ اعمال قواعد بازنویسی فعل و غیر فعل	۱۴۷
شکل ۳-۳ تولید همه کلمات	۱۴۷
شکل ۴-۳ قاعده سازنده مبدل استاندارد و قواعد کمکی دیگر	۱۴۸
شکل ۵-۳ قاعده سازنده مبدل همصدر و قواعد کمکی	۱۴۹
شکل ۶-۳ قاعده سازنده مبدل آوایی و قواعد کمکی	۱۵۰
شکل ۷-۳ قاعده سازنده مبدل آوایی و قواعد کمکی	۱۵۱
شکل ۸-۳ قاعده سازنده مبدل تقطیع و قواعد کمکی	۱۵۲
شکل ۹-۳ قاعده‌های سازنده مبدل تولید	۱۵۳

فهرست جداول

۹	جدول ۱-۲ مثال چند قاعده تصریفی برای یک کلمه
۱۰	جدول ۲-۲ جدول انتقال حالت اتوماتای شناسایی کلمه باران
۱۵	جدول ۳-۲ مشخصات منبع پرلکس
۱۷	جدول ۴-۲ تحلیل‌گرهای کلمه فارسی
۱۹	جدول ۵-۲ پردازش‌گرهای فارسی غیر رسمی
۲۳	جدول ۱-۳ نگارش‌های متنوع یک واحد چند قطعه‌ای
۲۳	جدول ۲-۳ نگارش‌های متنوع یک قطعه چند واحدی
۲۴	جدول ۳-۳ عدم قطعیت همزه در فارسی
۲۴	جدول ۴-۳ جایگزینی حروف هم‌صدا در فارسی غیر رسمی
۲۵	جدول ۵-۳ حذف واکه پایانی ۵ در فارسی رسمی
۲۵	جدول ۶-۳ حذف واکه پایانی ۵ در فارسی غیر رسمی
۲۷	جدول ۷-۳ پیکره داده‌های بکاررفته برای ارزیابی
۲۹	جدول ۸-۳ فهرست واژگان به کار رفته در تحلیل‌گر
۳۲	جدول ۹-۳ مرزنماهای به کار رفته در قواعد نگارشی
۴۰	جدول ۱-۴ شناسه‌های مضارع
۴۰	جدول ۲-۴ شناسه‌های ماضی
۴۱	جدول ۳-۴ واژه‌بست‌های ربطی
۴۱	جدول ۴-۴ واژه‌بست‌های مفعولی
۴۲	جدول ۵-۴ واژه‌بست فاعلی
۴۳	جدول ۶-۴ تک واژه‌ای محاوره
۴۶	جدول ۷-۴ - تصریف ویژه همه فعل‌های امری مفرد
۴۷	جدول ۸-۴ - تصریف ویژه برخی فعل‌های امری مفرد
۴۹	جدول ۹-۴ استثنای ترکیب نشدن برخی بن‌های غیر رسمی با شناسه‌های رسمی
۴۹	جدول ۱۰-۴ استثنای ترکیب نشدن برخی بن‌های رسمی با شناسه‌های غیر رسمی
۵۴	جدول ۱۱-۴ تغییرات برخی بن فعل‌های مضارع
۵۵	جدول ۱۲-۴ فعل‌های ناقص ساخت‌های ماضی
۵۵	جدول ۱۳-۴ تغییرات بن‌های پیشوندی مضارع
۵۶	جدول ۱۴-۴ تغییرات بن فعل‌های پیشوندی ماضی
۵۷	جدول ۱۵-۴ ساخت‌های متعدد همراه با واژه‌بست مفعولی
۵۸	جدول ۱۶-۴ مرزنماهای به کار رفته در قواعد نگارشی فعل
۶۲	جدول ۱-۵ وندهای نکرگی / موصولی
۶۳	جدول ۲-۵ واژه‌بست‌های شخصی
۶۴	جدول ۳-۵ نگارش مختلف وند معرفه
۶۴	جدول ۴-۵ حروف نمایشنده‌ند تکواز اضافه

۶۵	جدول ۵-۵ واژه‌بست‌های ربطی
۶۶	جدول ۶-۵ وندها و واژه‌بست‌های محاوره
۶۷	جدول ۷-۵ واژه‌بست‌های ربطی و بیژه
۷۸	جدول ۸-۵ مرزنماهای به کار رفته در قواعد نگارشی کلمات غیر فعلی
۸۲	جدول ۱-۶ حروف هم‌گروه که میتوانند جایگزین یکدیگر شوند
۸۳	جدول ۲-۶ قواعد آوایی
۸۵	جدول ۳-۶ کلمه و قاعده‌های بدست آمده از تقطیع
۹۱	جدول ۱-۷ بررسی گونه‌ای خطاهای
۹۲	جدول ۲-۷ ارزیابی تایپ‌های رسمی
۹۲	جدول ۳-۷ ارزیابی تایپ‌های غیر رسمی
۹۲	جدول ۴-۷ ارزیابی کل تایپ‌ها
۹۸	جدول ۱-۸ نحوه شناسایی فعل مجھول در پس‌پردازش
۹۹	جدول ۲-۸ بررسی تعداد تک‌واژه‌های ساختمان تصریف

فصل اول

مقدمه

۱-۱- تعریف موضوع

گونه رسمی فارسی، زبان رسمی کشور جمهوری اسلامی ایران است که در تمام اسناد حکومتی، قوانین، بخش‌نامه‌ها، مکاتبات و مکتوبات دولتی، آموزشی، روزنامه‌ها و نشریات کشور جمهوری اسلامی ایران و نه در متون ادبی و کهن^۱ به کار می‌رود.

زبان فارسی غیر رسمی، گونه‌ای از زبان فارسی است که در کنار و یا مقابله با زبان فارسی رسمی تعریف می‌شود. «در کنار» از این نظر که فارسی غیر رسمی شامل فارسی رسمی نیز می‌شود و پردازش فارسی غیر رسمی در واقع پردازش فارسی رسمی نیز هست. «در مقابل» هم به این معناست که می‌توان ساخت صرفی- نحوی دو گونه زبانی را به هم تبدیل کرد. البته بسیاری از ساختارها مشترک است و تنها در بسامد استفاده از آنها در این دو گونه زبانی تفاوت وجود دارد.

زبان فارسی غیر رسمی ساخت و سازی مختص به خود دارد و دارای ویژگی‌های ارتباطی نزدیک به گفتار و غیر رسمی است که بیشتر افراد جامعه‌ی ایران در گفتار و بخش گستره‌های از جامعه به عنوان نوشتار ارتباطی در شبکه‌های اجتماعی، پیام‌رسان‌های تلفنی، وبلاگ‌ها و گاهی سایتها و نامه‌های الکترونیکی استفاده می‌کنند.

از آنجایی که در گونه‌ی غیر رسمی فارسی از گونه‌ی رسمی هم استفاده می‌شود، بنابراین هر تحلیل‌گر کلمه برای فارسی غیر رسمی می‌بایست قادر به تحلیل کلمات فارسی رسمی هم باشد. بنابراین به طور کلی فارسی‌ای که در این پژوهش مدنظر است فارسی معاصر است.

در این پایان‌نامه یک تحلیل‌گر تصریفی برای فارسی معاصر ارائه خواهد شد. این تحلیل‌گر با گرفتن یک کلمه به عنوان ورودی، تک‌واژه‌های تصریفی، ریشه‌ی کلمه و جنس آنها را شناسایی کرده نحوه‌ی به هم پیوستن آنها را در قالب قاعده تصریفی زبان تولید می‌کند. خروجی این تحلیل‌گر می‌تواند یک یا چند قاعده تصریفی سازنده کلمه باشد (با توجه به وجود همنویسه‌های بسیار در زبان فارسی، در سطح تحلیل تصریفی مستقل از متن تنها می‌توان یک یا چند قاعده‌ی سازنده‌ی کلمه را تولید کرد و امکان انتخاب یک قاعده تصریفی از میان چند قاعده‌ی احتمالی و یا اولویت‌بندی قواعد تولید شده فراهم نیست).

۱-۲- اهمیت موضوع و کاربردهای آن

ابزارهای تجزیه و شناسایی زبان فارسی غیر رسمی برخلاف گسترش بسیار این گونه زبانی در رسانه‌های اجتماعی و فضای وب، توسعه بسیار اندکی یافته‌اند. به جز چند مورد اندک، پژوهشی (به منظور تحلیل رایانشی کلمه) در این زمینه انجام نگرفته است. برای تحلیل این گونه زبانی، تحلیل تصریفی از پایه‌ای ترین پردازش‌های است که در بسیاری از پردازش‌های زبانی سطح بالاتر دیگر و یا نرم‌افزارها و کاربردهای مبتنی بر پردازش زبان به کار می‌رود.

^۱ گرچه در کاربردهای رسمی و حتی غیر رسمی فارسی از این گونه زبانی استفاده می‌شود، اما این گونه ساختار خاص خود را دارد و چهارچوب نظری خود را برای لازم دارد که در حیطه این پژوهش نمی‌گنجد.

علت انتخاب جنبه تحلیلی برای این ابزار، نیاز به رویکردی عام منظوره در مواجهه با ساختمان کلمه است که به عنوان ابزاری بنیادی و پایه‌ای برای سایر پردازش‌های زبانی مناسب باشد. تعدادی از کاربردهای زبانی این ابزار به این شرح است: تولید زبان طبیعی، سیستم‌های پرسش و پاسخ، نگاشت گونه رسمی و غیر رسمی فارسی، ترجمه ماشینی قاعده‌بنیاد، تبدیل گفتار به نوشتار، آموزش فارسی به غیر فارسی زبانان.

گسترده‌گی موضوع و علاوه بر فارسی غیر رسمی، دربرگرفتن زبان فارسی رسمی نیز، هم به حجم کار و هم به پیچیدگی آن افروده است. زیرا پیاده سازی ابزاری نیاز به شناخت چهارچوب نظری مناسب برای پوشش کامل کلمه فارسی رسمی و سپس شناسایی قواعد کلمه فارسی غیر رسمی دارد. پس از روشن شدن این چارچوب پیاده‌سازی آن در قالب ابزاری رایانه‌ای که این تحلیل را انجام دهد، کاری گسترده و پر حجم است.

۱-۳- دستاوردهای پایان‌نامه

این پژوهش اولین پژوهشی است که منجر به ساخت ابزاری قاعده بنیاد برای تحلیل تصrifی فارسی غیر رسمی شده است. پژوهش مشابهی که توسط کارین مگردمیان (۲۰۰۸) انجام پذیرفته، به شناسایی و توضیح فارسی غیر رسمی محدود شده است و منجر به ساخت ابزار و یا توسعه (ارتقاء) تحلیل‌گر تصrifی رسمی او (مگردمیان، ۲۰۰۰) به غیر رسمی نشده است. پژوهش دیگر (تازه‌جانی و بحرانی، ۱۳۹۲) محدود به فعل غیر رسمی است و علی‌رغم تحلیل تصrifی فعل، برای تبدیل به معادل رسمی فعل طراحی شده است.

منابع واژگانی مورد نیاز برای تحلیل تصrifی گونه غیر رسمی، یعنی؛ واژگان، قواعد تصrifی و قواعد نگارشی سعی شده است به طور کامل پوشش داده شود. این منابع در این سطح برای اولین بار فراهم شده است. نوآوری‌های این پژوهش در مقایسه با پژوهش انجام شده قبلی (مگردمیان، ۲۰۰۶؛ ۲۰۰۸) به شرح زیر است.

ساخت فعل

- جایگاه تک‌واژه‌ای محاوره در ساختمان افعال مشخص شده است و عنصر تاکید آن در تحقیق مگردمیان پوشش داده نشده است (جدول ۴).
- فعل‌های پیشوندی که عمولاً به شکل سرهم نوشته می‌شوند و در زبان فارسی غیر رسمی بسیار رایج است و فعل‌های پیشوندی رسمی در این تحقیق پوشش داده شده است (۲-۴).
- فعل‌های ناقص غیر رسمی شناسایی و پوشش داده شده است (۴-۵).
- بن فعل‌های مضارع غیر رسمی در پذیرفتن شناسه سوم شخص مفرد رسمی و ساخت امری مفرد محدودیت دارند که این موضوع در قواعد به عنوان استثناء در نظر گرفته شده است (جدول ۴ و جدول ۹).
- بن فعل‌های رسمی در پذیرفتن شناسه سوم شخص غیر رسمی محدودیت دارند که این موضوع در قواعد به عنوان استثناء در نظر گرفته شده است (جدول ۴).
- تعریف ساختمان مجزا برای فعل لازم و فعل متعدد.

ساخت غیر فعلی

- تکواز تاکید در گره تکوازهای محاوره پوشش داده شده است (۷-۳-۵).
- واژه‌بستهای را+هم در ساختار اسامی و سایر کلماتی که از این ساختار بهره می‌گیرند شناسایی و پوشش داده شده است (۷-۳-۵).
- ساختمان تصریفی بعد از وند معرفه تعریف شده است (۴-۵).
- ساختمان تصریفی برای خمیر اشاره، مبهمن، شخصی، مشترک، شماره، پرسشی و حرف اضافه تعریف شده است (۱۲-۵، ۹-۵، ۸-۵، ۷-۵).

منابع واژگانی

- جمع‌آوری ۲۲۱ بن فعل غیر رسمی ساده و پیشوندی (۳۹ لازم، ۱۸۲ متعدی و ساختمان هر یک).
- جمع‌آوری ۴۶۳ واژه غیر فعلی غیر رسمی.
- جمع‌آوری ۴۹ فعل پیشوندی رسمی.
- جداسازی فعل‌های لازم و متعدی رسمی (۲۶۴ لازم، ۷۲۹ متعدی).
- اسم‌های عربی به کار رفته در فارسی رسمی؛ جمع‌های مكسر (۷۹۱ مورد) و جمع‌های عربی (بیش از ۶۵۰ مورد) استخراج شده است (۳-۵-۱ جمع).
- اسم‌های فارسی جانداران که وند جمع آن می‌پذیرد (۱,۹۵۱ مورد) استخراج شده است (۳-۵-۱ جمع).

جمع‌آوری پیکره و داده ارزیابی

- جمع‌آوری پیکرهای نزدیک به پنجاه هزار کلمه از تمام زیرسیاق‌های شناسائی شده فارسی غیر رسمی که تمام کلمات غیر رسمی آن شناسائی شده و قسم، تصریف و معادل رسمی آنها به شکل دستی به آن افزوده شده است.
- برای آزمون نیز از بخش ارزیابی این پیکره جملاتی حاوی بیش از چهار هزار کلمه انتخاب شده است که از ۱۷۸۶ کلمه یکتا ساخته شده است. این کلمات یکتا با کمک تحلیل‌گر و بررسی/اصلاح انسانی ۳۸۹۰ تایپ تولید کرده است که مربوط به کاربردهای مختلف این کلمات در بافت‌های مختلف است. از این دادگان برای ارزیابی این تحلیل‌گر استفاده شده است.

۱-۴- ساختار پایان‌نامه

ادبیات و کارهای پیشین در فصل مربوط (فصل پیش رو) بررسی می‌شود. ساختار اصلی طرح پیشنهادی در فصل راهکار پیشنهادی آمده است. اما جزئیات آن که قواعد استخراج شده کلمات فارسی و مبدل‌های ساخته شده است در فصل چهارم، فصل پنجم و فصل ششم قرار دارد. بخش‌های پیاده‌سازی آنها در پیوست گنجانده شده است. در فصل ارزیابی نیز علاوه بر اینکه طرح پیشنهادی با داده‌های واقعی ارزیابی می‌شود، شیوه جمع‌آوری پیکره به عنوان منبع تحقیق و همچنین داده

ارزیابی بررسی می‌شود. گزارش نتیجه و تحلیل آن نیز در این فصل توضیح داده می‌شود. در پایان خلاصه دستاورد این پژوهش، نواقص و روش‌های بهبود آن و کارهای آتی در فصل هشتم شرح داده می‌شود.

فصل دوم

کارهای پیشین و ادبیات موضوع

۱-۲- مقدمه

تحلیل تصریفی کلمه به عنوان یکی از پردازش‌های پایه‌ای زبان در مهندسی و پردازش زبان طبیعی شناخته می‌شود. کلمات زبان غالباً با تکوازهای تصریفی همراه هستند. اسم‌ها می‌توانند تکوازهای جمع، نکره‌گی و فعل‌ها می‌توانند با تکوازهای تصریفی شخص و شمار و زمان و غیره همراه باشند (در فارسی غیر رسمی، تکوازهای دیگر و ساختمان‌های متنوع‌تری نیز بکار می‌رود). وندها و واژه‌بستهای تصریفی که همراه کلمه به کار می‌روند معمولاً ساختار کلمه را دستخوش تغییر می‌کنند. این تغییر برای زبان‌های مختلف متفاوت است. تکواز تصریفی می‌تواند به ابتدا و یا انتهای کلمه افزوده شود، می‌تواند در میانه کلمه افزوده شود و یا می‌تواند ساختار ریشه کلمه را کامل درهم بربیزد. زبان فارسی از جمله زبان‌هایی است که تکوازهای تصریفی تنها به ابتدا و انتهای کلمات افزوده می‌شود.

۲-۲- تجزیه ساخت‌واژی یا تحلیل تصریفی؟

رویکردهای تحلیل رایانشی کلمه را نیز می‌توان به دو حوزه اصلی تقسیم کرد؛ رویکردهای آماری و رویکردهای قاعده بنیاد. هر کدام از این رویکردها بسته به کاربردی خاص می‌تواند مناسب باشد. گاهی هدف شناسایی دقیق ساختمان سازنده کلمه است، گاهی بدست آوردن ریشه یا بن‌واژه کلمه مهم است، گاهی تنها لازم است برچسب صرفی-نحوی^۱ کلمه خارج از واژگان یا تقطیع ساخت‌واژی کلمه، با استنباط از بافت (و شاید مستقل از زبان) مشخص شود. معمولاً روش‌های قاعده‌بنیاد برای مورد اول، روش‌های مکاشفه‌ای برای مورد دوم و روش‌های آماری برای مورد سوم بکار می‌رود.

۱-۲-۲- روش‌های آماری

روشهای آماری نظارتی^۲، بی‌نظارتی^۳ و نیمه‌نظرارتی^۴ معمولاً برای تقطیع تکوازهای کلمه استفاده می‌شوند. در صورتی که قصد داشته باشیم مسئله تحلیل تصریفی را با روش نظارتی حل کنیم، نیاز به دادگان بسیار برای هر مقوله صرفی-نحوی و هر قاعده تصریفی سازنده کلمه هست که حاشیه‌نویسی^۵ تصریفی را در سطح تکوازها مشخص کند. بنابراین روش‌های بی‌نظارت و نیمه نظارتی هم با این تعریف مناسب این نوع پردازش نیستند. علت دیگر مناسب نبودن چنین رویکردهایی برای تحلیل تصریفی، حذف شدن بسیاری از قواعد تصریفی ساخت کلمه کم‌کاربرد از عرصه تحلیل تصریفی است (رورک، ۲۰۰۷؛ شون و جورافسکی، ۲۰۰۸). به طور مثال روشی که یاروفسکی و وینستوفسکی (۲۰۰۰) استفاده کرده‌اند همین

^۱ Morphosyntactic

^۲ supervised

^۳ unsupervised

^۴ Semi-supervised

^۵ Annotation

گونه است و با استخراج قواعد از احتمالات (حالت تبدیل^۱) ساختار ساختواژی کلمات را مدل می‌کنند. این مدل ساختواژی بدست آمده شناختی زبان‌شناسانه از ساخت کلمه نیست و صرفاً الگوهای تقطیع^۲ و تجزیه کلمه است.

۲-۲-۲- روشهای مکاشفه‌ای

در روشهای مکاشفه‌ای^۳ که در ریشه‌یاب‌ها^۴ استفاده می‌شود برای سرعت بخشی از پروسه پیچیده و زمان بر تحلیل تصیری استفاده نمی‌شود و در عوض سعی می‌شود با بکارگیری الگوهای کلی و رایج، تکواژهای تصیری و اشتاقاقی آنها را از کلمات حذف کرد و به ریشه کلمات رسید. البته در بیشتر این روش‌ها از واژگان استفاده نمی‌شود، زیرا این کار علاوه بر بالا بردن سرعت، مشکل کامل نبودن واژگان تحلیل‌گرهای تصیری را بطرف می‌کند و در صورتی که قواعد مناسب انتخاب شده باشد مقدار فراخوانی را بالا می‌برد (اینگاسون، ۲۰۰۸). این کار دقت بالایی ندارد اما برای اموری نظیر بازیابی اطلاعات استفاده می‌شود. خیلی از ریشه‌های تولید شده ریشه‌های درست نیستند اما برای اموری نظیر بازیابی اطلاعات این امر مشکل ساز نیست. از معروف‌ترین ریشه‌یاب‌ها می‌توان ریشه‌یاب پورتر (۱۹۸۰) را نام برد. روش کار ریشه‌یاب پورتر ساده است، این ریشه‌یاب مجموعه‌ای از قواعد است که کلمات را تغییر می‌دهد. بخشی از قواعد این ریشه‌یاب در شکل ۱-۲ قابل مشاهده است.

ATIONAL → ATE (e.g., relational → relate)

ING → ∈ if stem contains vowel (e.g., motoring → motor)

LY → ∈ (e.g., hardly → hard)

ED → ∈ (e.g., walked → walk)

شکل ۱-۲ بخشی از قواعد الگوریتم ریشه‌یاب پورتر

کرووتز (۱۹۹۳) نشان داده است که استفاده از ریشه‌یاب پورتر می‌تواند عملکرد سیستم‌های بازیابی اطلاعات را افزایش دهد.

۳-۲-۲- روشهای قاعده‌بنیاد

روش شناسایی کلمات تصیری، ذخیره و نگهداری همه حالت‌های تصیری در واژگان و به جای تحلیل تصیری، جستجوی واژگان عملاً غیرممکن است، هم به این دلیل که حجم بسیار زیادی را اشغال می‌کند که نگهداری و جستجوی آن مشکل است و هم به دلیل زایا بودن زبان که همه کلمات زبان را نمی‌توان پوشش داد. البته در برخی کاربردها مانند بازشناسی

^۱ Transformation-based

^۲ Segmentation

^۳ heuristic

^۴ stemmer

گفتار از این روش استفاده می‌شود اما معمولاً نه برای همه واژگان و تنها برای تعداد محدودی از واژگان پرکاربرد این کار صورت می‌گیرد (جورافسکی و مارتین، ۲۰۰۸).

از طرف دیگر تحلیل تصrifی کلمات فارسی مستقل از متن همیشه همراه با تولید یک قاعده تصrifی مشخص نیست. بدون در نظر گرفتن بافت، برای بعضی از کلمات ممکن است چند قاعده سازنده وجود داشته باشد. بعضی از همنویسه‌ها نیز از این گونه کلمات تولید می‌کنند. چند نمونه از این نوع کلمات در جدول ۱-۲ آمده است.

جدول ۱-۲ مثال چند قاعده تصrifی برای یک کلمه

کلمه	قاعده تصrifی سازنده	مثال
خریدم	خرید (اسم) + م (شناسه شخصی اول شخص مفرد) خرید (بن فعل ماضی) + م (شناسه اول شخص ماضی مفرد)	خرید را جا گذاشتم. کتاب خریدم.
کردم	کُرد (اسم) + م (واژه‌بست ربطی اول شخص مفرد) کرد (فعل) + م (شناسه اول شخص ماضی مفرد)	من اصالتا کردم. من دیروز خرید کردم.
مردم	مرد(اسم) + م (واژه‌بست ربطی اول شخص مفرد) مُرد(بن فعل ماضی) + م (شناسه اول شخص ماضی مفرد) مرَدم (اسم)	من مردم. از تشنگی مردم. مردُم تهران.
نشستم	نشِست (بن فعل ماضی) + م ن (نفی) + شِست (بن فعل ماضی غیر رسمی) + م (شناسه اول شخص ماضی مفرد) ن (نفی) + شُست (بن فعل ماضی) + م (شناسه اول شخص ماضی مفرد)	کنار او نشتم. اونجا نشستم. ظرف‌ها را نشستم.

بنابراین تحلیل‌گرهای تصrifی مستقل از متن برای هر کلمه چند قاعده سازنده تصrifی تولید می‌کنند. اوتوماتای حالت متناهی برای تحلیل کلمه می‌تواند درستی کلمه و ساخت آن را تایید و یا آن را تولید کند، اما مبدل حالت متناهی، نوع عمومی‌تر و کامل‌تر آن است که می‌تواند علاوه بر شناسایی یا تولید، قاعده سازنده کلمه و یا کلمه حاصل از قاعده را تولید کند. یعنی با گرفتن کلمه قاعده تصrifی سازنده کلمه را تولید کند و با گرفتن قاعده تصrifی سازنده یک کلمه، آن کلمه را تولید کند (بیسلی و کارتونن، ۲۰۰۳). پی‌سی-کیمو (آرتورت، ۱۹۹۱) اولین مبدل حالت متناهی است که برای تحلیل تصrifی ساخته شد.

در دو روش اتوماتا و مبدل حالت متناهی سعی بر این است تا کلمات با ساختاری که در ذهن گویشور زبان وجود دارد بازسازی شود. همان ترتیبی که گویشور زبان کلمه را تولید و درک می‌کند، شناسایی شده و در تحلیل کلمه و یا تولید کلمه استفاده شود. بنابراین در روش‌های قاعده‌بنا برآگاهی از قواعدی که کلمات چگونه با تک‌واژه‌ای تصrifی ترکیب می‌شوند، واژگان ریشه‌ها و وندها و قواعدی که همنشینی بعضی واج‌ها و حروف را محدود می‌کند، لازم است.

۱-۳-۲-۲- اتوماتای حالت متناهی

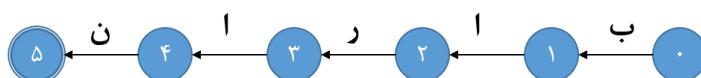
اتوماتای حالت متناهی^۱ روشی برای قاعده‌مند کردن مجموعه از حالت‌ها و حرکت بین آنهاست. هسته اصلی این سیستم جدولی از حالت‌ها و نحوه حرکت بین آنهاست. با این فرض که قصد داریم کلمه «باران» را در میان رشته‌ای از حروف که به عنوان ورودی سیستم هستند، شناسایی کنیم. برای مقایسه همه حروف ورودی و مقایسه آنها با حروف کلمه «باران» می‌توانیم از اتوماتای حالت متناهی استفاده کنیم. جدول انتقال چنین اتوماتایی در جدول ۲-۲ آمده است.

جدول ۲-۲ جدول انتقال حالت اتوماتای شناسایی کلمه باران

ورودی						حالت
«ن»	«ا»	«ر»	«ا»	«ب»		
.	.	.	.	۱	۰	
.	.	.	۲	۰	۱	
.	.	۳	۰	۰	۲	
.	۴	.	.	۰	۳	
۵	۰	.	۰	۰	۴	
.	.	.	.	۰	:۵	

اگر رشته ورودی در هر قسمت با موفقیت از حالت ۰ به حالت پایانی ۴ رسید، کلمه باران در ورودی یافت شده است و در غیر این صورت این کلمه یافت نشده است.

شکل ۲-۲ بهتر عملکرد جدول حالت‌های را نشان می‌دهد. شبکه اتوماتای حالت متناهی هم در شکل ۳-۲ (جورافسکی و مارتین، ۲۰۰۸) آمده است.



شکل ۲-۲ اتوماتای شناسایی کلمه «باران»

^۱ Finite-state automata

```

function D-RECOGNIZE(tape,machine) returns accept or reject
  indexBeginning of tape
  current-stateInitial state of machine
  loop
    if End of input has been reached then
      if current-state is an accept state then
        return accept
      else
        return reject
    elsif transition-table[current-state,tape[index]] is empty then
      return reject
    else
      current-statetransition-table[current-state,tape[index]]
      indexindex + 1
  end

```

شکل ۳-۲ شبکه کد الگوریتم اتوماتای حالت متناهی

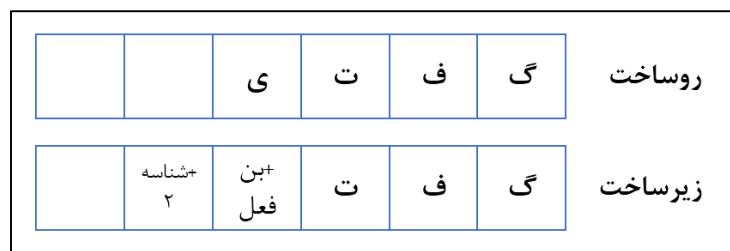
برای تحلیل گر تصریفی، از اتوماتا نمی‌توان استفاده کرد، چرا که اتوماتا برای شناسایی کلمه قابل پذیرش و یا تولید کلمه قابل پذیرش بکار می‌رود. اتوماتا فاقد ساختاری برای نگاشت چند رشته به یکدیگر است و بنابراین برای ساختار تحلیل گر که ساختمان تصریفی کلمه را می‌بایست شناسایی و تولید کند مناسب نیست.

۲-۳-۲-۲ - مبدل حالت متناهی

مبدل حالت متناهی نوع عمومی ترِ اتوماتای حالت متناهی است. مبدل قادر است دو رشته را به یکدیگر نگاشت کند. از این قابلیت می‌توان برای نگاشت تک‌واژها به نشانه‌ها یا برچسب‌های آنها استفاده کرد. در سطح کلمه نیز این نگاشت بین کلمه و تمام قطعات سازنده قاعده کلی کلمه صورت می‌گیرد. بنابراین ورودی مبدل کلمه و خروجی آن قاعده سازنده کلمه است. بر عکس آن هم ممکن است؛ یعنی ورودی قاعده سازنده کلمه و خروجی کلمه حاصل از آن قاعده باشد. بنابراین می‌توان از آن به عنوان تحلیل گر دو سطحی استفاده کرد به طوری که با گرفتن قاعده تولید کلمه، کلمه را تولید کند و با گرفتن کلمه، قاعده تولید کلمه را خروجی دهد. مبدل حالت متناهی می‌تواند برای تشخیص، تولید و تبدیل دو رشته به هم بکار رود. شکل ۴-۲ مبدل حالت متناهی برای تحلیل فعل ساده گذشته فارسی است. شکل ۵-۲ سطح زیرساخت و رو ساخت چنین مبدلی را نمایش می‌دهد.



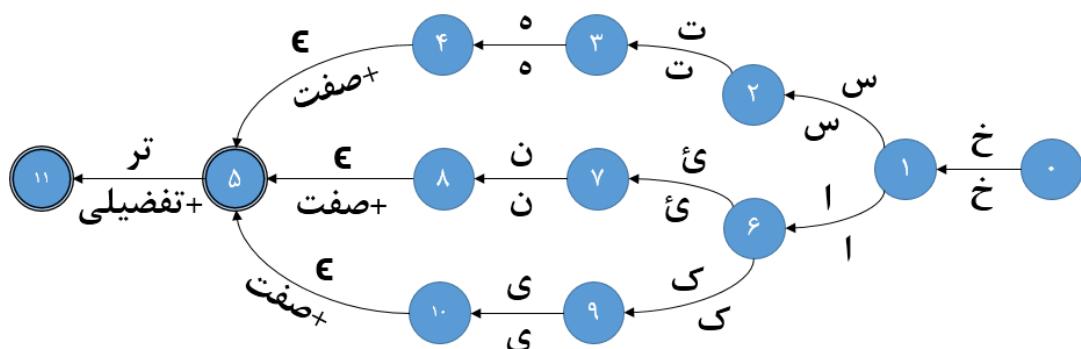
شکل ۲-۴ مدل حالت متناهی برای فعل ساده گذشته فارسی



شکل ۵-۲ لایه زیرساخت و روساختی که مدل به هم نگاشت می‌کند

در

شکل ۶-۲ یک مدل حالت متناهی برای چند صفت همراه با تکواز تصریفی تفضیلی (تر) نمایش داده شده است، قسمت پایین یال‌ها مربوط به زیر ساخت و قسمت بالای آنها مربوط به روساخت است.



شکل ۶-۲ مدل حالات متناهی برای وند جمع چند اسم

از آنجایی که سه منبع زبانی برای ساخت تحلیل گر تصريفی استفاده می‌شود، تا اینجا از دو منبع واژگان و قواعد تصريفی استفاده شده است، خروجی مبدل حالت این قسمت در واقع روساخت نیست بلکه حالت میانی‌ای است که هنوز به روساخت نرسیده است و نیاز است در مرحله دیگری که قواعد نگارشی را کنترل می‌کند به حالت روساخت تبدیل شود. تعدادی از قواعد نگارشی فارسی در شکل ۷-۲ آمده است.

شکل ۷-۲ تعدادی از قواعد نگارشی فارسی

مثال	قاعده
جزوه+م → جزوہام، جزوہ+ت → جزوہات گفته+م → گفتهام، گفته+ی → گفتهای، خسته+ید → خستهاید، خسته+ند → خسته‌اند	اضافه شدن ۱ بین حرف ۵ در پایان کلمه و قبل از واژه‌بست شخصی مفرد و تمام واژه‌بستهای ربطی غیر از سوم شخص مفرد
جزوه+مان → جزوہیمان، جزوہ+تان → جزوہیتان، جزوه+شان → جزوہیشان	اضافه شدن ۱ بین حرف ۵ در پایان کلمه و قبل از واژه‌بست شخصی جمع

بنابراین در دو مرحله حالت متناهی، قاعده از زیرساخت به روساخت می‌رسد. شکل ۸-۲ این مرحله‌ها را نشان می‌دهد.



شکل ۸-۲ سه لایه تحلیل تصريفی

۳-۲-۳-۲- منابع زبانی

در روش‌های قاعده‌بنیاد بخصوص پیاده‌سازی با مبدل و اتوماتی حالت متناهی، برای ساختن تصريفگر به سه منبع زبانی نیاز داریم (جورافسکی و مارتین، ۲۰۰۸).

- واژگان
- قواعد تصريفی
- قواعد نگارشی

واژگان یعنی تمام حالات تصريف نشده اسم‌ها، اسم‌های خاص، صفات، افعال و غیره. قواعد تصريفی مشخص می‌کند چه کلمه‌ای به چه تک واژی متصل می‌شود و چه نوع کلمه‌ای می‌سازد. قواعد نگارشی هم همنشینی واجها را در زبان مشخص

می‌کند، مثلاً اگر وند جمع ان به اسمی که با واکه ۵ خاتمه می‌یابد بچسبد، آن ۵ حذف شده و جای خود را به گ می‌دهد، مثل ستاره+ان ← ستارگان.

منابع مهمی که برای واژگان زبان فارسی موجود است به این شرح‌اند.

- واژگان زایای زبان فارسی (اسلامی و دیگران، ۱۳۸۳) که حاوی ۵۵ هزار مدخل است. هر مدخل با اطلاعات مربوط به صورت نوشتاری واژه در خط فارسی، ساخت واجی، مقوله واژگانی، الگوی تکیه، بسامد تکرار در پیکره متنی ۵ میلیونی بی‌جن‌خان (۲۰۱۱) مشخص شده است. تعداد مقوله‌های واژگانی ۳۲ عنوان است.
- در پروژه ترجمه ماشینی شیراز (امراپت^۱ و همکاران، ۲۰۰۰) با استفاده از تیم فرهنگ‌نویسان فارسی، نزدیک به ۵۰/۰۰۰ مدخل واژگانی، عبارت و اسم خاص جمع‌آوری شده است که حاوی اطلاعات نگارشی، برچسب اجزای سخن، ویژگی‌های نحوی هر لغت به همراه ترجمه انگلیسی سنس کلمات است. این اطلاعات در قالب ساختار ویژگی^۲ هر مدخل را توضیح داده و ذخیره شده است.
- پرلیکس (ساگوت و والتر، ۲۰۱۰) با استفاده از پیکره متن آزاد بی‌جن‌خان (۲۰۰۴)، ویکی‌پدیای فارسی^۳ و با کمک گرفتن از سخن‌وران فارسی زبان و مرجع دستوری فارسی لازارد (۲۰۰۶)، مجموعه‌ای از مدخل‌های فارسی را فراهم آورده‌اند. اطلاعات ساخت تصریفی هر مدخل و اطلاعات نحوی آن نیز همراه مدخل وجود دارد. محتوای این مجموعه در جدول ۳-۲ آمده است.

^۱ Amtrupt, Jan Willers

^۲ Feature structure

^۳ نسخه شانزدهم فوریه ۲۰۱۰ ویکی‌پدیای فارسی

جدول ۳-۲ مشخصات منبع پرلکس

مدخل تصریف شده	لمای مجزا	مدخل مجزا	طبقه
۱۹/۷۷۶	۱۳۹	۱۷۱	فعل
۱۷۷/۹۸۸	۹/۱۰۶	۹/۵۲۳	اسم
۳۳/۰۷۶	۱۰/۹۳۸	۱۰/۹۹۶	اسم خاص
۲۹۰/۰۵۳۷	۱۱/۸۳۵	۱۱/۸۷۲	صفت
۳/۳۲۳	۳/۱۲۰	۳/۳۲۲	دیگر طبقات
۵۲۴/۷۰۰	۳۳/۴۵۴	۳۵/۹۱۴	مجموع

۴-۳-۲-۲) ابزارهای پیاده‌سازی

زیراکس برای قاعده‌مند کردن تصریف کلمه، صورت‌بندی‌ای^۱ ابداع کرده (لکس^۲) که در آن واژگان و قواعد سازنده کلمات تعریف می‌شود. سپس این قواعد و کلمات را می‌توان به صورت یک مبدل حالت متناهی^۳ در حافظه کامپایل کرد. علاوه بر صورت‌بندی لکس، دستورالعمل^۴ دیگری نیز ابداع کرده است که هم به عنوان محیط نوشتن و پیاده کردن قواعد بازنویسی (نگارشی) مبدل‌ها و هم به عنوان بستر اجرای زبان قاعده‌مند^۵ بکار می‌رود. این زبان قاعده‌مند هم می‌تواند در مسیر ساخت کلمات و قواعد موثر باشد (و تغییرات لازم را در آنها ایجاد کند) و هم به شکل مستقل به صورت یک زبان مقید^۶ برای منطق مرتبه اول^۷ بکار رود. ابزارهای دیگری مثل جستجوگر واژه / قاعده در مبدل نیز در ساختمان این فناوری تعبیه شده است.

مشکلات این فناوری پشتیبانی نکردن کامل از یونیکد، تجاری بودن و متن آزاد نبودن آن است که استفاده از آن را دچار ملاحظه می‌کند.

۲-۲-۳-۴-۱) ابزار فوما

^۱ Formalism

^۲ Lexc

^۳ FST – Finite State Transducer

^۴ Script

^۵ Regular Expression

^۶ Formal

^۷ First Order Logic

ابزار فوما یک ابزار متن آزاد و رایگان است که صورت‌بندی لکس و زبان قاعده‌مند فناوری زیراکس برای ایجاد تغییر در مبدل‌ها و منطق مرتبه اول را پیاده کرده است. همینطور ابزار جتسجوگر واژه / قاعده را دارد. علاوه بر این کتابخانه‌هایی به زبان‌های سی، جاوا و پایتون برای استفاده از مبدل‌ها و یا ایجاد تغییر در آنها در سطح برنامه‌نویسی برای ابزاری مستقل را فراهم کرده است.

به دلیل سادگی و در دسترس بودن آن، این ابزار به عنوان ابزار پایه تحلیل‌گر ساخته شده در این پژوهش، استفاده شده است. گرچه به دلیل استفاده از صورت‌بندی لکس و زبان قاعده‌مند با همان استاندارد، امکان منتقل کردن این تحلیل‌گر به ابزاری مشابه که این استاندارد را پشتیبانی کند وجود دارد.

ابزار فوما این امکان را در اختیار ما قرار می‌دهد که علاوه بر یک مبدل، چندین مبدل در سیستم قرار دهیم و کلمه و رودی را از همه مبدل‌ها عبور دهیم و یا در صورتی که یک مبدل قاعده‌ای تولید نکرد به مبدل بعد از آن رجوع کنیم. حالت سومی هم وجود دارد که کاربر می‌تواند با استفاده از کتابخانه تعریف شده به صورت شخصی‌شده در مبدل‌ها جستجو انجام دهد و یا ترتیب و نحوه اجرا شدن مبدل‌ها را شخصی‌سازی کند.

۳-۲- کارهای انجام شده برای فارسی

این بخش به دو قسمت فارسی رسمی و فارسی غیر رسمی تقسیم می‌شود. به دلیل قاعده‌بندی بودن روش این پژوهش و ابزار آن، در بخش رسمی تنها ابزارهای قاعده‌بندی پوشش داده می‌شود، اما در بخش غیر رسمی تمام تلاش‌های انجام شده برای پردازش این گونه زبانی سعی شده تا پوشش داده شود.

۳-۱-۳- فارسی رسمی

در فارسی تحلیل‌گر صرفی (شمس‌فرد و همکاران، ۲۰۱۰) با استفاده از اتماتای حالات متناهی و دادگان واژگان زایا (اسلامی، ۱۳۸۴) تحلیل صرفی کلمات را انجام می‌دهد. این تحلیل‌گر دو اتماتا برای پیشوند و پسوند کلمه می‌سازد و ریشه کلمات را با روش‌های جدول جستجو^۱ و تناظر با قاعده مربوط به آن بررسی می‌کند. از برچسب ذخیره شده در واژگان زایا برای هر وند و ابزار برچسب‌زن اجزای سخن برای محدود کردن تحلیل‌های تولید شده برای کلمه و در نهایت رسیدن به یک تحلیل صحیح استفاده شده است. فراخوانی این تحلیل‌گر ۹۸٪ درستی آن روی ۶۰۰ کلمه ۹۳٪ گزارش شده است. در این تحلیل‌گر مرز تحلیل صرفی و تصريفی خلط شده است و امکان محدود کردن تحلیل‌گر به تنها یکی از دو حالت وجود ندارد.

در ساختار مبدل حالت متناهی نیز مگردمیان (۲۰۰۶، ۲۰۰۰) این ساختار را به کار برده است. برای واژگان، قواعد تصريفی و قواعد نگارشی از پروژه شیراز (امرپات، ۲۰۰۰) استفاده شده و با مبدل دو سویه زیراکس (بیسلی و کارتون، ۲۰۰۳) پیاده‌سازی شده است. برای رفع ابهام از قواعد تولید شده از هیچ روش مبتنی بر بافتی مثل برچسب اجزای سخن استفاده

^۱ Lookup Table

نشده است و تنها با استفاده از اطلاعات بسامد و آمار مربوط به هر کلمه و وند سعی شده که از میان قواعد هر کلمه، قاعده صحیح انتخاب شود. دقت این تحلیل گر برای پیکره ۷ مگاباتی ۹۵ درصد گزارش شده است. پارس مورف (مواجی و همکاران، ۱۳۹۰) نیز با استفاده از واژگان زایا (اسلامی و همکاران، ۱۳۸۳) برای واژگان و قواعد تصریفی ابزاری پیاده سازی کرده است که برای صرف کلمه (تصrif و ساخت واژه) به کار می رود. این ابزار قاعده بنیاد از زبان برنامه نویسی پایتون محض (بدون استفاده از مبدل یا اتوماتا) برای پیاده سازی خود بهره برده است که چندان بهینه به نظر نمی رسد. دقت این ابزار (دقت کلمه ای است که در مقاله این ابزار استفاده شده است اما مفهوم دقیق آن روشن نشده است) ۹۵ درصد گزارش شده است (حجم دادگان آزمون نیز مشخص نیست).

روحانیان و همکاران (۱۳۹۳) نیز با استفاده از مبدل حالت متناهی و منابع واژگان زایا، تحلیل گر تصریفی ای طراحی کرده است. دقت گزارش شده برآ آن ۹۵ درصد است و حجم داده آزمون و شیوه ارزیابی در آن مشخص نشده است.

جدول مقایسه ای این پیاده سازی ها در جدول ۴-۲ آمده است.

جدول ۴-۲ تحلیل گرهای کلمه فارسی

مقاله	رویکرد	وابستگی به بافت	مستقل از متن	نوع ابزار	منبع واژگان	از زایا	از پایتون	معایب
مگردمیان، ۲۰۰۰	تصrif	مستقل از متن	مبدل	پروژه شیراز	۷ مگاباتی متن خبری	دقت ۹۵٪ بر روی پیکره	پوشش کامل زبان رسمی	یکپارچه بودن نیاز به لاتین سازی و روودی و فارسی سازی خروجی
شمسفرد، ۲۰۱۰	تصrif و صرف جزئی	مستقل از متن	مبدل	اوژگان زایا به واژگان	برای ۶۰۰ کلمه فراخوانی ۹۸٪ و درستی ۹۳٪	مبتنی بر بافت بودن	مبتنی بر بافت بودن	* خلط اشتقاق و تصریف، * یکپارچه نبودن
مواجی، ۱۳۹۰	تصrif و صرف جزئی	مستقل از متن	مبدل	واژگان زایا	دقت ۹۵٪	واژگان زایا	پایتون	* خلط اشتقاق و تصریف، * استفاده نکردن از مبدل و کند بودن تحلیل صرفی در نسبت با شمس فرد
روحانیان، ۱۳۹۳	تصrif	مستقل از متن	مبدل	واژگان زایا	۹۵ درصد	یکپارچه بودن	عدم پوشش کامل زبان	

۲-۳-۲- فارسی غیر رسمی

مگردمیان (۲۰۰۸؛ ۲۰۰۶) برای تحلیل تصریفی فارسی غیر رسمی سعی داشته همان ساختار مبدل حالت متناهی ای را که برای فارسی رسمی استفاده کرده (مگردمیان، ۲۰۰۰)، توسعه دهد. برای این منظور تلاش کرده است تا ساختار تصریفی غیر رسمی را مانند ساختار کلمات رسمی تعریف کند. او برای این منظور داده های فراوانی را از وبلاگ های فارسی مختلف در یک پیکره جمع آوری کرده و سعی کرده با ارائه نمونه هایی از این پیکره، ساختار کلمات را تعریف کند. او سعی کرده ساختارهای تغییر آوایی رسمی به غیر رسمی را نیز بررسی کند و قاعده های این تغییرات را شناسایی کند. در موارد بسیاری

ساختار کلمات را به درستی تعریف کرده است. اما همه ساختارها را پوشش نداده و نتیجه پژوهش او نیز منجر به ساخت ابزار (یا ارتقاء^۱ تحلیل گر رسمی قبلی) نشده است.

تازه‌جانی و بحرانی (۱۳۹۲) نیز پژوهش محدودی تنها بر روی فعل غیر رسمی انجام داده‌اند. این پژوهش شامل ابزاری که با کمک پایتون پیاده سازی شده نیز می‌شود و علاوه بر تصریف فعل، معادل رسمی آن را تولید می‌کند. دقت این ابزار به برای افعال رسمی ۸۸٪ و برای روی افعال غیر رسمی ۸۵٪ گزارش شده است (۱۰۰ فعل آزمون).

آرمنی و شمس‌فرد (۱۳۸۹) با این استدلال که بسیاری از قواعد پیچیده تصریفی برای فارسی غیر رسمی بلااستفاده است و راهاندازی تحلیل گر کامل جزئیات زیادی لازم دارد، به جای استفاده از یک تحلیل گر تصریفی کامل، از یک ریشه‌یاب استفاده کرده‌اند. ریشه‌یابی که آنها برای این منظور استفاده کردند ریشه‌یاب مبتنی بر واژگان^۲ شریفلو و شمس‌فرد (۲۰۰۸) است. تک‌واژه‌های تصریفی غیر رسمی محدودی به این ریشه‌یاب افزوده‌اند^۳. روش کار آن به این صورت است که تک‌واژه‌های لیست شده را شناسایی و حذف می‌کند و تا جایی حذف تک‌واژه‌ها را ادامه می‌دهد که به ریشه معادل در واژگان رسمی و یا واژگان غیر رسمی 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، کلماتی که در قاعده تصریفشان ابهام وجود دارد، اولویت‌بندی می‌کند. دقت گزارش شده برای این ریشه‌یاب ۹۳ درصد است بر روی ۱۰۰ کلمه داده آزمون است. جدول مروری این پژوهش‌ها در جدول ۵-۲ آمده است.

^۱ Upgrade

^۲ تنها ارادت جمع و «ی» نکره را در اسامی و شناسه‌ها، خمایر مفعولی متصل و پیشوندهای فعلی مانند «نمی»، «می»، «ب» و «ن» را در افعال شناسایی می‌کند.

جدول ۵-۲ پردازش‌گرهای فارسی غیر رسمی

مقاله	رویکرد	وابستگی به بافت	نوع ابزار	منبع واژگان	ارزیابی	مزایا	معایب
مگردو میان، ۲۰۰۸	تصrif	-	-	-	-	پوشش نظری نسبی به عنوان اولین پژوهش در این زمینه	فقدان تولید ابزار یا ارتقاء تحلیل‌گر رسمی
تازه‌جانی، ۱۳۹۲	تصrif	مستقل از متن	پایتون	نامشخص	دقت٪۸۵	اولین کوشش برای تحلیل فعل غیر رسمی	* محدود بودن به فعل استفاده نکردن از مدل و کند ودن
آرمین، ۱۳۸۹	مکاشفه‌ای (ریشه‌یاب)	مبتنی بر بافت	نامشخص	بدون واژگان	دقت٪۹۵	سیک و سریع بودن سنتی بر بافت بودن	* تنها مناسب برای کاربردهای خاص * ارائه نکردن تحلیل کامل و قیقی

۴-۲- جمع‌بندی

در زمینه تحلیل تصrifی فارسی رسمی، کارهایی انجام پذیرفته که دو نمونه از کارهای مهم آن در این پژوهش بررسی شد. در زمینه تحلیل تصrifی غیر رسمی فارسی، غیر از پژوهش تحلیلی مگردو میان در این زمینه کار دیگری صورت نگرفته است. حوزه تصrif غیر رسمی را می‌توان در سه بخش بررسی کرد؛ اول، ساختمان تصrifی غیر رسمی فارسی به طور کامل بررسی شود و ساختارهای آن شناسایی شود. دوم، قواعد تغییرات آوایی که در سطح واژگان کلمه را تغییر می‌دهد و ممکن است کلمه خارج از واژگان تولید کند، شناسایی شود. سوم، منابع واژگانی مختص به فارسی غیر رسمی جمع آوری شود.

فصل سوم

روش پیشنهادی؛ ساختار

۱-۳- مقدمه

على رغم تغییرات آوایی و یکدست نبودن قوانین نگارشی در بسیاری از موارد، فارسی غیر رسمی در ساختار تصریفی خود از قواعدی تبعیت می‌کند. این گونه فارسی جدا از پذیرفتن ریشه متفاوت از رسمی در ساختار کلمه، وندها و واژه‌بسته‌های تصریفی تغییر یافته، تعداد تک‌واژه‌های تصریفی بیشتری نیز در مقایسه با فارسی رسمی می‌پذیرد و حتی برخی از این تک‌واژها مختص فارسی غیر رسمی است (مثل تک‌واژه‌ای محاوره). بنابراین تحلیل قاعده‌مند این گونه زبانی و در نظر گرفتن آن به عنوان یک گونه مستقل که نیاز به تحلیل خاص به خود دارد و حتی به عنوان گونه‌ای که گونه رسمی را به عنوان بخشی از خود در بر می‌گیرد، لازم است. همچنین تغییرات آوایی نیز بر محور قاعده‌هایی رخ می‌دهد که قابل شناسایی و پوشش است.

۲-۳- چالش‌های پردازش فارسی

مواردی که در این قسمت پوشش داده شده است، برخی مربوط به فارسی رسمی، برخی مربوط به فارسی غیر رسمی و برخی دیگر مربوط به هر دو گونه زبانی است. سه بخش ابتدائی همانطور که جورافسکی (۲۰۰۸) دسته‌بندی کرده است مربوط به سه قسمت مربوط به ساخت یک تحلیل‌گر دو سطحی است که برای زبان فارسی رسمی و غیر رسمی یکسان است. بخش‌های دیگر نیز گرچه مربوط به هر دو گونه است اما غالباً برای پوشش و شناسایی فارسی غیر رسمی بکار می‌روند.

۱-۴- فراهم کردن واژگان

فراهم کردن واژگان رسمی با توجه به منابع موجود کار دشواری نیست اما این واژگان با توجه به کاربردهای خاص تقسیم‌بندی شده‌اند و برای سازگار شدن با ساختار تحلیل‌گر تصریفی‌ای که هم کلمه رسمی و هم غیر رسمی را تحلیل کند، نیازمند تغییر است. با توجه به تغییرات واکه‌های پایانی می‌بایست به نحوی حروف پایانی آنها علامت گذاری شود تا امکان اعمال تغییرات مناسب را داشته باشد. اسم‌های عربی‌ای که در فارسی استفاده می‌شود، وندهای جمع مختص به خود را دارد. شناسایی این که کدام کلمه عربی کدام وند جمع را می‌پذیرد و کدام را نه می‌بایست انجام شود. جمع‌های عربی مکسر که بدون گرفتن وند جمع تبدیل به اسم جمع می‌شوند می‌بایست شناسایی شود. اسم‌های فارسی هم به همین صورت دارای دو نوع جمع جاندار و جمع معمولی است که جانداران می‌بایست در واژگان شناسایی شود.

برای واژگان غیر رسمی نیز منبعی وجود ندارد و می‌بایست با توجه به پیکره غیر رسمی و یا معادل سازی از روی واژگان رسمی، برای آن واژه استخراج کرد. با توجه به گستردگی کاربردی این گونه می‌بایست پیکره‌ای از زیر سیاق‌های آن تشکیل شود و همه واژگان بکار رفته فارسی و غیر فارسی پُرسامد استخراج شود. کلمات غیر رسمی به دو بخش تقسیم می‌شود؛ کلماتی که واحد واژگانی می‌سازند و کلماتی که در نگارش متفاوت از نگارش رسمی نوشته می‌شوند.

کلمات غیر رسمی که واحد واژگانی مستقل می‌سازند، با تغییر آوایی مخصوصی، به یک واحد واژگانی غیر رسمی تبدیل می‌شوند. کلماتی که در نگارش متفاوت از رسمی‌اند واحد واژگانی نمی‌سازند و برای پردازش آنها باید از روش‌های دیگری نظریه الگوریتم آوایی استفاده کرد.

۲-۲-۳- قواعد تصريفی

واژگان به همراه وندها و واژه‌بست‌های تصريفی طبق قواعد تصريفی زبان فارسي می‌توانند کلمه تولید کنند. اين قواعد برای فارسي رسمی می‌بايست تمام امكانات کلمه فارسي رسمی را پوشش دهد. شناسايي عناصر سازنده کلمه، ساختار مناسب ترکيب اين عناصر سازنده برای توليد و شناسايي کلمه رسمی فارسي ضروري است. کلمات فارسي غير رسمی از طرف ديگر علاوه بر ساختمان کلمه رسمی (واژگان و قاعده سازنده) که آن را كاملا به کار می‌گيرد، از اجزای غير رسمی و يا ساختمان غير رسمی که متفاوت از ساختمان کلمه رسمی است، استفاده می‌كند. ساختمان کلمه غير رسمی، وند و واژه‌بست‌های بيشتری می‌پذيرد و ساختمان پيچide تری تولید می‌كند.

۳-۲-۳- ساختار جامع و مانع

برای پياده‌سازی رايانيه‌اي قواعد کلمه فارسي غير رسمی نياز به قواعد جامع و پوشش همه حالات و امكانات تصريف آنهاست. در پژوهش مگردو ميان (۲۰۰۸)، نمونه‌هایي از زبان فارسي غير رسمی ذکر شده اما ساختارهای جامع برای آنها تعریف نشده است. مثلاً اينکه واژه‌بست هم در کجای کلمه می‌تواند قرار بگيرد و در همنشيني با چه وندها و واژه‌بست‌های ديگري می‌تواند قرار بگيرد روش نشده است. ساختار جامع برای فعل‌ها و کلمات غير فعلی می‌بايست تمام حالت‌هاي ممکن ساخت کلمه رسمی و غير رسمی را پوشش دهد. در مورد کلمه رسمی پيش‌تر اين ساختار پوشش داده شده است (مگردو ميان، ۲۰۰۰؛ اسلامي و همكاران، ۱۳۸۳؛ گيوي و همكاران، ۱۳۹۱). اما در مورد فارسي غير رسمی تنها دو پژوهش توسط مگردو ميان صورت گرفته که منجر به ساخت ابزار تحليل‌گر تصريفی نشده است و ساختارهای جامع کلمه فارسي غير رسمی نيز در آنها تعریف نشده است.

از طرف ديگر ساختارها ممکن است دامنه‌اي گسترده‌تر از چيزی را که آن ساخت یا برچسب لازم دارد تولید کند. به طور مثال کلمه **رفiqshonim** شامل اسم به همراه واژه‌بست شخصی سوم شخص جمع و واژه‌بست ربطی اول شخص جمع است. با توجه به جايگاه واژه‌بست‌های محاوره، قاعده **اسم+وشخصی ۶+وربطی ۲+هم** نيز می‌تواند تولید شود که در ساختار اسمی وجود ندارد و بـی معنی است. بنابراین در کنار جامع بودن ساختارها (بالا بردن فراخوانی^۱، محدوديت‌هایي نيز بـاید بر آنها اعمال کرد تا درستی^۲ (صحت) قواعد نيز حفظ شود.

^۱ Recall

^۲ Precision

۴-۳-۴- فاصله‌گذاری

فاصله، نیم‌فاصله و اتصال، سه حالت مرز نما در بین کلمات و یا اجزای کلمات هستند. این سه منشاء اصلی نگارش‌های متنوع فارسی و مشکل‌ساز در شناسایی واحدهای یک‌دست فارسی است. این مشکل هم در مرحله جداسازی و هم در مرحله تحلیل تصریفی کلمه رخ می‌دهد.

اجزای قطعه‌های چند واحدی^۱ و واحدهای چند قطعه‌ای^۲ در فارسی رسمی هم با فاصله و هم با نیم‌فاصله و گاهی ترکیبی از فاصله و اتصال یا نیم‌فاصله و اتصال نوشته می‌شود (جدول ۱-۳ و جدول ۲-۳). بنابراین در نظر گرفتن آن به عنوان یک واحد کل یا چند واحد به سادگی امکان پذیر نیست.

از انواع دیگر قطعه‌های چند واحدی افعال و اسم و صفت‌های غیر بسیط هستند که وند و واژه‌بست تصریفی می‌پذیرند. در زبان رسمی و غیر رسمی این وندها و واژه‌بستها با فاصله، نیم‌فاصله و اتصال نوشته می‌شوند (جدول ۲-۳).

جدول ۱-۳ نگارش‌های متنوع یک واحد چند قطعه‌ای

حروف ربط گروهی	صفت مرکب
آن چنان که	آنچنان ساز
آن چنانکه	آنچنان ساز
آنچنانکه	آنچنان ساز
آنچنان که	

جدول ۲-۳ نگارش‌های متنوع یک قطعه چند واحدی

فعل		صفت	اسم	اسم مشتق مرکب	صفت مشتق
می‌بروی	خوردمشون	زیبایی‌ش	کتابها	دل‌پیچه	فهرستوار
می‌روی	خوردم‌شون	زیبایی‌ش	کتاب‌ها	دل‌پیچه	فهرست‌وار
می‌روی	خوردمشون	زیبایی‌ش	کتاب‌ها	دل‌پیچه	فهرستوار

۴-۵-۲- 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 غیر رسمی رخ می‌دهد.

^۱ MUT – Multi Units Token

^۲ MTU – Multi Tokens Unit

همزه یکی از حروف الفبای فارسی به حساب می‌آید که به صورت مستقل و چسبان در استاندارد یونیکد، دو نویسه^۱ مجزا به حساب می‌آید (ء و ئ)، اما در زبان فارسی به شکل مستقل یک حرف به حساب می‌آید. 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 (جدول ۳-۳).

جدول ۳-۳ عدم قطعیت همزه در فارسی

مثال	نوع جایگزینی	
زمینه خشونت	زمینه خشونت	جایگزینی همزه با حرف اصلی
مسوول / مسئول	مسوول	جایگزینی همزه با همزه
املا، اعضا	إملا، أعضاء	جایگزینی همزه با حرف اصلی
انشا	إنشاء	حذف همزه مستقل در پایان
املایی	املائی	جایگزینی همزه با حرف اصلی

در فارسی غیر رسمی حروف‌های هم‌صدا با احتمال بالایی جایگزین یکدیگر می‌شوند (جدول ۴-۳).

جدول ۴-۳ جایگزینی حروف هم‌صدا در فارسی غیر رسمی

غلط نویسی	شكل صحیح	حروف هم‌صدا
گزشته، ذربان، خبان، ضرف	گذشته، ضربان، زبان، ظرف	ذ، ز، ض، ظ
ترف، طماس	طرف، تماس	ت، ط
صالح، سابون، صبت	ساحل، صابون، ثبت	س، ص، ث
غفل، قذا	قفل، غذا	غ، ق
هوله، هستی	حوله، هستی	ح، ه
مسئول / مسیول، عرزش، هوایی	مسئول، ارزش، هوایی	ؤ، ا، ئ، ئی، ع، ی

حرف ھ در نقش واکه در صورتی که در پایان یک تکواز قرار بگیرد، در بسیاری از موارد، در هنگام اتصال به تکوازی دیگر حذف می‌شود (جدول ۵-۳).

^۱ Character

جدول ۵-۳ حذف واکه پایانی ۵ در فارسی رسمی

جمع با ان	جمع	پیوسته	جدا	قسم کلمه
	مفرد	بشكل	به شکل	اسم
ستارگان	ستاره	بطوریکه	به طوری که	حرف ربط گروهی
آیندگان	آینده	یک بیک	یک به یک	شماره / قید
آزادگان	آزاده	بما	به ما	ضمیر شخصی
بافندگان	بافنده	بکجا	به کجا	ضمیر پرسشی
بندگان	بنده	بکسی	به کسی	ضمیر مبهم

حرف ۵ به عنوان واکه در پایان تک واژه در یک ساخت غیر رسمی در هنگام اتصال به تک واژی دیگر حذف می شود (جدول ۶-۳).

جدول ۶-۳ حذف واکه پایانی ۵ در فارسی غیر رسمی

متصل	جزا	قسم کلمه
خستس	خسته	صفت+وربظی ۳
باقیماندس	باقیمانده	اسم+وربظی ۳
این همس	این همه	ض. اشاره+وربظی ۳
همشون	همه	ض. مبهم+وشخصی ۶
پنیمون	پنبه	اسم+وشخصی ۵
برندش	برنده	صفت+وشخصی ۳

واکه ۵ برای بازنمایی کسره اضافه (در فارسی غیر رسمی) در پایان کلمه منتهی به همخوان و یا واکه ی گاهی استفاده می شود (جدول ۴-۵).

۶-۲-۳ - کلمات به هم چسبیده

این خطای املایی رایجی است که شناسایی و تحلیل تصrifی کلمات را با مشکل مواجه می سازد. در هر دو گونه رسمی و غیر رسمی رخ می دهد. کلماتی که منتهی به یکی از حروف و، ر، ز، ڙ، ڏ، ڌ باشد با احتمال بیشتری ممکن است به

کلمه بعد از خود به شکل متصل حروف چینی شود. علت این امر یک نوع خطای نگرشی است که در هنگام سریع حروف چینی کردن رخ می‌دهد. گرچه این نوع اتصال بخش زیادی از این خطا را پوشش اما شامل همه آن نمی‌شود. برخی حرف‌های اضافه و ضمایر نیز تحت تاثیر این اتصال قرار می‌گیرند و گاهی حروف واکه خود را در محل این اتصال از دست می‌دهند. مانند ازین \leftarrow از+این، ترا \leftarrow تو+را و موارد مشابه.

۷-۲-۳- خطای املایی

خطای املایی یکی از مشکلات اجتناب ناپذیر در همه پردازش‌های متنی است. در فارسی رسمی و به خصوص فارسی غیر رسمی چنین مسئله‌ای می‌تواند به صورت جدی مشکل‌ساز شود. البته بخش عمده‌ای از این خطاهای بنا بر دسته‌بندی کوکیک (۱۹۹۲) که خطاهای را به سه دسته حروف چینی، شناختی و آوایی تقسیم کرده است، منشا آوایی و شناختی دارد. در این صورت اگر برای سایر مشکلات این بخش (چالش‌های پردازش فارسی) چاره‌جویی شود، بخش عمده این خطاهای که آوایی و نگرشی هستند برطرف می‌شود. اما دسته دیگر خطاهای یعنی خطاهای ناشی از حروف چینی نیز در بعضی موارد می‌تواند پردازش فارسی را دچار اخلال کند.

۳-۱- پیکره فارسی معاصر

برای انجام این پژوهش هم نیاز به داده‌هایی برای بررسی گونه زبان غیر رسمی و رسمی وجود داشت و هم برای ارزیابی نهایی نیاز به داده مناسب بود. بنابراین سعی شد تمام زیر سیاق‌های فارسی غیر رسمی ابتدا شناسایی شود و سپس برای هر کدام متونی تهیه شود و سپس برای انجام پژوهش آماده سازی و در صورت نیاز برچسب‌گذاری شود. حاصل کار داده‌های زیر سیاق‌های فارسی غیر رسمی است که طبق جدول ۷-۳ جمع آوری شده است. سعی شده تا این داده‌ها پوشش تمام زیر سیاق‌های زبان فارسی غیر رسمی باشد. مجموعه پیکره جمع آوری شده نزدیک به پنجاه هزار قطعه است که از آن حدود سی درصد برای ارزیابی کنار گذاشته شده است. از مابقی این داده‌ها برای بررسی ساخته‌های تصویری، آمارگیری از میزان توزیع ساخته‌های مختلف، بررسی و استخراج ساختمان‌های غیر رسمی و استخراج واژگان غیر رسمی استفاده شده است. تمام کلمات غیر رسمی این پیکره به شکل دستی معادل نویسی شده است و لمای آن نیز استخراج شده است. ساختار پیکره‌ای که به طور متوازن از این زیر سیاق‌ها (که همگی مربوط به زبان فارسی غیر رسمی معاصر است) استفاده می‌کند از نوع پیکره‌ی نمونه‌گیری شده^۱ است و نمونه‌های هر سیاق زبان فارسی غیر رسمی به شکل تصادفی^۲ هم در مرحله جمع آوری اولیه و هم در مرحله گزینش از هر سیاق انتخاب شده است (مکانی و هارדי، ۲۰۱۱).

^۱ Sampling corpus

^۲ Random sampling

جدول ۷-۳ پیکره داده‌های بکاررفته برای ارزیابی

شماره	سیاق	موجود در پیکره
۱	گفت و گوی غیر خیالی	گفت و گوی عادی (خودانگیز)
۲		گفت و گوی تلفنی
۳		✓ مصاحبه
۴		منظاره
۵		صحنه دادگاه
۶	گفت و گوی خیالی	✓ رمان (اصل فارسی)
۷		✓ نمایشنامه (اصل فارسی)
۸		✓ فیلم‌نامه (اصل فارسی)
۹		✓ زیرنویس (ترجمه به فارسی)
۱۰		✓ زیرنویس (اصل فارسی)
۱۱	سخنرانی	✓ نامه‌ی شخصی
۱۲		✓ خاطرات روزانه
۱۳		✓ شعر محاوره
۱۴	شوندگی	✓ از پیش آماده شده
۱۵		✓ خودانگیز
۱۶	شوی رادیو و تلویزیونی	✓ اجرا از روی متن
۱۷		اجرای خودانگیز یا نیمه خودانگیز
۱۸	وبلاگ	✓ پست‌ها
۱۹		✓ پاسخ‌ها
۲۰	شبکه‌های اجتماعی	✓ پست‌ها
۲۱		✓ پاسخ‌ها
۲۲	پیام‌رسان تلفنی	✓ مکالمه‌ی شخصی
۲۳		✓ مکالمه‌ی گروهی
۲۴		✓ گروه اطلاع رسانی
۲۵	نظر کاربر فضای مجازی	✓ مصرف کننده‌ی کالا یا خدمات
۲۶		✓ مقالات خبری و متفرقه

بنابراین سعی شده است تا هم تنوع گونه و سیاق و هم نسبت متون حفظ شود و با این حساب خطای سوءگیری^۱ و خطای تصادفی^۲ کنترل و به حداقل برسد (بایبر، ۱۹۹۳-الف؛ بایبر، ۱۹۹۳-ب). البته انتخاب تصادفی و یا متنوع از متون باعث نشده تا متن‌های انتخاب شده آنقدر کوتاه باشند که انسجام نداشته باشند. در واقع سعی شده از هر بخش متنی به طور

تصادفی انتخاب شود اما با در نظر داشتن این نکته که متن انتخابی کامل و منسجم باشد و زنجیره واژگانی^۳ طولانی باشد (منینگ و شوتز ۱۹۹۹). این نکته با توجه به حجم اولیه کم متون سیاق‌ها تا جایی که امكان پذیر بوده رعایت شده است. اگرچه هدف اولیه از جمع‌آوری این متون استخراج کلمات و یا استخراج ساختار تصrifی کلمات بوده است اما رعایت توازن ایجاب می‌کند که تمامی موارد ذکر شده برای یک پیکره‌ی اسنپ‌شات^۴ که هم دارای توازن^۵ و هم دارای ویژگی نمایندگی^۶ زبان فارسی غیر رسمی باشد در طراحی این پیکره رعایت شود (مکانی و هارדי، ۲۰۱۱).

مدیوم بعضی از این سیاق‌ها گفتار و بعضی دیگر نوشتار است. آنها یک گفتاری هستند به صورت پیاده شده به متن در پیکره گنجانده شده است اما بدون علائم گفتاری نظیر توقف و یا گویش و غیره به کار رفته و تنها گوینده مشخص شده است.

جملات انتخاب شده برای ارزیابی نیز به همین صورت از بخش ارزیابی این پیکره انتخاب شده است. یعنی از هر قسمت یک یا دو جمله به صورت تصادفی انتخاب شده است به نوعی که همه معیارهای ذکر شده در جمع‌آوری پیکره نیز در این مرحله رعایت شود.

۱-۳- چارچوب پردازش فارسی

چارچوب پیشنهادی امکان پردازش کلمه فارسی رسمی و غیر رسمی را توسط یک مبدل حالت متناهی امکان‌پذیر می‌سازد. در این قسمت ابتدا واژگان رسمی و غیر رسمی استفاده شده در این تحلیل‌گر شرح داده می‌شود، بعد از آن نحوه و شکل ورودی تحلیل‌گر مشخص می‌گردد، سپس نوع دسته‌بندی یا قسم کلماتی که باید انتظار داشته باشیم این تحلیل‌گر تولید کند، تشریح می‌شود. بعد از آن ساختار تصrifی کلمات به طور اجمالی توضیح داده می‌شود (توضیح کامل این ساختارها در فصل چهارم و فصل پنجم گنجانده شده است). سپس ساختار کلی تحلیل‌گر و مبدل‌های ثانویه توضیح داده می‌شود (ساختار تفصیلی این بخش در فصل ششم توضیح داده می‌شود).

۱-۱-۳- کلمه فارسی قابل پذیرش برای تحلیل‌گر

ورودی تحلیل‌گر تصrifی یک کلمه است. این کلمه برای ساده‌تر شدن مرحله جداسازی^۷ کلمه‌ای تک قطعه‌ای و نه چند قطعه‌ای^۸ در نظر گرفته شده است. تنها قسمت‌های وابسته به هسته مثل وندها و واژه‌بستها و چند مورد پیشوند می‌توانند به هسته متصل شوند. فعل‌های مرکب و کلمات چند قطعه‌ای را می‌توان در پس پردازش، بعد از تحلیل تصrifی هر قسمت،

^۷ Lexical chain

^۸ Snapshot

^۹ Balanced

^{۱۰} Representative

^{۱۱} Tokenization

^{۱۲} MTU – Multi Tokens Unit

انجام داد. فعل‌های ساده و فعل‌های پیشوندی همگی در این تحلیل گر پشتیبانی می‌شوند. فعل‌های ساده و پیشوندی‌ای که از بیش از یک کلمه ساخته می‌شوند، گرچه در قواعد گنجانده شده‌اند اما همگی از تولید نهایی کنار گذاشته شده‌اند. زیرا این افعال (ماضی بعید، ماضی بعد، مستقبل و ماضی الترامی) عملاً در مرحله جداسازی غیر قابل شناسایی‌اند. شناسایی گروه کامل این افعال نیاز به تحلیل تصريفی دارد و جداساز نمی‌تواند چنین پردازشی انجام دهد. اما در صورتی که اجزای آن توسط جداساز به درستی جدا شود و تک‌تک کلمات با تحلیل گر تصريفی، تحلیل تصريفی شود، آنگاه در پس‌پردازش می‌توان به راحتی این گروه افعال چند بخشی را شناسایی کرد.

بنابراین برای افعال واحد قابل پذیرش در تحلیل گر صرفاً یک کلمه و نه گروه کلمات است. در مورد کلمات غیر فعلی نیز مینا کلمه تک واحدی یا چند واحدی است، اما از آنجا که کلمات مرکب هم در واژگان وجود دارد پذیرفتن آنها نیز ممکن است (واحد چند قطعه‌ای) اما حالت مطمئن‌تر تحلیل تک‌کلمه و سپس شناسایی گروه کلمات در پس‌پردازش است.

۲-۱-۳- واژگان

برای واژگان رسمی از واژگان زایایی زبان فارسی (اسلامی و همکاران، ۱۳۸۳) استفاده شده است. البته دسته‌بندی، اندکی تغییر یافته است و بن فعل‌های ادبی و کم کاربرد کنار گذاشته شده است. از میان فعل‌ها، فعل‌های لازم و متعددی جداگانه برای این پژوهش جدا شده است.

جدول ۸-۳ فهرست واژگان به کار رفته در تحلیل گر

غیر رسمی	رسمی	قسم کلمه
۸۰	۷۹۱ جاندار، ۱،۹۵۱ (۲۷،۹۴۰ مکسر، ۶۵۰ سایر جمع‌های عربی)	اسم
۴۰	۱۶،۶۵۰	صفت
۲۲۷	۱،۳۷۴	قید
۲۴	۸۹	حرف ربط
۲	۲۱۳	حرف ربط گروهی
۱۷	۴۸	صفت شمارشی
۴	۲۷	حرف اضافه
-	۱۳۷	حرف اضافه گروهی
۳	۱۰	ضمیر شخصی
۸	۲۷	ضمیر یا صفت مبهم
۱۰	۲۲	ضمیر یا صفت اشاره
۴	۲۸	ضمیر یا صفت پرسشی
۵	۱،۰۵۷۷	اسم خاص مکان
۱۳	۱،۹۵۰	اسم خاص اشخاص

-	۲.۷۱۵	اسم خاص فامیل
۲۳	۱۹۳	شبیه جمله
۳	۲۹	شاخص
(۸۷ ۷۸ لازم، ۹ متعدي)	(۳۷۹ ۲۹۶ لازم، ۸۳ متعدي)	بن مضارع ساده
(۸۹ ۷۶ لازم، ۱۳ متعدي)	(۵۱۶ ۳۵۶ لازم، ۱۶۰ متعدي)	بن ماضی ساده
(۲۳ ۱۱ لازم، ۱۲ متعدي)	(۴۹ ۳۷ لازم، ۱۲ متعدي)	ماضی پیشوندی
(۲۲ ۱۶ لازم، ۶ متعدي)	(۴۹ ۴۰ لازم، ۹ متعدي)	مضارع پیشوندی

کلمه‌های غیر رسمی به دو دسته واژگان واحد و کلمات رسمی که با نگارش متفاوت نوشته می‌شوند تقسیم می‌شود. واحدهای واژگانی در واژگان قرار می‌گیرد اما آنها بایستی که نگارش متفاوت دارد در صورتی که در واژگان گنجانده نشده باشد می‌باشد. گاهی کلماتی خارج از واژگانند، آنها را نیز با کمک الگوریتم‌های آوای می‌توان شناسایی کرد (جدول ۲-۶). واحدهای واژگانی غیر رسمی با معادل سازی از نمونه رسمی‌شان در صورت امکان به دست آمده است و گاهی از پیکره جمع‌آوری شده برای گونه غیر رسمی استخراج شده است.

در مورد فعل‌های ساده، اغلب آنها از روی واژگان زایا معادل سازی شده‌اند. بن فعل‌هایی که مربوط به گونه ادبی و کهن بودند از دسته بن‌های رسمی هم حذف گردیده است. فعل‌های پیشوندی، ساده و حتی گاهی مرکب در واژگان زایا از هم متمایز نیست. این دسته‌بندی در این پژوهش انجام گرفته است. برخی از فعل‌های پیشوندی غیر رسمی از واژگان ویراستار (کاشفی و همکاران، ۱۳۸۹) و برخی دیگر از پیکره جمع‌آوری شده برای این پژوهش استخراج شده است.

کلمات غیر فعلی، از برخی از واژگان رسمی‌ای که در اختیار بوده معادل سازی شده است و برخی دیگر از پیکره جمع‌آوری شده برای این پژوهش استخراج شده است. البته دسته‌بندی همه کلمات آنطور که در بخش برچسب‌های تحلیل گر آمده تغییر کرده است.

به طور کلی ۵۳،۰۲۹ واژه غیر فعلی رسمی و ۹۹۳ فعل رسمی با استفاده از واژگان زایا و اصلاح یا افزودن از منابع دیگر به تحلیل گر افزوده شده است. با توجه به منبع واژگان رسمی و تبدیل آنها به واژگان غیر رسمی و همینطور استخراج از پیکره فارسی غیر رسمی، ۴۶۳ واژه غیر فعلی و ۲۲۱ فعل نیز به دست آمده است که در ساختمان تحلیل گر استفاده شده است. فهرست کامل این واژگان در جدول ۸-۳ آمده است.

۳-۱-۳- برچسب تولیدی تحلیل گر برای کلمات

ساختار قاعده‌ای که تحلیل گر تولید می‌کند حاوی قسم ریشه کلمه در سمت راست همراه با علامت مساوی (=) است که بعد از آن ریشه کلمه قرار می‌گیرد. در مورد افعال این برچسب دقیقاً نشان دهنده برچسب کل کلمه است اما در مورد سایر کلمات همیشه اینطور نیست زیرا ممکن است به طور مثال یک اسم با پذیرفتن واژه‌بست ربطی به مسنده تبدیل شده باشد، یا یک صفت با اشتقاء صفر به اسم تبدیل شده و در ساختمان اسم تولید شده باشد.

همینطور در مواردی که کلمه با توجه به جایگاهش در جمله می‌تواند برچسب دیگر بپذیرد اما برچسب اصلی آن با توجه به دسته‌بندی در واژگان صورت گرفته است. این دسته‌بندی تا جای ممکن سعی شده که همه برچسب‌های اصلی ریشه‌ها را پوشش دهد اما از طرف دیگر سعی شده نقش‌هایی که کمتر توسط آن کلمه ایفا می‌شود و یا نقش اصلی آن نیست، در این برچسب‌ها گنجانده نشود. مثلاً صفت‌ها یا اسم‌ها که در موقعیت‌هایی می‌توانند نقش قید (مشترک) داشته باشند اما از آنجاییکه این نقش اصلی آنها نیست، این برچسب برای اسمی و صفت‌ها استفاده نشده است.

۳-۱-۴- قواعد و ساختمان تحلیل گر تصریفی

قواعد سازنده کلمات فارسی معاصر اعم از فعل و غیر فعل در این قسمت قرار دارد.

۳-۱-۴-۱- تصrif فعل

ساختار افعال سعی شده به شکل کامل پوشش داده شود. این به این معنی است که تمامی ساختهای فعل رسمی و غیر رسمی ساده و پیشوندی که در واقع می‌توان عنوان فعل فارسی امروزین به آن داد به همراه بن فعل‌ها در این ساختار گنجانده شده است. فعل‌ها همگی فعل ساده و یا پیشوندی هستند. فعل‌های چند بخشی (ماضی بعید، بعد، ماضی التزامی و مستقبل) در این ساختمان پوشش داده شده‌اند اما به این دلیل که جداسازی آنها در مرحلی پیش‌پردازش کار دشواری است و نیاز به تحلیل تصریفی دارد از ساخت نهایی کنار گذاشته شده‌اند و شناسایی آنها می‌باشد در پس‌پردازش؛ بعد از شناسایی تک‌تک کلمات در تحلیل گر تصریفی، صورت پذیرد. در فصل چهارم (ساختمان تصریفی فعل) این ساختار به طور کامل از لحاظ نظری توضیح داده شده است و جزئیات استثنائات آن نیز به تفصیل بیان شده است. حتی پیاده سازی آن نیز به طور کامل در پیوست در بخش مربوطه شرح داده شده است.

۳-۱-۴-۲- تصrif کلمه غیر فعلی

برای کلمات غیر فعلی ساختمان جامع و مانعی طراحی شده است. ساختمان اصلی برای اسم‌ها طراحی شده است و سایر کلمات این بخش نیز به طور موردنی یا نسبتاً کامل از ساختمان اسمی در ساخت تصریفی خود استفاده می‌کنند. تغییرات واکه‌ای در مرز تک‌واژها، وندها و واژه‌بست‌ها نیز در این بخش پیاده شده است. به طور مثال واکه ۵ در پایان اسم جانداری که وند جمع آن می‌پذیرد حذف شده و تبدیل به حرف گ می‌شود. مثل تبدیل ستاره به ستارگان. یا حذف ۵ در پایان خسته و قبل از واژه‌بست سوم شخص مفرد س؛ خستس. یا حذف ۵ در پایان همه و قبل از واژه‌بست شخصی سوم شخص جمع شون؛ همشون. در فصل پنجم (تحلیل تصریفی کلمه غیر فعلی) این ساختمان به تفصیل تشریح و تبیین می‌گردد.

۳-۱-۴-۳- قواعد نگارشی

برای کلمات رسمی قواعد نگارشی بر اساس فرهنگ املایی زبان فارسی (صادقی و مقدم، ۱۳۸۵) طراحی شده است. اما برای کلمات غیر رسمی این قواعد با توجه به قواعد تصریفی به دست آمده است. این قواعد نگارشی در فصل‌های مربوط

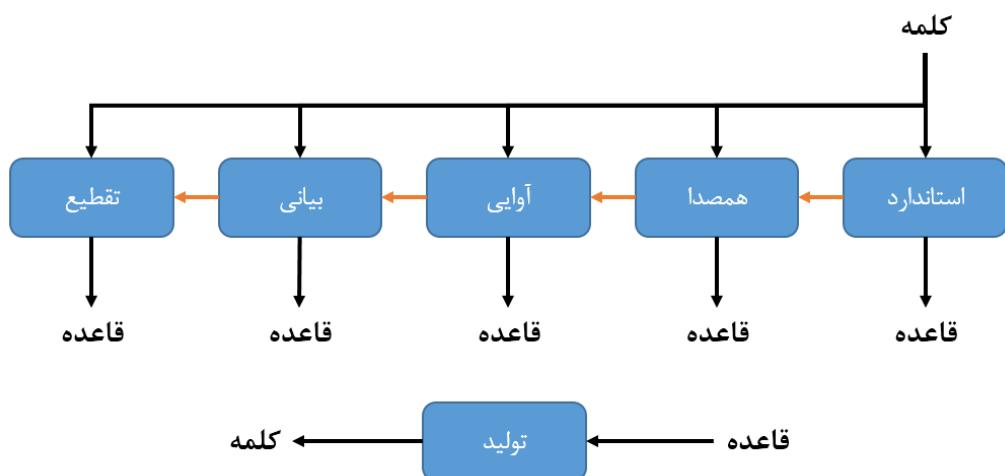
به قواعد تصريفی در هر بخش مشخص شده است. برای مشخص شدن نوع اتصال یا فاصله گرفتن تکوازهای یک کلمه در نسبت با هم از نویسه‌های ویژه‌ای استفاده شده است که در هر بخش نوع اتصال‌ها را مشخص می‌کند. این نویسه‌ها و کارکرد هر یک در جدول ۹-۳ توضیح داده شده است.

جدول ۹-۳ مرznماهای به کار رفته در قواعد نگارشی

نویسه	کارکرد
*	دو تکواز را بدون هیچ فاصله‌ای به هم متصل می‌کند.
×	دو تکواز را با نیم‌فاصله به هم متصل می‌کند.
∧	بین دو تکواز یک فاصل کامل قرار می‌دهد.
¤	<p>در بین تکواز منتهی به واکه ۵ و واژه‌بست شخصی قرار می‌گیرد. یکبار واکه ۵ قبل از خود را حذف می‌کند و بدون فاصله واژه‌بست شخصی را به تکواز قبل از خود می‌چسباند و یکبار دیگر بدون حذف، خود تبدیل به نیم‌فاصله می‌شود.</p> <p>مثل، همه+شون: همشون، همهشون (مختص کلمات غیر فعلی).</p> <p>در بین تکواز منتهی به واکه ۵ و واژه‌بست ربطی سوم شخص مفرد به قرار می‌گیرد. یکبار واکه ۵ قبل از خود را حذف می‌کند و واژه‌بست را بدون هیچ فاصله‌ای به تکواز قبلی می‌چسباند و یکبار دیگر بدون حذف، خود تبدیل به نیم‌فاصله می‌شود. مثل، خسته+س: خستس، خستهس (مختص کلمات غیر فعلی).</p> <p>در بین تکواز منتهی به واکه ۵ و وند جمع جانداران قرار می‌گیرد و واکه ۵ را حذف می‌کند حرف گ را جایگزین آن می‌کند. مثل، ستاره+ان: ستارگان، بنده+ان: بندگان (مختص کلمات غیر فعلی).</p>

۳-۱-۴-۴- ساختمان کلی تحلیل گر

از آنجایی که فارسی امروزین (رسمی و غیر رسمی) زوایای متفاوتی دارد، برای سهولت و مدیریت بر تحلیل کلمات آن می‌پایست از چند مبدل متنوع استفاده کرد (شکل ۱-۳).



شکل ۱-۳ ساختار کلی تحلیل‌گر و میدل‌ها

قواعد تصرف سازنده کلمات رسمی و غیر رسمی در مبدل استاندارد گنجانده شده است. مبدل استاندارد تنها اختصاص به قواعد تصرفی فارسی، رسمی، و غیر رسمی دارد.

در صورتی که حروف غیر قطعی در کلمات استفاده شده باشد طوری که مبدل استاندارد خروجی صحیحی تولید نکند، مبدل هم‌صدا کلمه را تحلیل تصویری می‌کند. این مبدل همچنین واکه ۵ را که در نقش کسره اضافه در پایان کلمه قرار می‌گیرد، شناسایی می‌کند.

در صورتی که ریشه کلمه در واژگان مبدل استاندارد یا هم‌صدا نبود و یا تعییرات آوایی (جدول ۲-۶) آنچنان که در مبدل آوایی، تعریف شده است در کلمه رخدید این مبدل آن را شناسایی کرده و قاعده سازنده آن را تولید می‌کند.

میدا، بیانه، نیز برای شناسایی کلماتی، که تاکید تکرار بر روی حروف آن (خ م دهد، بکار م م ده) ود. مثا، چیزی؟

در مبدل پایانی سلسله مبدل‌های شناسایی^۹، از مبدل تقطیع استفاده می‌شود. این مبدل کلمه‌هایی که به یکدیگر چسبیده باشد، احتمالاً کلمات جدا شده، احتمالاً گانه تحلیل تصرف می‌کند.

برای تولید نیز از مبدل جدگانه‌ای استفاده شده است تا امکان ساده کردن قواعد کلمات فراهم شود. در این مبدل قاعده سازنده جمع (با توجه به پنج نوع وند جمع عربی و فارسی) ساده‌تر شده است. همینطور علامت جمع مکسر برای اسم‌های عربی، حالت غیر قطعی برای فاصله، نیم‌فاصله و اتصال، حالت غیر قطعی برای حروف ۱ و آ و اعراب کلمات در قاعده سازنده آنها حذف شده است. حالت غیر قطعی حروف ۱ و آ برای قواعد ایجاد شده است. توضیحات کامل‌تر را می‌توان در فضای مربوط به میدا. ها دنیا . ک.د.

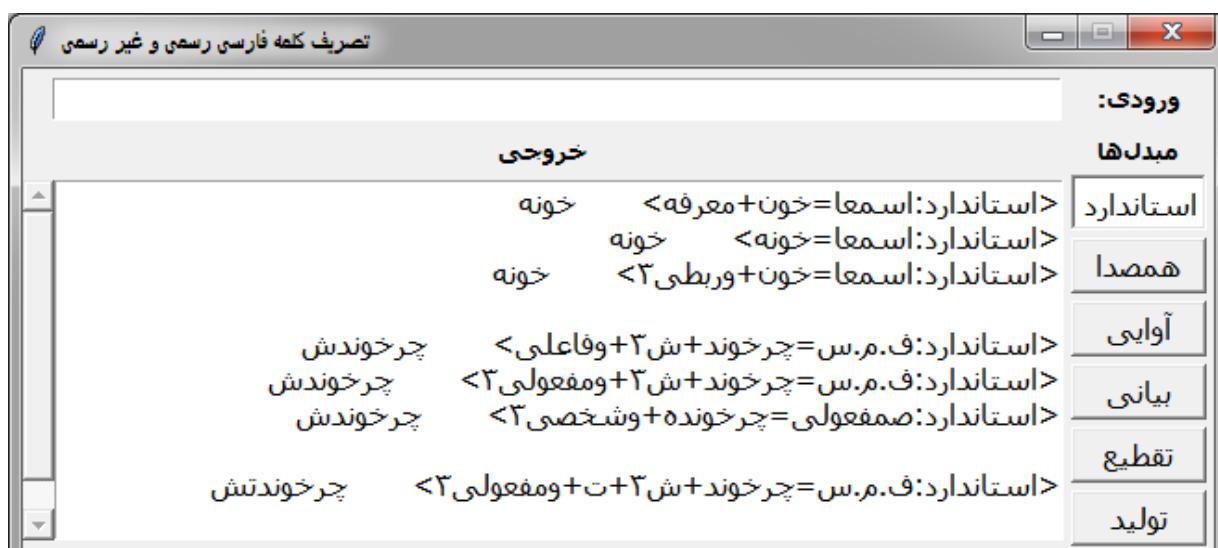
Recognition

جدای از تقسیم کار بین مبدل‌ها، دلیل دیگر جدا کردن آنها کم حجم‌تر شدن فایل هر یک بود. همه این مبدل‌ها همیشه استفاده نمی‌شود و قرار دادن همه آنها در کنار هم درون یک فایل پر حجم مفید نیست.

۳-۱-۵- ابزار پیاده‌سازی

برای پیاده‌سازی تحلیل‌گر تصیری از ابزار فوما^{۱۰} استفاده شده است که پیاده‌سازی متن-آزاد فن‌آوری زیراکس و همه زیرشاخه‌های آن برای قاعده‌مند کردن تحلیل تصیری و تحلیل کلمه است. این صورت‌بندی‌ها^{۱۱} و دستورالعمل‌ها^{۱۲} در ادبیات تحلیل تصیری به عنوان استاندارد شناخته می‌شود بنابراین امکان انتقال همه این ساختارها به ابزار دیگری که این استانداردها را پشتیبانی کند فراهم است. توضیحات مکمل در فصل دوم آمده است.

هر کدام از مبدل‌ها را جداگانه و یا با هم می‌توان استفاده کرد. برای دسترسی آزمایشی به مبدل‌ها ابزاری به زبان پایتون طراحی شده است که می‌تواند نقش رابط را بین کاربر و مبدل‌ها ایفا کند. تصویر این واسط را در شکل ۲-۳ مشاهده می‌کنید.



شکل ۲-۳ ابزار واسط کاربر و مبدل‌ها

۲-۳- نوآوری

پژوهش مگردو میان (۲۰۰۶؛ ۲۰۰۸) گرچه منتهی به ساخت تحلیل‌گر تصیری نشده و ساختمان‌های جامع و مانع کلمات غالباً شناسایی نشده است اما برای مشخص شدن مواردی که در این پژوهش برای اولین بار انجام پذیرفته، آن پژوهش مینا قرار گرفته است. توضیح و پیاده‌سازی هر قسمت برای رجوع در پایان هر کدام آمده است.

^{۱۰} Foma

^{۱۱} Formalism

^{۱۲} Script

ساخت فعل

- جایگاه تک واژه‌های محاوره در ساختمان افعال مشخص شده است و عنصر تاکید آن در تحقیق مگردویان پوشش داده نشده است (جدول ۶-۴).
- فعل‌های پیشوندی که معمولاً به شکل سرهم نوشته می‌شوند و در زبان فارسی غیر رسمی بسیار رایج است در این تحقیق پوشش داده شده است (فصل چهارم).
- فعل‌های ناقص غیر رسمی شناسایی و پوشش داده شده است (۴-۱-۴-۶-فصل چهارم).
- بن فعل‌های مضارع غیر رسمی در پذیرفتن شناسه سوم شخص مفرد رسمی و ساخت امری مفرد محدودیت دارند که این موضوع در قواعد به عنوان استثناء در نظر قرار گرفته شده است (جدول ۴-۸ و جدول ۴-۹).
- بن فعل‌های رسمی در پذیرفتن شناسه سوم شخص غیر رسمی محدودیت دارند که این موضوع در قواعد به عنوان استثناء در نظر قرار گرفته شده است (جدول ۴-۱۰).
- تعریف ساختمان مجزا برای فعل لازم و فعل متعدد.

ساخت غیر فعلی

- تک واژه تاکید در گره تک واژه‌های محاوره پوشش داده شده است (فصل پنجم).
- واژه‌بسته‌های را+هم در ساختار اسامی و سایر کلماتی که از این ساختار بهره می‌گیرند شناسایی و پوشش داده شده است (فصل پنجم ۵-۲-۷).
- ساختمان تصریفی بعد از وند معرفه تعریف شده است (فصل پنجم).
- ساختمان تصریفی برای ضمیر اشاره، مبهوم، شخصی، مشترک، شماره، پرسشی و حرف اضافه تعریف شده است (فصل پنجم).

منابع واژگانی

- جمع‌آوری ۲۲۱ بن فعل غیر رسمی ساده و پیشوندی (۳۹ لازم، ۱۸۲ متعدد و ساختمان هر یک).
- جمع‌آوری ۴۶۳ واژه غیر فعلی غیر رسمی.
- جمع‌آوری ۴۹ فعل پیشوندی رسمی.
- جداسازی فعل‌های لازم و متعدد رسمی (۲۶۴ لازم، ۷۲۹ متعدد).
- اسم‌های عربی به کار رفته در فارسی رسمی، جمع‌های مكسر (۷۹۱ مورد) و جمع‌های عربی (بیش از ۶۵۰ مورد) استخراج شده است (بخش جمع فصل پنجم).
- اسم‌های فارسی جاندار که وند جمع ان می‌پذیرد (۱،۹۵۱ مورد) استخراج شده است (بخش جمع فصل پنجم).

جمع‌آوری پیکره و داده ارزیابی

- جمع‌آوری پیکرهای نزدیک به پنجاه هزار کلمه از تمام زیرسیاق‌های شناسائی شده فارسی غیر رسمی که تمام کلمات غیر رسمی آن شناسائی شده و قسم، تصریف و معادل رسمی آنها به شکل دستی به آن افزوده شده است.
- برای آزمون نیز از بخش ارزیابی این پیکره جملاتی حاوی بیش از چهار هزار کلمه انتخاب شده است که از ۱۷۸۶ کلمه یکتا ساخته شده است. این کلمات یکتا با کمک تحلیل‌گر و بررسی / اصلاح انسانی ۳۸۹۰ تایپ تولید کرده است که مربوط به کاربردهای مختلف این کلمات در بافت‌های مختلف است. از این دادگان برای ارزیابی این تحلیل‌گر استفاده شده است.

۳-۳- راهنمای فصل‌های دیگر روش پیشنهادی

ساختمان تفصیلی تصریف فعل و غیر فعل در فصل‌های بیش رو توضیح داده شده است. نوآوری‌ها، ساختارهای جدید و استثنایات، همگی به شکل کامل در این دو فصل (فصل چهارم و فصل پنجم) و فصل بعد (فصل ششم) از آنها که مربوط با ساختمان کلی تحلیل‌گر و مبدل‌های گنجانده شده است.

۴-۳- جمع‌بندی

روش پیشنهادی برای تحلیل تصریفی فارسی معاصر ابتدا شناخت ساختار فارسی غیر رسمی از لحاظ نظری و سپس پیاده‌سازی ابزاری قاعده بنیاد با استفاده از مبدل حالت متناهی برای هر دو گونه فارسی رسمی و غیر رسمی است. منابع زبانی شامل واژگان، قواعد تصریفی و قواعد نگارشی می‌شود.

در مورد واژگان، از واژگان زایا برای بخش رسمی استفاده شده است. برای بخش غیر رسمی، معادل سازی برخی از واژگان زایا و استخراج کلمه از پیکره جمع‌آوری شده مورد استفاده قرار گرفته است.

برای قواعد تصریفی و نگارشی در بخش رسمی از واژگان زایا (اسلامی و همکاران، ۱۳۸۳) و مگردویان (۲۰۰۰) استفاده شده است و برای بخش غیر رسمی از مگردویان (۲۰۰۸) و نیز استخراج از پیکرهای که جمع‌آوری شده، استفاده شده است.

فصل چهارم

روش پیشنهادی؛ تصریف فعل

۱-۴ - مقدمه

فعل را در زبان فارسی از جهت ساختمان به چند دسته می‌توان تقسیم کرد: فعل‌های ساده، فعل‌های پیشوندی، فعل‌های مرکب، فعل‌های پیشوندی مرکب، فعل‌های ناگذر یک شخصه و عبارت‌های فعلی (حسن احمدی گیوی و حسن انوری، ۱۳۹۱). فعل‌هایی که در این تحلیل گر پوشش داده می‌شود، تنها فعل‌های ساده و فعل‌های پیشوندی است. فعل‌های مرکب از دو کلمه یا بیشتر تشکیل شده‌اند که کلمه اسمی / صفتی جداگانه و کلمه فعلی نیز به طور جداگانه قابل تحلیل تصریفی است. فعل‌های پیشوندی مرکب نیز می‌تواند به دو بخش اسمی / صفتی و فعل پیشوندی تقسیم و سپس توسط تحلیل گر تصریفی، تجزیه شود. فعل‌های ناگذر یک شخصه درست به همین ترتیب به دو بخش کلمه غیر فعلی و فعل سوم شخص مفرد گذشته تقسیم می‌شود که تحلیل تصریفی هر یک جداگانه انجام می‌پذیرد. در آخر عبارت‌های فعلی نیز با تبدیل به کلمات مستقل، قابل تصریف توسط تحلیل گر تصریفی هستند.

فعل‌های مجھول هم از دو بخش تشکیل شده‌اند؛ صفت مفعولی به همراه ساختهای مختلف فعل ساده شدن (یا گردیدن، گشتن که در معنای شدن به کار می‌روند) (حسن احمدی گیوی و حسن انوری، ۱۳۹۱). بنابراین شناسایی آن‌ها نیز به صورت بخش بخش در تحلیل گر تصریفی صورت می‌گیرد و شناسایی گروه آن در پس‌پردازش انجام می‌گیرد. فعل‌های ساده و پیشوندی ماضی التزامی، ماضی بعد، ماضی مستقبل، فعل‌هایی هستند که از بیش از یک کلمه ساخته می‌شوند. اجزای سازنده هر کلمه این افعال (ماضی ساده، مضارع ساده و صفت مفعولی) در قسمت مربوط تولید می‌شوند. برای شناسایی / تولید این افعال توسط تحلیل گر تصریفی هر دو روش قابل استفاده است. به شکل گروهی این افعال تعریف شده‌اند اما به دلیل مشکل شناسایی در سطح جداساز کنار گذاشته شده‌اند. اما فایل آنها و قواعد آنها به طور جداگانه وجود دارد و قابل استفاده است. برای افروzen آنها کافیست در مرحله کامپایل شدن مبدل‌ها خطی از دستور که آنها را فراخوانی می‌کند فعل کرد.

تفاوت فعل رسمی و غیر رسمی

نحوه ساخت فعل‌های رسمی فارسی در تحلیل گر تصریفی براساس دستور زبان فارسی (حسن احمدی گیوی و حسن انوری، ۱۳۹۱) انجام می‌پذیرد. جزئیات این ساختمان در ادامه توضیح داده شده است.

ساخت فعل‌های غیر رسمی مانند فعل‌های رسمی است. همان قواعد را همراه با تعدادی قواعد افزوده که مختص فارسی غیر رسمی است، دارد. علاوه بر ریشه فعل‌های غیر رسمی، می‌توان از ریشه فعل‌های رسمی، شناسه، واژه‌بست مفعولی و تک‌واژه‌ای عطف، تاکید و هم، در این ساخت‌ها استفاده کرد.

ساختار فعل‌های غیر رسمی با بررسی تحقیقی که به صورت ناتمام در این زمینه انجام شده (کارین مگردویان، ۲۰۰۸)، منابع مشترک فعل رسمی که ذکر شد، و بررسی پیکره جمع‌آوری شده، صورت پذیرفته است.

به طور کلی تفاوت فعل رسمی فارسی با غیر رسمی به این صورت است؛ در ساخت فعل غیر رسمی حداقل یک عنصر غیر رسمی باید وجود داشته باشد. عنصرهای غیر رسمی عبارتند از: بن غیر رسمی فعل، شناسه غیر رسمی، داشتن واژه‌بست مفعولی، داشتن تک واژه‌ای محاوره (تک واژه‌ای تاکید ۱ یا ها، واژه‌بست عطف و واژه‌بست هم به معنای هم).

به طور مثال فعل گفتیشان گرچه از بن رسمی گفت، شناسه‌ی (رسمی) و واژه‌بست شخصی رسمی شان استفاده کرده است اما دارای ساختی غیر رسمی است. زیرا فعل‌های رسمی واژه‌بست مفعولی نمی‌پذیرند (کارین مگردومنیان، ۲۰۰۸/۲۰۰). واژه‌بست مفعولی که خود از جنس واژه‌بست‌های شخصی است، این فعل را به یک فعل غیر رسمی تبدیل کرده است. یا فعل رفتش که از یک بن رسمی و شناسه سوم شخص مفرد Ø تشکیل شده و یک فعل ماضی ساده رسمی را می‌سازد اما به این دلیل که واژه‌بست فاعلی شش (محرم اسلامی و صدیقه علیزاده، ۱۳۸۸) پذیرفته یک فعل غیر رسمی به حساب می‌آید.

فعل‌های پیشوندی

ساخت این فعل‌ها تفاوتی با فعل‌های ساده ندارد و تنها عنصر پیشوندی در ابتدای فعل ساده (قبل از همه عنصرهای سازنده فعل) قرار می‌گیرد. فعل‌هایی مانند برآوردن، برچیدن، فرارسیدن، وارفتن فعل‌های پیشوندی‌ای هستند که غالباً در نوشتار رسمی و غیر رسمی به شکل سرهم نوشته می‌شود.

ساختمان این فعل‌ها مانند فعل‌های رسمی ساده است. تنها تفاوت در ساختهای امری و مضارع التزامی اتفاق می‌افتد. در این ساختها برخی از فعل‌ها بدون پیشوند ب بکار می‌روند و برخی دیگر به شکل اختیاری پیشوند ب را استفاده می‌کنند. مثل در(ب) رو، برانگیخت، بر(ب) انداز، برگزیدند، برشمارید.

ساختار فعل‌های پیشوندی غیر رسمی هم مانند فعل‌های غیر رسمی است. بن فعل‌ها از روی بن فعل‌های رسمی پیشوندی معادل سازی شده‌اند و برخی فعل‌هایی که معادل صریحی در زبان رسمی ندارند نیز به آنها اضافه شده‌اند. مانند وايسادن، پاشدن^{۱۳}، وايسوندن.

۴-۲-۴- اجزای ساختمان فعل

در ساختمان فعل‌های رسمی ساده و پیشوندی که در ادامه ساختشان توضیح داده خواهد شد، از بن‌های فعل ماضی و مضارع استفاده می‌شود. این بن‌ها از واژگان زایا (محرم اسلامی و همکاران، ۱۳۸۳)، وبراستیار (کاشفی، نصیری و کتعانی، ۱۳۸۹)، واژگان پرلکس (سگوت و والتر، ۲۰۱۰) و پیکره جمع‌آوری شده بدست آمده است. با تحلیل و بررسی تعدادی از این بن فعل‌ها، آنها یکی که قدیمی، کم‌کاربرد و یا ادبی بوده از این لیست حذف گردیده است. معیار برای حذف یا نگه داشتن واژگان، پیکره جمع‌آوری شده، جستجو در لغتنامه‌های زبان فارسی (محمد دبیر سیاقی، ۱۳۹۰؛ محمد معین، منیژه امیریان، ۱۳۸۲) و ششم زبانی نگارنده بوده است.

^{۱۳} گرچه فعل پاشدن یک فعل مرکب است اما به دلیل کاربرد زیاد و نگارش به هم پیوسته می‌توان آن را با ساختار فعل پیشوندی تحلیل تصریفی کرد.

بن فعل‌های حال و گذشته غیر رسمی از پیکره جمع آوری شده و همچنین با معادل‌سازی بن فعل‌های رسمی بدست آمده است. شناسه‌ها و واژه‌بست‌ها نیز از تحلیل بLAG‌های فارسی (کارین مگردویان، ۲۰۰۸)، واژگان زایا (محرم اسلامی و همکاران، ۱۳۸۳)، دستور زبان فارسی (حسن احمدی گیوی و حسن انوری، ۱۳۹۱) و از پیکره جمع آوری شده بدست آمده است. شناسه‌های فعل در جدول ۱-۴ و جدول ۲-۴ آمده است. واژه‌بست‌های ربطی در جدول ۳-۴، واژه‌بست‌های مفعولی در جدول ۴-۴، واژه‌بست فاعلی در جدول ۵-۴، و تک‌واژه‌های محاوره در جدول ۶-۴ گنجانده شده است.

جدول ۱-۴ شناسه‌های مضارع^{۱۴}

بعد از واکه ی		بعد از واکه و		بعد از واکه آ		غیر رسمی		رسمی		شخص	شمار
غیر رسمی	رسمی	غیر رسمی	رسمی	غیر رسمی	رسمی	غیر رسمی	رسمی	غیر رسمی	رسمی	و شمار	
*ام / ×ام		*یم		*م		*یم		*م		ش ۱	
×ای		*بی		*ی		*بی		*ی		ش ۲	
به / *به	*د	به *	*د	به / د*	د	به / د*	د	به *	د	ش ۳	
*ایم		*بیم		*یم		*بیم		*یم		ش ۴	
*این	*اید	*بین	*اید	*بین / *بین	بید	*بید	بید	*بین	بید	ش ۵	
*ان / ×ان	*ند / ×اند	*ین	*یند	*ن	*ند	*یند	*ن	*ن	*ند	ش ۶	

جدول ۲-۴ شناسه‌های ماضی^{۱۵}

مثال	شناسه رسمی	شناسه غیر رسمی	شخص و شمار
رفتم، می‌دیدم	*م		ش ۱
رفتی، می‌دیدی	*ی		ش ۲
رفت، می‌دید	Ø		ش ۳
رفتیم، می‌دیدیم	*بیم		ش ۴
رفتین، می‌دیدین	*بین	*بید	ش ۵
رفتن، می‌دیدن	*ن	*ند	ش ۶

^{۱۴} Present Inflections

^{۱۵} Past Inflections

جدول ۴-۳ واژه‌بستهای ربطی

مثال	واژه‌بست ربطی		شخص و شمار
	غیر رسمی	رسمی	
رفته‌ام، می‌گفته‌ام، شسته بوده‌ام	*م	ام*	وربطی ۱
رفته‌ای، می‌گفته‌ای، شسته بوده‌ای	*ی*	ای*	وربطی ۲
رفته‌ست، می‌گفته است، شسته بوده است	*س / اس / Ø*	ست، *	وربطی ۳
رفتایم، می‌گفتایم، شسته بوده‌ایم	*یم*	ایم*	وربطی ۴
رفته‌اید، می‌گفته‌اید، شسته بوده‌اید	*این / *ین	اید*	وربطی ۵
رفته‌اند، می‌گفته‌اند، شسته بوده‌اند	*ان / *ن	اند*	وربطی ۶

واژه‌بستهای ربطی‌ای^{۱۶} که در افعال به کار می‌رond محدود به جدول ۴-۳ هستند، اما واژه‌بستهایی که در ساخت کلمات غیر فعلی به کار می‌رond بدلیل قرار گرفتن بعد از همخوان و واکه‌های مختلف از جدول زیر اندکی متنوع‌ترند که در قسمت مربوط به کلمات غیر فعلی آمده است.

جدول ۴-۴ واژه‌بستهای مفعولی

مثال	واژه‌بست مفعولی	شخص و شمار
کشوندیم، می‌کوبوندنم	*م	ومفعولی ۱
کشوندمت، می‌کوبوندت	*ت	ومفعولی ۲
کشوندمش، می‌کوبوندنش	*ش	ومفعولی ۳
کشوندیمون، می‌کوبوندنمون	*مان / *مون	ومفعولی ۴
کشوندمتون، می‌کوبوندنتون	*تان / *تون	ومفعولی ۵
کشوندمشون، می‌کوبوندنشون	*شان / *شون	ومفعولی ۶

استفاده از واژه‌بست مفعولی محدود به گفتار غیر رسمی و زبان ادبی است که گاهی به صورت محدود در متون رسمی هم استفاده می‌شود (کارین مگردویان، ۲۰۰۸/۲۰۰۰). بنابراین این نوع ساخت فعل در قسمت فعل غیر رسمی گنجانده شده است. این واژه‌بستها در ساختار فعل‌ها هیچگاه پس از واکه^۱، و قرار نمی‌گیرند بنابراین تنها موارد موجود در جدول ۴-۴ در فعل‌های غیر رسمی استفاده می‌شود.

^{۱۶} Copula Clitics

فعل‌های متعددی (گذرا^{۱۷}) می‌توانند واژه‌بست فعلی مفعولی^{۱۸} و یا واژه‌بست فعلی بپذیرند. در مقابل فعل‌های لازم (ناگذر) تنها می‌توانند واژه‌بست فعلی بپذیرند. در فعل‌های گذرا پذیرفتن واژه‌بست فعلی یا واژه‌بست مفعولی سوم شخص مفرد هر دو منجر به افزود شدن ش به انتهای بن فعل می‌شود که ابهام ایجاد می‌کند.

$$\text{فعل گذرا} + \left[\begin{array}{l} (\text{واژه بست مفعولی}) \\ (\text{واژه بست فعلی}) \end{array} \right] + \dots$$

$$\text{فعل ناگذر} + \left[\begin{array}{l} (\text{واژه بست فعلی}) \end{array} \right] + \dots$$

جدول ۴-۵ واژه‌بست فعلی

لامس یا متعدد بودن	وجه فعل	سوم شخص مفرد
لازم و متعدد	ماضی ساده	اومند*ش
لازم و متعدد	ماضی استمراری	میومد*ش
لازم و متعدد	ماضی بعید	اومنده بود*ش
لازم و متعدد	ماضی التزامی	اومنده باشد*ش
لازم	مضارع اخباری و التزامی	بیادش / میاد*ش

واژه‌بست فعلی^{۱۹} در فعل‌های سوم شخص بکار می‌رود. همه فعل‌ها چه متعددی (گذرا) و چه لازم (ناگذر) می‌توانند این واژه‌بست را پذیرند اما در فعل‌های گذرا این واژه‌بست فعلی با واژه‌بست مفعولی تداخل دارد (در توزیع تکمیلی با واژه‌بست مفعولی قرار دارد) و ابهام ایجاد می‌کند (محرم اسلامی و صدیقه علیزاده، ۱۳۸۸).

استفاده از این واژه‌بست فعلی محدود است به این ساختهای و سایر فعل‌ها که از این فعل‌های ساده استفاده می‌کنند (ملموس، مرکب، پیشوندی و غیره).

فعل‌های مضارع نیز تنها بخش فعل‌های لازم آن این واژه‌بست را می‌پذیرند و فعل‌های متعددی مضارع نمی‌توانند چنین ساختی تولید کنند.

^{۱۷} در واقع فعل‌های لازم و فعل‌های دووجهی واژه‌بست فعلی می‌پذیرند اما از آنجا که اغلب فعل‌های متعددی در زبان غیر رسمی با این واژه‌بست استفاده می‌شوند بنابراین به جای فعل‌های دووجهی، تمامی فعل‌های متعددی به این ساختار افزوده شده‌اند.

^{۱۸} Object Clitics

^{۱۹} Subject Clitic

مثال: رفتش (ناگذر)، ریختش (دووجهی^۱)، خوردش (گذرا)

جدول ۴-۶ تکوازهای محاوره

نام	واژه بست	واژه معادل	مثال
هم	م*	هم	(اگر) می‌گفتیم (گوش نمی‌کرد)
عطاف	و*	و	اومنو (نشست)، وايسادو (خندید)
تاکید*	ها / آ*	-	مي زنمتا، مي خوردا

* طبق شقاقی (۱۳۹۴)، تاکید یک واژه‌بست است، اما علت اینکه در عنوان وند آمده است، این است که در پژوهش شقاقی این تکواز تنها منحصر به فعل فرض شده است که در آن تکیه کلمه را به خود جذب نمی‌کند (از میان چند آزمون مختلف برای تشخیص وند بودن یا واژه‌بست بودن) و در این پژوهش علاوه بر فعل و ساختهای اسنادی به کلمه‌های غیر فعلی نیز می‌تواند متصل شود. در انتقال به اسم‌ها، این تکواز تکیه کلمه را جذب می‌کند. بنابراین از نظر واژه‌بست بودن ظاهرا این تکواز دو رفتار متفاوت در افعال و اسمای از خود بروز می‌دهد. البته بررسی نظری این مورد در حوزه پژوهش فعلی نمی‌گنجد و صرفاً شناسایی آن در این پژوهش مد نظر است. گرچه این تکواز در این پژوهش به عنوان وند در نظر گرفته شده است، استفاده کنندگان از این تحلیل گر به قطعه + تاکید در قواعد خروجی تحلیل گر، هر عنوانی که مناسب می‌دانند می‌توانند اطلاق کنند.

تکوازهای محاوره تنها در ساختهای غیر رسمی به کار می‌روند و آخرين قسمت در ساختمان کلمات غیر رسمی هستند. تکوازهای این گروه از لحاظ معنایی متفاوت از یکدیگرند. قرار گرفتن آنها در یک گروه به این دلیل است که در توزیع تکمیلی هم قرار دارند.

واژه‌بستهای **هم** و **عطاف** بر اساس تحلیل وبلاگ‌های فارسی (کاربن مگردمیان، ۲۰۰۸) تعریف شده است، اما تکواز تاکیدی و همینطور ساختمان این گروه که اجزای آن در توزیع تکمیلی هم هستند، با بررسی پیکره جمع‌آوری شده برای این پژوهش، شناسایی و استخراج شده است.

۴-۳- ساختمان فعل ساده

گرچه مبنای تحلیل گر تصریفی کلمه واحد است اما فعل‌های ساده گاهی در ساختمان خود از دو کلمه یا حتی سه کلمه (ماضی ابعد) استفاده می‌کند (واحدهای چند قطعه‌ای^۲). این کلمه‌های اضافه، فعل‌های کمکی‌ای هستند که در مرحله تولید تولید به راحتی با دادن قاعده سازنده آنها به تحلیل گر تصریفی، تولید می‌شوند اما در مرحله شناسایی اگر توسط جداساز^۳

^۱ به فعل‌هایی که هم به شکل گذرا و هم ناگذر بکار می‌روند فعل دووجهی یا دوگانه می‌گویند (گیوی و انوری، ۱۳۹۱).

^۲ MTU – Multi Tokens Unit

^۳ Tokenizer

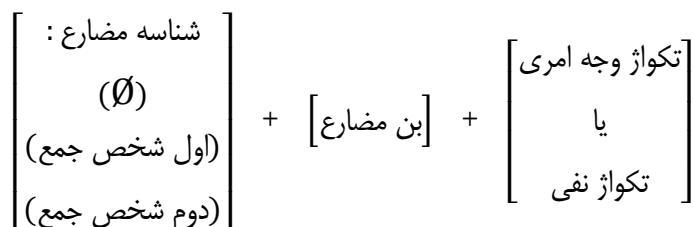
درست جدا شوند، تحلیل گر تصریفی هم به راحتی می‌تواند آنها را شناسایی کرده و قاعده سازنده آنها را تولید کند، اما در صورتی که به هر دلیلی همه کلمات سازنده آن به تحلیل گر داده نشود، تحلیل گر به صورت جداجدا قاعده سازنده هر کلمه را تولید می‌کند و برای شناسایی واحد کامل فعل می‌بایست با ابزاری دیگر و با توجه به بافت و قواعد تولید شده (از تحلیل گر تصریفی) آنها را شناسایی کرد. تمامی قاعده‌های سازنده تک‌تک کلمات اینگونه فعل‌ها در بخش پیاده‌سازی به طور کامل توضیح داده شده است.

این فعل‌ها (ماضی بعید، ماضی التزامی و مستقبل) جدای از فعل‌های تک کلمه‌ای تعریف شده‌اند و در مرحله کامپایل مبدل می‌توان آنها را حذف و یا اضافه کرد (که از ساخت نهایی تحلیل گر کنار گذاشته شده‌اند اما ساختارها در کد موجود است).

در ساخت غیر رسمی ساخته‌های مستقبل و ماضی بعد وجود ندارند.

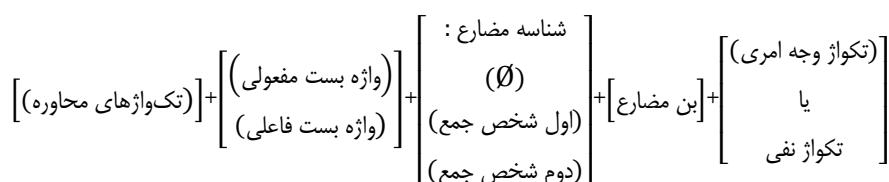
۴-۱-۳- امری

قاعده سازنده فعل امری^۱ رسمی:



مثال: برو، نخوانید، بشین

قاعده سازنده فعل امری غیر رسمی:



مثال: برو، نخوانیدشان، بشینین

استثناء:

برخی بن‌های مضارع می‌توانند بدون تکواز ب نیز و یا فقط بدون آن، فعل امری بسازند. این بن‌ها عبارتند از باش / دار (همیشه بدون تکواز ب استفاده می‌شود)، کن / شو (هم با تکواز ب و هم بدون آن می‌تواند فعل امری بسازد)، مثال: (درست رفتار) کن / کنید / بکن / بکنید / بکنین، (آنجا) باش، (آرام) باشین، (خفه) شو، (گم) شو، (خبر) دار، (نگه) دارید.

^۱ Imperative

فعل‌های امری غیر از ساخت دوم شخص مفرد در سایر ساخت‌ها، منطبق بر فعل‌های التزامی است. جدول ۷-۴ شbahet ساخت امری و التزامی را نشان می‌دهد. در ساخت امری رسمی، می‌توان به طور کلی فعل امری را تنها به این ساخت مفرد محدود کرد و شناسایی دیگر شخص و شماره‌های امری را در پس‌پردازش از روی بافت فعل التزامی رسمی انجام داد. اما در فعل امری غیر رسمی تفاوتی بین ساخت امری و التزامی وجود دارد، و آن نپذیرفتن واژه‌بست هم در ساخت فعل امری غیر رسمی در مقابل فعل مضارع التزامی غیر رسمی است.

ساختار فعل‌های امری جمع در قواعد لکس تعریف شده است و در فایل فعل‌ها موجود است اما از ساخت نهایی کنار گذاشته شده است.

جدول ۷-۴ - تصریف ویژه همه فعل‌های امری مفرد

۶	۵	۴	۳	۲	۱	امری یا التزامی	ساخت فعل
بروند	بروید	برویم	برود	بروی	بروم	التزامی	فعل ساده رسمی
-				-	-	امری	
بگویند	بگویید	بگوییم	بگوید	بگویی	بگویم	التزامی	فعل ساده غیر رسمی
-				-	-	امری	
برن	برین	بریم	بره	بری	برم	التزامی	فعل پیشوندی رسمی
-				-	-	امری	
بگن	بگین	بگیم	بگه	بگی	بگم	التزامی	فعل پیشوندی غیر رسمی
-				-	-	امری	
بشن	بشین	بشیم	بشه	بشی	بشم	التزامی	فعل پیشوندی رسمی
-				-	-	امری	
درروند	درروید	دررویم	دررود	درروی	درروم	التزامی	فعل پیشوندی غیر رسمی
-				-	-	امری	
پاشن	پاشین	پاشیم	پشه	پاشی	پاشم	التزامی	فعل پیشوندی رسمی
-				-	-	امری	
وايسن	وايسین	وايسیم	وايسه	وايسی	وايسم	التزامی	فعل پیشوندی غیر رسمی
-				-	-	امری	
دررن	دررین	درریم	درره	درری	دررم	التزامی	فعل پیشوندی رسمی
-				-	-	امری	

تصریف خاص فعل امری دوم شخص مفرد به اینجا ختم نمی‌شود و در ساخت غیر رسمی نیز برخی بن‌های غیر رسمی در آن مورد استفاده قرار نمی‌گیرد (جدول ۸-۴). این الگو در مورد یک بن رسمی هم صدق می‌کند، گرچه در این مورد در ساخت مفرد امری، در عوض از فعل مرکب استفاده می‌شود (سطر آخر جدول ۸-۴).

جدول ۸-۴ - تصریف ویژه برخی فعل‌های امری مفرد

اول شخص مفرد	دوم شخص جمع	بن فعل	نوع فعل	
			غیر رسمی	رسمی
برید / برین	بریم	برو	ر	رو
بشدید / بشین	بشیم	بشو	ش	شو
بگید / بگین	بگیم	بگو	گ	گو
بدید / بدین	بدیم	بده	د	دد
پاشین / پاشید	پاشیم	پاشو	پا-ش	پا-شو
بگریید	بگرییم	گریه کن	-	گر / گری

همانطور که ملاحظه می‌شود چهار فعل غیر رسمی چهار سطر اول و فعل غیر رسمی پیشوندی سطر پنجم، در ساخت دوم شخص مفرد از بن رسمی استفاده می‌کنند و در سایر ساخت‌ها همان بن فعل غیر رسمی را بکار می‌گیرند. در سط آخر نیز بن فعل **گر / گری** است که ساخت امری اول شخص جمع و دوم شخص جمع آن استفاده می‌شود. مثل **بگریید، بگرییم**. اما ساخت دوم شخص مفرد ندارد و در عوض از فعل مرکب **گریه کن** استفاده می‌شود. این فعل‌ها در ساخت سوم شخص مفرد مضارع اخباری و التزامی هم نمی‌توانند شناسه رسمی ۵ پذیرند، که از این حیث رفتار این پنج فعل و فعل **وایسادن** در امری مفرد، مضارع اخباری و التزامی سوم شخص مفرد مانند هم است (جدول ۹-۴).

۴-۳-۲- مضارع

دو ساخت مضارع التزامی و اخباری تنها در پیشوند با هم متفاوتند. ساخت التزامی پیشوند **ب** یا **نفی ن** می‌پذیرد و ساخت اخباری پیشوند **می** یا **نفی نمی** می‌پذیرد. بقیه فعل ساخت یکسانی دارد.

۴-۳-۱- مضارع التزامی^۱

رسمی:

$$\left[\begin{array}{c} \text{تکواز وجه التزامی} \\ \text{یا} \\ \text{تکواز نفی} \end{array} \right] + \left[\begin{array}{c} \text{بن مضارع} \\ + \\ \text{شناشهای مضارع} \end{array} \right]$$

^۱ Present Subjunctive

مثال: بروم، نخوانید، بشینی.

غیر رسمی:

$$\left[\begin{array}{c} \text{(تکواز وجه التزامی)} \\ \text{+} \\ \text{بن مضارع} \end{array} \right] + \left[\begin{array}{c} \text{(واژه‌بستهای مفعولی)} \\ \text{+} \\ \text{شناسه‌های مضارع} \end{array} \right] + \left[\begin{array}{c} \text{(تکوازهای محاوره)} \\ \text{+} \\ \text{(واژه‌بست فاعلی)} \end{array} \right]$$

مثال: ببرم، نخونیدش، بشینن

فعل‌های مضارع التزامی‌ای که بدون پیشوند ب ساخته می‌شوند در قسمت مضارع ساده تولید می‌شوند.

۴-۳-۲- مضارع اخباری^۱

رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ \text{+} \\ \text{بن مضارع} \end{array} \right] + \left[\begin{array}{c} \text{تکواز می} \\ \text{+} \\ \text{شناسه‌های مضارع} \end{array} \right]$$

مثال: می‌روم، نمی‌رونده، می‌نشینید

فعل می‌باشد که در نقش اسناد ظاهر می‌شود و سایر شخص و شمارهای آن در این بخش تولید می‌شود.

غیر رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ \text{+} \\ \text{بن مضارع} \end{array} \right] + \left[\begin{array}{c} \text{(تکواز می)} \\ \text{+} \\ \text{شناسه‌های مضارع} \end{array} \right] + \left[\begin{array}{c} \text{(واژه‌بستهای مفعولی)} \\ \text{+} \\ \text{(تکوازهای محاوره)} \end{array} \right]$$

مثال: می‌رم، نمی‌رن، می‌شینین

بن‌های رسمی و غیر رسمی با همه شناسه‌های رسمی و غیر رسمی می‌توانند ترکیب شوند. دو مورد استثناء وجود دارد که یکی در جدول ۹-۴ آمده است. این بن فعل‌های غیر رسمی درست مانند رفتارشان در ساخت فعل امری مفرد (جدول ۸-۴) که با معادل رسمی جایگزین می‌شد، این بار نیز فعلی تولید نمی‌کند و شناسه غیر رسمی را تنها می‌پذیرند.

مورد دیگر بن‌های رسمی است که در ساخت سوم شخص مفرد، شناسه غیر رسمی نمی‌پذیرند. این بن فعل‌ها همگی از دسته بن‌هایی‌اند که منتهی به حرفی هستند که در نقش همخوان ظاهر شده است اما در بن‌های دیگر می‌توانند واکه هم باشند (جدول ۱۰-۴). همه بن‌های این بخش منتهی به و و هم در نقش همخوان و هم در نقش واکه ظاهر می‌شوند) که در اینجا در نقش همخوان بکار رفته‌اند.

^۱ Present Indicative

جدول ۴-۹ استثنای ترکیب نشدن برخی بن‌های غیر رسمی با شناسه‌های رسمی

شناسه سوم شخص مفرد		بن فعل		نوع فعل
غیر رسمی	رسمی	غیر رسمی	رسمی	
بره	* برد	ر	رو	غیر رسمی
بشه	* بشد	ش	شو	
بگه	* بگد	گ	گو	
بده	* بدد	د	دد	
واشه / واسته	وايسه / وايسيه *	واي-س / وا-سنس	-	پیشوندی
پاشه	* پاشد	پا-ش	پا-شو	

جدول ۱۰-۴ استثنای ترکیب نشدن برخی بن‌های رسمی با شناسه‌های غیر رسمی

شناسه سوم شخص مفرد		بن فعل رسمی	شناسه سوم شخص مفرد		بن فعل رسمی
غیر رسمی	رسمی		غیر رسمی	رسمی	
* کاهه	کاهد	کاه	* دهه	دهد	دد
* نهه	نهد	نه	* خواهه	خواهد / (می) خواه	خواه
* شوه	شود	شو (شن)	* پژوهه	پژوهه	پژوه
* تراوه	تراود	تراو	* جهه	جهد	جه
* رهه	رهد	رو (رفتن)	* رهه	رهد	ره
* شیوه	شیود	شیو	* ستوهه	ستوه	ستوه

۴-۳-۲-۳- مضارع ساده^۱

رسمی:

$$[\text{بن مضارع}] + [\text{شناسه‌های مضارع}]$$

مثال: دارم، باشم، (دیده) شوم.

ریشه‌های محدودی در این ساخت استفاده می‌شود، مگر در زبان ادبیانه که در این تحلیل گر پوشش داده نشده است.

^۱ Simple Present

غیر رسمی:

$$\left[\text{بن مضارع} + \left[\begin{array}{l} \text{شناسه‌های مضارع} \\ \text{(تکوازهای محاوره) } \end{array} \right] \right] + \left[\begin{array}{l} \text{(واژه‌بستهای مفعولی)} \\ \left(\text{بن مضارع} + \left[\begin{array}{l} \text{شناسه‌های مضارع} \\ \text{(تکوازهای محاوره) } \end{array} \right] \right) \end{array} \right]$$

مثال: دارمشون، باشه، (دیده) شم.

بخشی از فعل‌های کمکی^۱ در این بخش ساخته می‌شود. مثل: [باش / خواه / شو / دار] + شناسه مضارع.

مثال: دارم، باشم، (دیده) شوم.

علاوه بر آن فعل‌های التزامی با بن کن (باش / خواه / شو / دار) هم استثناءً در این قسمت ساخته می‌شود زیرا در برخی ساختها ممکن است بدون تکواز التزامی "ب" استفاده شوند. مثل: (درست) رفتار کنم، (اگه خوب) شم.

۴-۳-۳- ماضی

تفاوت فعل ماضی ساده و ماضی استمراری در پذیرفتن پیشوند استمراری می‌در ماضی استمراری است.

۴-۳-۳-۱- فعل ماضی ساده^۲

رسمی:

$$\left[\text{(تکواز نفی)} \right] + \left[\text{بن ماضی} \right] + \left[\text{شناسه‌های ماضی} \right]$$

مثال: رفتم، رفتند، نخندیدید، هستم (اسناد)، نیستید (اسناد)

غیر رسمی:

$$\left[\text{(تکواز نفی)} \right] + \left[\begin{array}{l} \text{بن ماضی} \\ \left[\begin{array}{l} \text{شناسه‌های ماضی} \\ \text{(تکوازهای محاوره) } \end{array} \right] \end{array} \right] + \left[\begin{array}{l} \text{(واژه‌بستهای مفعولی)} \\ \left(\text{بن ماضی} + \left[\begin{array}{l} \text{شناسه‌های ماضی} \\ \text{(تکوازهای محاوره) } \end{array} \right] \right) \end{array} \right]$$

مثال: رفتم، رفتش، نخندیدین، هستن (اسناد)، نیستین (اسناد)، هستش

بخشی از فعل‌های کمکی در این بخش ساخته می‌شود. مثل: [بود / خواست / شد / داشت] + شناسه ماضی.

فعل‌های اسنادی غیر از است و هست (و سایر تصریف‌های آنها) در این قسمت ساخته می‌شود (بن‌های شد، گشت، گردید، بود).

۴-۳-۳-۲- فعل ماضی استمراری^۳

رسمی:

$$\left[\text{(تکواز نفی)} \right] + \left[\text{تکواز می} \right] + \left[\text{بن ماضی} \right] + \left[\text{شناسه‌های ماضی} \right]$$

مثال: می‌رفتم، نمی‌خندیدید، می‌گفتند.

^۱ Auxiliaries

^۲ Simple Past

^۳ Imperfect

غیر رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{شناسه‌های ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بستهای مفعولی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{(تکواز می)} \\ + \end{array} \right]$$

$$\left[\begin{array}{c} \text{(وژه‌بست فاعلی)} \\ + \end{array} \right]$$

مثال: می‌رفتم، نمی‌خندیدم، می‌گفتمن، می‌پیچوندم

۴-۳-۳-۱- ماضی نقلی^۱

دو فعل ماضی نقلی و نقلی مستمر تنها در پذیرفتن پیشوند استمراری می‌در فعل مستمر با هم متفاوتند.

رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{تکواز ه} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه بستهای ربطی)} \\ + \end{array} \right]$$

مثال: خریده‌ام، رفته‌اند، نخندیده‌اید، گفته (است).

غیر رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بستهای مفعولی)} \\ + \end{array} \right]$$

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بست فاعلی)} \\ + \end{array} \right]$$

مثال: خریده‌ام، رفته‌ن، نخندیده‌این، گفته (است)

۴-۳-۳-۲- فعل ماضی نقلی مستمر^۲

رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بستهای ربطی)} \\ + \end{array} \right]$$

مثال: می‌رفته‌ام، نمی‌خندیده‌ای، می‌گفته‌اند، می‌خریده است

غیر رسمی:

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بستهای مفعولی)} \\ + \end{array} \right]$$

$$\left[\begin{array}{c} \text{(تکواز نفی)} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{بن ماضی} \\ + \end{array} \right] \left[\begin{array}{c} \text{(وژه‌بست فاعلی)} \\ + \end{array} \right]$$

مثال: می‌رفته‌ام، نمی‌خندیده‌ای، می‌گفته‌ان، می‌خریده (است)

۴-۳-۴- فعل‌های چند قطعه‌ای ساده

فعل‌های این بخش در ساختارها تعریف شده است اما بدون در اختیار داشتن جداساز مناسب نمی‌توان از این ساختها استفاده کرد.

^۱ Present Perfect

^۲ Compound Imperfect

۴-۳-۱- ماضی التزامی^۱

رسمی:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های مضارع]$$

مثال: رفته باشم، نخورده باشد، گفته باشید

غیر رسمی:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های مضارع] + [(تکوازهای محاوره)] + [(واژه‌بستهای مفعولی)]$$

مثال: رفته باشم، نخورده باشه، گفته باشین

۴-۳-۲- ماضی بعيد^۲

رسمی:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های ماضی]$$

مثال: رفته بودم، خریده بود، گفته بودید

غیر رسمی:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های ماضی] + [(تکوازهای محاوره)] + [(واژه‌بستهای مفعولی)]$$

مثال: رفته بودم، خریده بود، گفته بودش

۴-۳-۳- ماضی ابعد^۳

تنها ساخت رسمی دارد:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های ربطی]$$

مثال: رفته بوده‌ام، خریده بوده است، گفته بوده‌اید

۴-۳-۴- مستقبل^۴

تنها ساخت رسمی دارد:

$$[(تکواز نفی) + [بن ماضی] + [باش] + [شناسه‌های مضارع]$$

مثال: خواهم رفت، نخواهد رفت، خواهید گفت

^۱ Past Subjunctive

^۲ Pluperfect

^۳ Double Compound

^۴ Future

۴-۳-۵- سایر فعل‌ها

سایر فعل‌های باقی مانده در این قسمت قرار دارد. فعل‌های اسناد و وجهی در قواعد تحلیل گر تعریف شده است، اما فعل‌های ملموس می‌بایست در پس‌پردازش شناسایی شود.

۴-۱-۵- اسناد

فعل اسنادی^۱ دارای ساختی رسمی است و سایر بن فعل‌هایی که فعل اسنادی می‌سازند در بخش ماضی ساده تولید می‌شوند.

$$[\text{شناسه‌های ماضی}] + [\text{ست}] + [\text{(تکواز نفی)}]$$

مثال: هست، است، نیست، هستند، نیستید، سنت.

۴-۲-۵- فعل وجهی

فعل وجهی^۲ تنها می‌تواند تصریف نفی و پیشوند می بپذیرد و از مجموعه‌ای از واژگان تشکیل شده است. مثال: (ن)بایست، (ن)باید، (ن)می‌توان، همینطور بتوان و نتوان مربوط به این بخش هستند^۳.

۴-۳-۵- فعل‌های ملموس

فعل‌های ماضی^۴ و مضارع مستمر^۵ (ملموس) و فعل ناقص ماضی ملموس نقلی می‌توانند از فعل‌های ساده، مرکب، پیشوندی و پیشوندی مرکب ساخته شوند. بنابراین نحوه ساخت و یا شناسایی آنها درست مانند شناسایی تک‌تک کلمات سازنده آن (садه، مرکب، پیشوندی وغیره) است. در ادامه ساخت تک‌تک آنها توضیح داده شده است.
ماضی ملموس:

$$\{\text{ماضی ساده داشت}\} + \{\text{فعل ماضی استمراری}\}$$

که در آن شناسه‌های دو فعل باید مطابقت داشته باشند.

مثال: داشتند (درس) می‌خوانند، داشتم (از خرید) بر می‌گشتم
مضارع ملموس:

$$\{\text{مضارع ساده دار}\} + \{\text{فعل مضارع اخباری}\}$$

که در آن شناسه‌های دو فعل باید مطابقت داشته باشند.

مثال: دارم می‌روم، دارد (سریع) می‌دود

^۱ Copula

^۲ Modal

^۳ فعل‌های می‌شود و می‌شه به معنای می‌توان و می‌تواند هم فعل وجهی محسوب می‌شود اما نقش اصلی آن مضارع اخباری است و در همان بخش نیز تولید می‌شود. تشخیص نقش‌های دیگر این افعال با توجه به بافت و در پس‌پردازش صورت می‌گیرد.

^۴ Past Progressive

^۵ Present Progressive

ماضی ملموس نقلی:

{داشته (ماضی نقلی سوم شخص مفرد داشت) + {فعل ماضی نقلی متمر سوم شخص مفرد}

که در آن شناسه‌های دو فعل باید مطابقت داشته باشند.

این فعل ناقص است و تنها همین شناسه را می‌پذیرد. مثال: داشته می‌رفته، داشته بر می‌گشته، داشته (از سر کوچه) می‌آمده.

۴-۴-۳- موارد ویژه

این قسمت شامل تصریف‌های ویژه و یا ناقص می‌شود. موارد این قسمت متفاوت از قواعد نگارشی است که در بخش بعد توضیح داده می‌شود.

۴-۴-۱- فعل‌های ناقص

فعل‌هایی را که همه ساختها و زمان‌های آنها متداول نیست فعل ناقص می‌گویند (حسن احمدی گیوی، حسن انوری؛ ۱۳۹۱). برخی بن فعل‌های غیر رسمی در همه ساختها استفاده نمی‌شوند و مانند فعل‌های ناقص محدود به تعدادی از ساختها هستند. جدول ۱۲-۴ و جدول ۱۱-۴ ساخت این افعال را نشان می‌دهد. این بن فعل‌ها در صورتی که پیشوند استمراری می، نفی ن و التزامی ب در ابتدای فعل به کار رود، تغییر می‌کنند. این تغییر عموماً به حذف حرف اول بن فعل می‌انجامد. در برخی موارد نیز منجر به حذف حرف و اضافه شدن واکه‌ای به جای آن می‌شود.

فعل‌های مضارع همواره پیشوندهای ذکر شده را به همراه دارند، بنابراین همیشه از شکل تغییر یافته بن فعل استفاده می‌کنند.

فعل‌های مستقبل و ماضی بعد جزو فعل‌های غیر رسمی نیستند اما برای روشن شدن موضوع در جدول ۱۲-۴ گنجانده شده‌اند.

جدول ۱۱-۴ تغییرات برخی بن فعل‌های مضارع

بن فعل	گذار	شون	شین	انداز	افت	افت	اف	اف	بن تغییر یافته
ذار	ذارم	شون	شین	نداز	فت	وفت			
نمی‌ذاری	نمی‌ذارم	نمی‌شونیم	نمی‌شینیم	نمیندازی	نمیقته	میوقتن			مضارع اخباری
بدارم	بسونه	بسین	بداری	نمینداری	نمیقیم	نمیوقتن			مضارع التزامی
نذاری	نَشونیم	نَشینیم	نذاره	بیفتحیم	بیوفتن				امروز
بزار	بسونیم	بسین	بندازید	بیفتح	بیوفت				
نذارین	نَشونید	نَشین	نذار	نیفتحیم	نیوفتن				

جدول ۱۴-۴ فعل‌های ناقص ساخت‌های ماضی

بن فعل	گذاشت	نشوند	نشست	انداخت	افتاد	افتاد	آورد	اوmd
بن تغییر یافته	ذاشت	شوند	شست	نداخت	فتاد	وقتاد	ورد	وmd
ماضی ساده	گذاشتم	نشوندم	نشستم	انداختم	افتادم	افتادین	اوردم	اوmdm
	-	-	-	-	-	-	نیوردم	نیومدی
ماضی استمراری	میداشتم	میشوندم	میشستم	مینداختم	میفتادم	میوفتادین	میوردم	میومد
	-	-	-	-	-	-	نمیوردم	نمیومدن
صفت معمولی، ماضی نقلی، بعید، بعد و التزامی	گذاشته	نشونده	نشسته	انداخته	افتاده	افتاده	آورده	اوmdه
	-	-	-	-	-	-	نیورده	نیومده
ماضی نقلی مستمر	میداشته	میشونده	میشسته	مینداخته	میفتاده	میوفتاده	میورده	میومده
	-	-	-	-	-	-	نمیورده	نمیومده
مستقبل	گذاشت	نشوند	نشست	انداخت	افتاد	افتاد	آورد	اوmd
	-	-	-	-	-	-	اورد	اوmd

در فعل‌های پیشوندی این محدودیت کمتر است و به دلیل وجود پیشوند، بن‌های تغییر یافته غیر از چند مورد، در همه ساخت‌ها استفاده می‌شوند (جدول ۱۳-۴ و جدول ۱۴-۴).

جدول ۱۳-۴ تغییرات بن‌های پیشوندی مضارع

بن فعل	در-افت	در-افت	در-افت	باز-گذار
بن تغییر یافته	در-رفت	در-رفت	در-رفت	باز-ذار
مضارع اخباری	در-میفته	در-میفته	در-میفته	باز-میداریم
	-	-	-	باز-نمیدارین
مضارع التزامی	در-بیفته	در-بیفته	در-بیفته	باز-بذارم
	-	-	-	باز-نذارم
امری	در-بیفت	در-بیفت	در-بیفت	باز-بذار
	-	-	-	باز-نذارین

* بن فعل پیشوندی درافتادن، هم با پیشوند التزامی ب و هم بدون آن استفاده می‌شود. در ساخت بدون پیشوند التزامی بن تغییر یافته استفاده نمی‌شود.

جدول ۴-۴ تغییرات بن فعل‌های پیشوندی ماضی^۱

باز-گذاشت	در-اوید	ور-اوید	در-آورد	در-افتاد	در-افتاد	بن فعل
باز-ذاشت	در-ومند	ور-ومند	در-وردن	در-وقتاد	در-وقتاد	بن تغییر یافته
باز-گذاشتمن	در-ومندم	ور-ومند	در-وردم	در-افتادین	در-افتادن	+ ماضی ساده
باز-نذاشتی	در-نیومندم	ور-نیومندم	در-نیوردم	در-نیوفتادن	در-نیفتادیم	-
باز-میداشت	در-میومند	ور-میومند	در-میوردم	در-میوفتادین	در-میفتادن	+ ماضی استمراری
باز-نمیداشت	در-نمیومندن	ور-نمیومند	در-نمیوردن	در-نمیوفتاد	در-نمیفتادم	-
باز-گذاشتنه	در-ومنده	ور-ومنده	در-ورده	در-افتاده	در-افتاده	+ ماضی نقلی، بعید،
باز-نذاشتنه	در-نیومنده	ور-نیومنده	در-نیورده	در-نیوفتاده	در-نیفتاده	- ابعاد و التزامی
باز-میداشته	در-میومنده	ور-میومنده	در-میورده	در-میوفتاده	در-میفتاده	+ ماضی نقلی مستمر
باز-نمیداشته	در-نمیومنده	ور-نمیومنده	در-نمیورده	در-نمیوفتاده	در-نمیفتاده	-
گذاشت	در-اوید	ور-اوید	در-آورد	در-افتاد	در-افتاد	+ مستقبل
گذاشت	در-اوید	ور-اوید	در-آورد	در-افتاد	در-افتاد	-

۴-۴-۲- حرف واسطه

در فعل‌های متعددی غیر رسمی سوم شخص که واژه‌بست مفعولی می‌پذیرند معمولاً بین واژه‌بست و فعل، حرف ت به شکل واسطه قرار می‌گیرد تا بین شناسه فعل و مفعول آن که واژه‌بست مفعولی است تداخل ایجاد نشود (جدول ۴-۴).

^۱ نمونه‌های موجود در این جدول و برخی جداول دیگر این قسمت با توجه به مشاهده داده‌های جمع‌آوری شده انتخاب شده است. برای بررسی دقیق‌تر می‌بایست با آمارگیری از موقعه هر یک مشخص شود کدام نمونه‌ها می‌توانند معیار قرار بگیرند و یا همه موارد به عنوان بازنمایی درست انتخاب شود.

جدول ۱۵-۴ ساختهای متعدد همراه با واژه‌بست مفعولی

مثال	سوم شخص مفرد	ساخت
(می) خوردن (می) خورد*تشون	(می) + بن ماضی + واژه‌بست مفعولی (می) + بن ماضی + ت + واژه‌بست مفعولی	ماضی ساده و استمراری
خورده بودش	بن ماضی + ه + بود + واژه‌بست مفعولی	ماضی بعید
خورده باشدشون	بن ماضی + ه + باشد + واژه‌بست مفعولی	ماضی التزامی
(می) خورده شان (می) خورده*تشون	(می) + بن ماضی + ه + واژه‌بست مفعولی (می) + بن ماضی + ه + ت + واژه‌بست مفعولی	ماضی نقلی و نقلی مستمر
(می / ب) خوره شان (می / ب) خوره*تشون (می / ب) خوردشون (می / ب) خور*تشون	(می / ب) + بن مضارع + ه + واژه‌بست مفعولی (می / ب) + بن مضارع + ه + ت + واژه‌بست مفعولی (می / ب) + بن مضارع + د + واژه‌بست مفعولی (می / ب) + بن مضارع + ت + واژه‌بست مفعولی	مضارع التزامی و اخباری منتھی به همخوان
(می / ب) پایه شون (می / ب) پایه*تشون (می / ب) پایدشون (می / ب) پای*تشون	(می / ب) + بن مضارع + یه + واژه‌بست مفعولی (می / ب) + بن مضارع + یه + ت + واژه‌بست مفعولی (می / ب) + بن مضارع + د + واژه‌بست مفعولی (می / ب) + بن مضارع + ی + ت + واژه‌بست مفعولی	منتھی به واکه <u>ا</u>
(می / ب) بویه شون (می / ب) بویه*تشون (می / ب) بویدشون (می / ب) بوی*تشون	(می / ب) + بن مضارع + یه + واژه‌بست مفعولی (می / ب) + بن مضارع + یه + ت + واژه‌بست مفعولی (می / ب) + بن مضارع + ید + واژه‌بست مفعولی (می / ب) + بن مضارع + ی + ت + واژه‌بست مفعولی	منتھی به واکه <u>و</u>

۴-۵- قواعد نگارشی

قواعد نگارشی آخرین لایه‌ای است که به روساخت منتهی می‌شود. برخی تغییرات آوایی و حذف و اضافه حروف و یا واکه‌ها در این قسمت انجام می‌گیرد.

برای یادآوری دوباره جدول مشخص کننده مرznماها و جایگزین آنها در جدول ۱۶-۴ آمده است. این مرznماها در ساختارهای جدول ۱-۴ شناسه‌های مضارع، جدول ۲-۴ شناسه‌های ماضی، جدول ۳-۴ واژه‌بستهای ربطی، جدول ۴-۴ واژه‌بستهای مفعولی، جدول ۵-۴ واژه‌بست فاعلی، جدول ۶-۴ تکوازه‌های محاوره و جدول ۱۵-۴ ساختهای متعدد همراه با واژه‌بست مفعولی به کار رفته است.

جدول ۴-۶ مرزنماهای به کار رفته در قواعد نگارشی فعل

نویسه	کار کرد
*	دو تکواز را بدون هیچ فاصله‌ای به هم متصل می‌کند.
×	دو تکواز را با نیم فاصله به هم متصل می‌کند.
۸	بین دو تکواز یک فاصله کامل قرار می‌دهد.

۴-۵-۱- تغییرات واژه‌بست محاوره هم در انتهای برخی افعال

در فعل‌های غیر رسمی تنها در ساخت سوم شخص مفرد و بدون واژه‌بست مفعولی، کلمه به یک واکه، آنهم حرف ۵ ختم می‌شود. واژه‌بست هم در صورتی که بعد از این فعل بکار رود به صورت واژه کامل **هم** و نه واژه‌بست آن درمی‌آید.

۴-۵-۲- حرف ی میانجی در آغاز

برای بن فعل‌هایی که با **الف** شروع می‌شوند و قرار است پیشوند مضارع التزامی / امری **ب** و یا نفی **ن** بپذیرند، حرف ی میانجی بکار می‌رود. دامنه تاثیر این قاعده همه فعل‌ها هستند، زیرا بن فعل‌های بسیاری، چه فعل‌های مضارع و چه ماضی با حرف **الف** شروع می‌شوند.

استثنای این قاعده بن فعل‌هایی است که با **ای** شروع می‌شود؛ این بن‌ها **ایست** و **ایستن** هستند که می‌توانند فعل‌های **بایست**، **نایstem**، **بایستانم**، **نایستانند** و غیره بسازند.

همینطور فعل‌هایی که دارای بن فعل تغییر یافته هستند در صورت نیاز بجای **ب** و **ن**، **بی** و **نی** می‌گیرند.

۴-۵-۳- فاصله دادن ۵ از حرف بعدی

حرف ۵ در هر نقشی که ظاهر شود در صورتی که قبل از وند و یا واژه‌ای قرار بگیرد می‌بایست از آن فاصله بگیرد. بنابراین در تمام بافت‌ها در صورتی که حرفی از وند و یا واژه بعدی به حرف ۵ از وند و یا کلمه قبل بچسبد، با فاصله جدا می‌شود.

۴-۶- جمع‌بندی

در این فصل تمام فعل‌های ساده و پیشوندی (و حتی فعل‌های چند قطعه‌ای مثل مستقبل، ماضی التزامی، ماضی بعید و ماضی بعد) رسمی و غیر رسمی پوشش داده شد. سایر فعل‌ها با توجه به چند قطعه‌ای بودن در پس پردازش با استفاده از ابزارهای دیگر شناسایی می‌گردد، اما شناسایی تک تک این قطعه‌ها وظیفه تحلیل‌گر تصریفی است. در قسمت فعل‌های غیر رسمی ساختارهای ویژه‌ای نظری فعل‌های ناقص وجود دارد که همگی در این بخش پوشش داده شد. قواعد نگارشی برای فعل‌ها نیز در این فصل تعریف شد.

فصل پنجم

روش پیشنهادی؛ تصریف کلمه غیر

فعلی

۱-۵ - مقدمه

از آنجایی که تمامی کلمات این بخش از ساختمان اسمی در ساخت خود استفاده می‌کنند، آنها را در کنار هم و در یک فصل قرار می‌دهیم. دسته‌بندی کلمات غیر فعلی براساس تعریف دستور زبان فارسی (حسن احمدی گویی و حسن انوری، ۱۳۹۱) انجام پذیرفته است. اما در هر بخش کلمات غیر فعلی به دو دسته رسمی و غیر رسمی تقسیم می‌شوند. ساختمان هر یک از کلمات رسمی نیز با اندکی تغییر بر اساس دستور زبان فارسی، واژگان زایا (محرم اسلامی و همکاران، ۱۳۸۳) تعریف شده است. ساختمان کلمات غیر رسمی تا اندازه‌ای مبتنی بر تحلیل و بلاغ‌های فارسی (کارین مگردومیان، ۲۰۰۸)، تحلیل پیکره بنیاد منابع جمع‌آوری شده و شم زبانی نگارنده تعریف شده است.

دسته‌بندی کلمات غیر فعلی

کلمات این بخش به اسم، صفت، قید، عدد، حرف ربط، حرف اضافه، ضمیر / صفت (اشارة، مبهوم، پرسشی، مشترک، شخصی) تقسیم می‌شوند. می‌بایست در نظر داشت که نقش اصلی کلمات در نظر است و نه نقش مفهومی و یا نحوی که بنا به موقعیت می‌تواند در جمله به کار رود.

به طور مثال، بعضی اسمی و صفت‌ها می‌توانند نقش قید (مشترک) را هم در جمله بپذیرند، اما این نقش بسته به جمله و ساختار نحوی آن است، بنابراین شناسایی این نقش با توجه به مستقل از متن بودن تحلیل‌گر ممکن نیست. از طرف دیگر صفت‌ها می‌توانند با اشتغال صفر به اسم تبدیل شده و در همان ساختمان اسم با بکارگیری همه ظرفیت این ساختمان بکار روند. این ساختار نیز مورد توجه قرار نمی‌گیرد و همچنان با همان برچسب صفت قاعده تولید می‌کند. برای شناسایی آنها می‌توان در پس‌پردازش اقدام مناسب را انجام داد.

در مورد صفت و ضمیر (اشارة، مبهوم، پرسشی) نیز که بسته به بافت تفاوت‌شان روشن می‌شود، به طور کلی ضمیر و صفت از عنوان آنها حذف شده (در قاعده سازنده) و تنها بخش دوم آن باقی مانده است. در صورتی که بدون هیچ‌وند و واژه‌بستی به کار روند با توجه به بافت یا صفت هستند و یا ضمیر، اما در صورتی که قاعده و واژه‌بست در ساختمان آنها به کار رود قطعاً ضمیر هستند که در این صورت می‌بایست در پس‌پردازش این موارد روشن شود.

در موارد جزئی‌تر نیزی نکره از موصولی تفکیک شده است، هر کدام از این ساختها ویژگی‌های خاص خود را دارد اما در مواردی هم این دوساخت در سطح تصریف با هم همپوشانی دارند (با قطعه +نم در قاعده مشخص می‌شود) و امکان تمیز آنها نیست که در این موارد می‌بایست در پس‌پردازش این شناسایی با توجه به بافت صورت گیرد.

کلمه غیر فعلی رسمی و غیر رسمی

درست مانند بخش افعال، در این بخش نیز کلمات از ریشه‌های رسمی و غیر رسمی ساخته شده است. ساختارها هم به دو قسمت ساختارهای رسمی و غیر رسمی تقسیم می‌شود. کلمه رسمی شامل ریشه و ساختار رسمی است. کلمه غیر رسمی در ساختمان خود باید حداقل یک عنصر غیر رسمی (ریشه و یا ساختار) داشته باشد.

۲-۵-۱-۲-۵ اجزای ساختمان کلمه غیر فعلی

علاوه بر ریشه که در واژگان هر بخش گنجانده شده است، وندها و واژه‌بسته‌های تصریفی‌ای نیز وجود دارد که در این قسمت پوشش داده شده است.

۲-۵-۱-۲-۵ جمع

اسم‌های فارسی در ساختار رسمی با وندهای **ها** و **ان** جمع بسته می‌شوند. از اسم‌های فارسی آنها‌ی که جاندارند با استفاده از **ان** و **ها** (و در بعضی موارد که اسم منتهی به واکه **ه** است با **گان**) جمع بسته می‌شوند و بقیه اسم‌ها با **ها** جمع بسته می‌شوند.

اسم‌های عربی رایج در فارسی نیز با وندهای **ون**، **ین**، **ات** و **جات** جمع بسته می‌شوند. برخی اسم‌های عربی هم جمع مکسر هستند و وند جمع ندارند. برخی از این اسم‌ی در عربی جمع هستند اما در فارسی دو باه علامت جمع فارسی می‌پذیرند. برخی اسم‌ها نیز با چند وند جمع متفاوت جمع بسته می‌شود.

اسم‌های جمع مکسر عربی نیز که در زبان فارسی استفاده می‌شود با قطعه **+جم** در قاعده سازنده مشخص شده است. در ساختار غیر رسمی نیز علاوه بر ساختهای فوق از الف (۱) به تنها‌ی استفاده می‌شود.

واژگان اسمی بکار رفته در تحلیل‌گر بیش از ۲۷ هزار واژه است. برای شناسایی اسم‌های عربی، جمع‌های مکسر، وندهای جمع پذیرنده توسط هر اسم عربی و برای شناسایی اسم‌های جاندار فارسی که می‌توانند وند جمع **ان** پذیرند از لغتنامه معین (معین و امیریان، ۱۳۸۲) و پیکره بی‌جن‌خان (۲۰۱۱) استفاده شده است.

روش کار شناسایی اسم‌ها و وندهای جمع عربی به این صورت بوده است که ابتدا واژگان تحلیل‌گر را با توجه به عربی بودن آنها در لغتنامه معین، به عنوان اسم عربی مشخص کرده‌ایم (لغتنامه معین اسم‌های عربی را مشخص کرده است). سپس این اسم‌های عربی را با تک‌تک وندهای جمع عربی جمع بسته‌ایم و سپس آنها را در پیکره بی‌جن‌خان بررسی کرده‌ایم. آنها‌ی که در این پیکره وجود داشتند به عنوان وند جمع قبل قبول برای آن اسم عربی ذخیره کرده‌ایم. مثلاً اسم مؤسس هم می‌توان جمع **ات** و **هم** جمع **ین** پذیرد و کلمه‌های مؤسسات و مؤسسین را بسازد.

در مورد اسم‌های جاندار فارسی نیز از همین روش استفاده کرده‌ایم. و اسم‌ها‌ی که وند جمع **ان** را می‌پذیرند در اسامی شناسایی کرده‌ایم. مثل مؤسس که می‌تواند با وند جمع جاندار فارسی آن جمع شود و کلمه مؤسسان را بسازد.

۲-۵-۲-۵ نکره / موصولی

تکواز نکره و موصولی که در ظاهر یکسان هستند در جدول ۱-۵ آمده است.

جدول ۱-۵ وندهای نکرگی / موصولی

مثال	وند نکره یا موصولی	اسم منتهی به
کتابهایی، اعضا	بی / ئی / عی	واکه الف
باموبی، بانویی	بی / ئی / عی	واکه واو
باغچه‌ای، دگمه‌ای	ای / ئی / عی	واکه ه
دوگانگی‌ای، اپیدمی‌ای	ای / ای	واکه ی
کتابی، رفتاری، اقامتگاهی	ی	سایر حروف

۳-۲-۵- واژه‌بست شخصی

این واژه‌بست‌ها شامل واژه‌بست‌های ملکی، مفعولی، تفکیکی و غیر شخصی می‌شود (کارین مگردمیان، ۲۰۰۸). از آنجایی که تمایز آنها در سطح مفهومی انجام می‌گیرد و در سطح تصریف همیشه قابل تفکیک نیستند، بنابراین همه آنها در بخش کلمات غیر فعلی با عنوان واژه‌بست‌های شخصی نام‌گذاری می‌شوند.

ملکی^۱ مانند، کتابم، لباسم، مدادهات، املاکشون.

مفعولی^۲ مانند، جلومون، خوراکشون، باهام، برashون (و در ساخت افعال به کار می‌رود مانند، خوردشون، خریدنست).

تفکیکی^۳ مانند، بینشون، اونیکیاتون، هممون.

غیر شخصی^۴ مانند، خوشم (اومد)، خوابش (برد).

واژه‌بست‌های شخصی در جدول ۲-۵ آمده است.

^۱ Possessive clitics

^۲ Object clitics

^۳ Partitive clitics

^۴ Impersonal clitics

جدول ۲-۵ واژه‌بست‌های شخصی

شمار	رسمی بودن	همخوان	واکه واو	واکه ۱۵	واکه الف	واکه های	واکه ۵
۱	رسمی	کتابه: *م	لبویم: *یم	پنبه‌ام: *ام	بهایم: *یم	سیزیم: *م	بهام: *
	غیر رسمی		لبووم: *م	پنبه‌م: *م	پن‌م: *م		
۲	رسمی	کتابت: *ت	لبویت: *یت	پنبهات: *ات	بهایت: *یت	سیزیت: *ت	بهات: *
	غیر رسمی		لبوت: *ت	پنبه‌ت: *ت	پنت: *ت		
۳	رسمی	کتابش: *ش	لبویش: *یش	پنبه‌اش: *اش	بهایش: *یش	سیزیش: *ش	بهاش: *
	غیر رسمی		لبوش: *ش	پنبه‌ش: *ش	پنس‌ش: *ش		
۴	رسمی	کتابمان: *مان	لبویمان: *یمان	پنبه‌یمان: *یمان	بهایمان: *یمان	سیزیمان: *مان	بهامان: *
	غیر رسمی		لبومن: *مون	پنبه‌من: *مان	بهامان: *مان		بهامون: *
۵	رسمی	کتابتان: *تان	لبویتان: *یتان	پنبه‌یتان: *یتان	بهایمان: *یتان	سیزیمان: *تان	بهاتان: *
	غیر رسمی		لبوتون: *تون	پنبه‌تون: *تون	بهاتون: *تون		بهاتون: *
۶	رسمی	کتابشان: *شان	لبویشان: *یشان	پنبه‌یشان: *یشان	بهایشان: *یشان	سیزیشان: *شان	بهاشان: *
	غیر رسمی		لبوشون: *شون	پنبه‌شون: *شون	بهاشون: *شون		بهاشون: *

۴-۲-۵ معرفه

تک‌واژه معرفه که با وند ۵ به اسم متصل می‌شود با توجه به بافت حرفی مانند جدول ۳-۵ استفاده می‌شود. این وند اسم فارسی را شناسه می‌کند و فقط منحصر به ساخت غیر رسمی است.

^۱ واکه ۵ قبل از واژه‌بست شخصی می‌تواند حذف شود (اختیاری). ساخت و ساز آن در قسمت قواعد نگارشی انجام می‌شود.

جدول ۳-۵ نگارش مختلف وند معرفه

وند	مثال	حرف
ه*	کتابه	همخوان
هه / هعه	لبوئه	واکه واو
ههه / هعه	پنبههه	واکه ه
هههه / هعه	بهائه	واکه الف
ههههه	نظميه	واکه ی

۲-۵-۵ اضافه

كسره اضافه طبق جدول ۴-۵ به کلمات افزوده می‌شود.

جدول ۴-۵ حروف نمایشنده‌ند تکواز اضافه

نمایشن تکواز اضافه		
ه (خطای نگرشی)	ی	حرف پایانی
كتابه	-	همخوان
-	بدگویی	واکه واو
-	بردهی	واکه ه
-	هیولا	واکه الف
نظمیه	-	واکه ی

۶-۲-۵ واژه‌بست ربطی

این واژه‌بست‌ها با توجه به بافت حرفی‌ای که در آن ظاهر می‌شوند در جدول ۵-۵ آمده‌اند. این واژه‌بست ساخت اسنادی را به اسم‌ها و غالباً سایر قسم کلمه‌هایی که در این بخش آمده است اضافه می‌کند.

جدول ۵-۵ واژه‌بستهای ربطی

شمار	رسمی بودن	همخوان	واکه او	واکه ه	واکه الف	واکه ی
۱	رسمی	مریضم: *م	بدگوام: *ام	بردهام: ×ام	هیولاام: *ام	نظمیم: *م نظمیام: ×ام
	غیر رسمی		بدگوام: *ام	بردهم: ×م	هیولايم: *یم هیولاام: *م	
۲	رسمی	مریضی: *ی	بدوگوای: *ای	بردهای: ×ای	هیولاای: *ای	نظمیای: ×ای
	غیر رسمی		بدگویی: *بی		هیولاای: *بی	
۳	رسمی	مریضه: *ه	بدگوست: *ست	بردهست: ×ست	هیولات: *ست	نظمیست: *ست نظمیه: *ه
	غیر رسمی		بدگوئه: *ئه	بردنس: ×س	هیولاس: *س	
۴	رسمی	مریضیم: *یم	بدگوایم: *ایم	بردهایم: ×ایم	هیولاایم: *ایم	نظمیایم: ×ایم نظمییم: *بیم
	غیر رسمی		بدگوییم: *بیم	بردهیم: ×بیم	هیولایم: *بیم	
۵	رسمی	مریضید: *ید	بدگواید: *اید	بردهاید: ×اید	هیولااید: *بید	نظمیاید: ×اید نظمیبید: *بید
			بدگویید: *بید	بردهبید: ×بید		
۶	غیر رسمی	مریضین: *ین	بدگواین: *این	بردهاین: ×این	هیولااین: *ین	نظمیاین: ×این نظمیین: *بین
			بدگوین: *بین	بردهبین: ×بین		
۷	رسمی	مریضند: *ند	بدگواند: *اند	بردهاند: ×اند	هیولايند: *بند	نظمیند: *ند نظمیاند: ×اند
			بدگوند: *ند	بردهند: ×ند		
۸	غیر رسمی	مریضن: *ن	بدگوان: *ان	بردهان: ×ان	هیولان: *ن	نظمین: *ن نظمیان: ×ان
			بدگون: *ن	بردهن: ×ن		

۷-۲-۵- وندها و واژه‌بستهای محاوره

وندها و واژه‌بستهای محاوره شامل تاکید، عطف، را و هم است که در انتهایی ترین قسمت کلمات به ساختار افزوده می‌شوند و در جدول ۵-۶ آمده‌اند.

جدول ۶-۵ وندها و واژه‌بست‌های محاوره

حروف پایانی	ها (تاكيد)*	ا (تاكيد)*	و (اعطف)	رو (را)	و (را)	رم (را+هم)	م (هم)
همخوان	کتابها: *ها	کتاب: *ا	کتابو: *و	کتابرو: *رو	کتابو: *و	کتابارم: *رم	کتابام: *م
واكه واو	لبوها: *ها	لبو: *ا	لبوو: *و	لبورو: *رو	لبوو: *و	لبورم: *رم	-
واكه ی	نظميهها: *ها	نظميا: *ا	نظميو: *و	نظميرو: *رو	نظميو: *و	نظميرم: *رم	نظميم: *م
واكه الف	کتابها: *ها	-	کتابهao: *و	کتابارو: *رو	کتابهao: *و	کتابارم: *رم	کتابام: *م
واكه ھ	پنبهها: *ها	*ها	پنبهرو: *رو	پنبهو: *و	پنبهرو: *رو	پنبههرو: *رم	-

* طبق شفاقی (۱۳۹۴)، تاکید یک واژه‌بست است، اما علت اینکه در اینجا به عنوان وند آمده است، این است که در پژوهش شفاقی این تکواز تنها منحصر به فعل فرض شده است که در آن وزن کلمه را به خود جذب نمی‌کند (از میان چند آزمون مختلف برای تشخیص وند بودن یا واژه‌بست بودن) و در این پژوهش علاوه بر فعل و ساخت‌های اسنادی به کلمات غیر فعلی نیز می‌تواند متصل شود. در اتصال به اسم‌ها، این تکواز تکیه کلمه را جذب می‌کند. بنابراین از نظر واژه‌بست بودن ظاهرا این تکواز دو رفتار متفاوت در افعال و کلمات غیر فعلی از خود بروز می‌دهد. البته بررسی نظری این مورد در حوزه پژوهش فعلی نیست و صرفاً شناسایی آن در این پژوهش اهمیت دارد. گرچه این تکواز در این پژوهش به عنوان وند در نظر گرفته شده است، استفاده کنندگان از این تحلیل‌گر به قطعه +تاکید در قواعد خروجی تحلیل‌گر، هر عنوانی که مناسب می‌دانند می‌توانند اطلاق کنند.

۸-۲-۵- واژه‌بست ربطی و یژه

در ساخت‌های اسنادی محاوره واژه‌بست‌های ویژه‌ای وجود دارد که در دو ساخت مختلف استفاده می‌شود. این واژه‌بست غیر از بکار رفتن تک حرف **ش** در همه ساخت‌های آن در ساخت سوم شخص مفرد هم با واژه‌بست ربطی معمولی متفاوت است. این ساخت ویژه در جدول ۷-۵ آمده است. ساخت سوم شخص مفرد بدون هیچ واژه‌بستی (تمهی) نیز می‌تواند ساخته شود.

جدول ۷-۵ واژه‌بست‌های ربطی و بیزه

شمار	واژه‌بست	مثال
ربطویزه ۱	*شم	کوشم، ایناهاش
ربطویزه ۲	*شی	کوشی، اوناهاشی
ربطویزه ۳	*ش، *شش، Θ	کو، کوش، کوشش، ایناها، ایناهاش، ایناهاشش
ربطویزه ۴	*شیم	کوشیم، اوناهاشیم
ربطویزه ۵	*شید، *شین	کوشین، اوناهاشین
ربطویزه ۶	*شن، *شند	کوشن، ایناهاشن

۹-۲-۵- تک واژه منتهی به واکه

در کلمات غیر رسمی امکان حذف یا تغییر حروف واکه در انتهای تک واژه‌ها بخصوص در هنگام اتصال به تک واژه دیگر بسیار زیاد است. مثلاً کلمه **همشون** از دو تک واژه **همه+شون** تشکیل شده است. یا وقتی اسم **ستاره** با جمع **ان جمع** بسته می‌شود، واکه **۵** پایانی حذف شده و حرف **گ** جایگزین آن می‌شود و اسم جمع **ستارگان** ساخته می‌شود. یا صفت **خسته** همراه با واژه‌بست ربطی غیر رسمی سوم شخص مفرد با حذف **ه** تبدیل به کلمه **خستس** می‌شود. همینطور واژه‌بست‌ها و وندها بعد از همخوان و واکه‌ها با شکل‌های متفاوتی افزوده می‌شوند. این موضوع در زیربخش‌های همین قسمت مشخص است (مثل **وربطی، وشخصی، معرفه و غیره**).

بنابراین تمامی واژگان غیر فعلی با مشخص شدن حرف پایانی آنها که همخوان و یا واکه‌ای بخصوص است مشخص شده است. همینطور هر وند یا واژه‌بستی که به ریشه اصلی می‌پیوندد در صورتی که منتهی به واکه خاصی باشد به ساخت جدید کلمه افزوده می‌شود و به این صورت علامت واکه ریشه را پاک می‌کند. این روند تا مرحله پایانی ساخت کلمه ادامه پیدا می‌کند.

شناسایی واکه‌های پایانی واژگان با استخراج تلفظ واژگان زایا (اسلامی و همکاران، ۱۳۸۳) امکان پذیر شده است. این واژگان برای تک‌تک کلمات تلفظ آن را نیز به همراه دارد. بنابراین افزودن نشان برای کلمات منتهی به واکه با استفاده از اطلاعات واژگان زایا ممکن شده است.

۳-۵- ساختمان‌های غیر فعلی

ساختمان اسم و سایر کلمه‌های این بخش که به صورت جزئی و یا کامل از ساختمان اسم بهره می‌برند در این قسمت قرار دارد.

۱-۳-۵- اسم

ساختمان اسم رسمی آنطور که در واژگان زایی زبان فارسی (محرم اسلامی و همکاران، ۱۳۸۳) تعریف می‌شود، (با اندکی تغییر) به صورت زیر است.

$$\left[\begin{array}{c} (\text{نکره}) \\ (\text{موصولی}) \\ (\text{وشخصی}) \\ (\text{اضافه}) \end{array} \right] + \left[\begin{array}{c} (\text{وربطی}) \end{array} \right] + \text{اسم} + (\text{جمع})$$

مثال: کتاب، آلدگیست، آدم‌هاییند (که).

ساختمان جامع اسم‌های غیر رسمی نیز به صورت زیر است. این ساختمان تمامی ساختهای ممکن اسم غیر رسمی و سایر قسم کلمه‌هایی را که در این بخش از ساختمان اسم به صورت جزئی یا کلی استفاده می‌کنند، پوشش می‌دهد.

$$\left[\begin{array}{c} \left[\begin{array}{c} (\text{تاكيد}) \\ (\text{هم}) \\ (\text{اعطف}) \\ (\text{اضافه}) \end{array} \right] \\ + \left[\begin{array}{c} (\text{ربطی}) \\ (\text{را}) \end{array} \right] \end{array} \right] + \left[\begin{array}{c} \left(\begin{array}{c} (\text{جمع}) \\ (\text{نکره/موصولی}) \end{array} \right) + (\text{شخصی}) \\ (\text{معرفه}) \end{array} \right] + \text{اسم}$$

مثال: (-همه طرفداراشون سطح پایین! +نه! اون) طرفدارایشون (که اهل مطالعه‌ن، خوبن)، کتابارم (بیر)، خونه‌شونا. البته مسیرهایی در ساختمان بالا وجود دارد که هرگز در ساختمان اسم بکار نمی‌رود. برای محدود کردن این ساخت و بالا بردن درستی^۱ اسمی ساخت مانع اسمی را تعریف می‌کنیم و در عوض از آن استفاده می‌کنیم. ساختمان مانع باید در عین داشتن جامعیت ساخت بالا، محدودیت‌هایی برای بستن مسیرهای خطای آن داشته باشد.

^۱ Precision

$$\begin{aligned}
 & \left[\left[\left(\begin{array}{c} (\text{تاكيد}) + (\text{ربطي}) \\ (\text{تاكيد}) + (\text{را}) \\ (\text{هم}) \\ (\text{عطف}) \end{array} \right) + \text{شخصي} \right] \right] \\
 & + \dots \left[\left[\begin{array}{c} (\text{ربطي}) \\ (\text{هم}) \\ (\text{را}) \\ (\text{هم}) \end{array} \right] + \text{شخصي} \right] + \text{موصولي} \\
 & + \left[\left[\begin{array}{c} (\text{تاكيد}) \\ (\text{ربطي}) + (\text{ربطي}) \\ (\text{عطف}) \end{array} \right] + (\text{نكره} + \text{ربطي}) \right] + \text{(جمع)} \\
 & + \left[\begin{array}{c} *(\text{هم}) \\ (\text{تاكيد}) \\ (\text{ربطي}) + (\text{تاكيد}) \\ (\text{عطف}) \\ (\text{اضافه}) \end{array} \right] + \text{اسم} \\
 & + \left[\left[\begin{array}{c} (\text{تاكيد}) \\ (\text{ربطي}) \end{array} \right] + (\text{را}) \right] + \text{(معرفه)}
 \end{aligned}$$

* ساخت را+هم (که با هم واژه‌بست هم را می‌سازد) در دو حالت تولید می‌شود؛ اول در حالتی که کلمه وند معرفه بپذیرد، در این صورت اگر این وند در نگارش هم نوشته نشود، وجود آن به علت جذب تکیه کلمه مشهود است، دوم در حالتی که وند جمع به اسم مفرد افزوده شده باشد. در واقع این ساخت ظاهرا تنها محدود به اسم جمع و معرفه است. هر کدام به صورت تکی تکی می‌توانند برای مفرد بکار روند.

ساختهای دیگر این فصل به صورت جزئی یا کلی از این ساخت استفاده می‌کنند.

ساختمان اسم شامل اسم جا، شخص و فamilی هم می‌شود (موجودیت یا اسم خاص).

۵-۳-۲- صفت

ساختمان صفت رسمی نیز بر اساس واژگان زایا در ادامه آمده است. این ساختمان با پذیرفتن وندهای اختیاری تفضیلی و عالی تولید می‌شود. در عین حال با اشتقاق صفر می‌تواند به اسم تبدیل شود و از ساختمان تصریفی اسم بهره ببرد.

$$\text{صفت} + \left[\begin{array}{c} (\text{نکره}) \\ (\text{ربطی}) \\ (\text{موصولی}) \\ (\text{شخصی}) \\ (\text{اضافه}) \end{array} \right] + (\text{جمع}) + \left[\begin{array}{c} (\text{تکواز صفت تفصیلی ساز}) \\ (\text{تکواز صفت عالی ساز}) \end{array} \right]$$

مثال: زیبا، بهترین ست، فقیرند.

ساختمان غیر رسمی صفت هم به این شکل است.

$$\left[\begin{array}{c} \text{(تاكيد)} \\ \text{(ربطي)} \\ \text{(هم)} \\ \text{(را)} \\ \text{(عطف)} \\ \text{(اضافه)} \end{array} \right] + \left[\begin{array}{c} \text{(جمع)} \\ \text{(نكره / موصولی)} \\ \text{(شخصی)} \\ \text{(معرفه)} \end{array} \right] + \left[\begin{array}{c} \text{(تکواز صفت تفصیلی ساز)} \\ \text{(تکواز صفت عالی ساز)} \end{array} \right] + \text{صفت}$$

مثال: بهترینشونه، پدیده‌ها، دیوونه‌س، خسته‌ام (هم).

این ساخت به صورت محدودتر همان ساخت مانع اسامی است با این تفاوت که تک واژه‌ای صفت عالی‌ساز و تفضیلی‌ساز نیز در آن یکار می‌روند.

٥-٣-٣-٣-٣

ساختار رسمی قید مختص طبق دستور زبان فارسی (حسن احمدی گیوی و حسن انوری، ۱۳۹۱) هیچگون وندی نمی‌پذیرد و به شکل انفرادی بکار می‌رود. مثال: ابداً، مثلاً، عجولانه.

قید مختص غیر رسمی می‌تواند وند تاکید و واژه‌بست هم نیز بپذیرد.

تاكید + قید

مثل: بعذنهم (بعدا هم)، بدقتا، اصلن.

۳-۴- صفت یا خمیر اشاره

تفاوت صفت و ضمیر اشاره در این است که صفت اشاره همراه با اسم بعد از خود بکار می‌رود اما ضمیر به شکل مستقل بکار می‌رود و جای اسم را می‌گیرد. از این حیث شناسایی آنها تنها در جمله و با توجه به بافت قابل تشخیص است. بنابراین تحلیل گر تصrifی مستقل از متن قادر به تمیز آنها نیست. اما در صورتی که هرگونه ساخت تصrifی مرتبط بپذیرد، در اینصورت ضمیر اشاره است. ساخت رسمی ضمیر اشاره در ادامه آمده است. این ساختار تفاوت چندانی با ساختار اسم ندارد، تفاوت ساخت اضافه و ساخت معرفه است که در ضمیر اشاره وجود ندارد.

$$+ \begin{bmatrix} (\text{نکره}) \\ (\text{موصولی}) \\ (\text{شخصی}) \end{bmatrix} + (\text{جمع}) + \text{اشاره}$$

مثال: این، آن، اینانند، آن‌هایشانند.

ساخت غیر رسمی نیز مانند اسم غیر رسمی است، و علاوه بر ساخت اضافه، ساخت معرفه نیز در آن وجود ندارد.

$$\begin{bmatrix} (\text{تاكيد}) \\ (\text{هم}) \end{bmatrix} + \begin{bmatrix} (\text{ربطي}) \\ (\text{را}) \\ (\text{عطف}) \end{bmatrix} + (\text{جمع}) + (\text{نکره}/\text{موصولی}) + (\text{شخصی}) + \text{اشاره}$$

مثال: ایناییشونم (که)، اونارم، اینارو، همینایی (که).

ساخت مانع صفت / ضمیر اشاره نیز به صورت زیر است.

$$\left[\begin{array}{c} \left(\begin{array}{c} (\text{تاكيد} + \text{ربطي}) \\ (\text{را} + \text{تاكيد}) \\ (\text{هم}) \\ (\text{عطف}) \end{array} \right) + \text{شخصی} \\ \dots \\ \left[\begin{array}{c} \left(\begin{array}{c} (\text{ربطي}) \\ (\text{هم}) \end{array} \right) + \text{شخصی} \\ (\text{را} + \text{هم}) \end{array} \right] + \text{موصولی} \\ \left(\begin{array}{c} (\text{تاكيد}) \\ (\text{عطف}) \end{array} \right) + (\text{نکره} + \text{ربطي}) \\ \left[\begin{array}{c} *(\text{هم}) \\ (\text{را}) \end{array} \right] + (\text{تاكيد}) \\ (\text{ربطي}) + (\text{تاكيد}) \\ (\text{عطف}) \end{array} \right] + \text{اشاره} + (\text{جمع})$$

* ساخت را+هم (که با هم واژه‌بست رهم را می‌سازد) تنها در صورتی که کلمه مفرد نباشد و یا خاصیت مفرد بودن برای آن وجود نداشته باشد و یا وند جمع پذیرفته باشد استفاده می‌شود. وند جمع خاصیت مفرد بودن این اشاره‌ها را از بین می‌برد.
مثال: اینو (این را)، اینم (این+هم)، اینارم (این+جمع+را+هم). این‌همرم (این‌همه+را+هم). هر کدام از این واژه‌بست‌ها به صورت جدا می‌توانند برای مفرد بکار روند.

اشاره‌های مفرد: این، آن، همین، همان، این قدر، اینقدر، انقدر، آن دیگری، آن یکی، این یکی، اینجور، اینطور، همینطور، همینجور، همانطور، همانجور، همونجور.

اشاره های غیر مفرد: این همه، آن همه، آن گونه

بخشی از ضمایر نیز ساخت تصریفی، محدودتری دارد.

بخشی از ضمایر نیز ساخت تصریفی محدودتری دارند

اشاره + ربطی + تاکید

مثال: چنینند، این چنینم (هم).

وازگان اشاره‌های این ساخت استثنائی شامل این موارد می‌شود: چنین، چنان، این‌چنین، آن‌چنان، این‌سان.

مورد استثنایی دیگر نیز ضمایر اشاره این و اوⁿ است که به شکل ویژه‌ای صرف می‌شود.

این / اون + ا + ها + (وربٹویژه)

مثال: اپناها، اوناهاش، اوناهاشش، اپناهاشن.

وازگان این ساخت استثنائی تنها محدود به این و اون است.

۳-۵- صفت یا ضمیر مبهم

مانند صفت و خصایر اشاره، صفت و ضمیر مبهم نیز میتوان به شکل انفرادی و یا در ساختمان تصریف به کار رود.

$$\left[\begin{array}{c} \text{مبهمن} \\ + \\ \text{شخصی (ربطی)} \\ + \\ \text{(اضافه)} \end{array} \right]$$

هر کس، دیگر، هیچ‌جیز ند.

ساختار غیر سممی میهمان نیز به این صورت است.

مبهم + $\left[\begin{array}{c} \left(\begin{array}{c} \left(\begin{array}{c} \left(\begin{array}{c} \text{اعطف} \\ \left(\begin{array}{c} \text{اضافه} \end{array} \right) \end{array} \right) + \left(\begin{array}{c} \text{ربطى} \\ \left(\begin{array}{c} \text{تاكيد} \end{array} \right) \end{array} \right) \\ \left(\begin{array}{c} \text{هم} \\ \left(\begin{array}{c} \text{تاكيد} \end{array} \right) + \left(\begin{array}{c} \text{را} \\ \left(\begin{array}{c} \text{هم} \end{array} \right) \end{array} \right) \end{array} \right] + \text{شخصى} \end{array} \right]$

* کلماتی که می‌وصولی می‌پذیرند می‌توانند ساخت را+هم تولید کنند. در غیر اینصورت از تک تک آنها جداگانه می‌توانند استفاده کنند ولی، با هم این امکان وجود ندارد.

مثال: هرکیم، هیشکیو، هرچیرم(را+هم).

واژگان مبهم به این شرح است: دیگه، هیشکس، هیشکسی، هیشکی، چنتا، چنتایی، هیچیز، خیلیا (جمع)، دیگر، کسی، هرکس، فلان، بهمان، هیچ کس، هیچ چیز، چندتا، یکی، یکیا (جمع)، هرکی، هرچی، خیلی، کمی، قدری، لختی، اندکی، بسیاری.

مبهم ساخت استثنائی ای هم دارد که هیچ ساخت تصریفی ای نمی‌پذیرد و مستقل استفاده می‌شود. واژگان این ساخت به این شرحند: هر، هرکه، هرچه، چیز، چندیدن، چندان، همگان، دیگران، هیچ.

۶-۳-۵- صفت یا ضمیر پرسشی

صفت و ضمیر پرسشی رسمی به طریق زیر ساخته می‌شود.

$$\left[\begin{array}{c} \text{(نکره)} \\ \text{+ (ربطی)} \\ \hline \text{(شخصی)} \end{array} \right] + \text{پرسش + (جمع)}$$

مثال: کدام، کدامی، کدامهایشانند.

ساختار غیر رسمی آن نیز به شکل زیر است.

$$\left[\begin{array}{c} \left[\begin{array}{c} \text{(ربطی)} \\ \text{+ (را)} \\ \hline \text{(هم)} \end{array} \right] + \text{شخصی} \\ \left[\begin{array}{c} \text{(نکره + (ربطی))} \\ \text{* \quad \quad \quad (را) + (هم)} \\ \hline \text{(ربطی) + (تایید)} \end{array} \right] \end{array} \right] + \text{پرسش + (جمع)}$$

* مانند ساختار اسم‌ها و اشاره، در اینجا نیز ساختهای مفرد نمی‌توانند ساخت را+هم را با هم بکار ببرند.

مثال: کدومیشونه، چندمی، کدومیاشونی.

واژگان پرسش به این شرحند. هر کدام به صورت تکی تکی می‌توانند برای مفرد بکار روند.

- پرسش‌های مفردی که می‌توانند وند جمع بپذیرند: کدام، کدوم، چندم.
- پرسش‌های مفردی که نمی‌توانند وند جمع بپذیرند: کدامین، چندمین، کدامیک.
- پرسش‌های غیر مفردی که نمی‌توانند وند جمع بپذیرند: چقدر، چه قدر، چقد، چه مقدار، چه اندازه‌ای.
- پرسش‌هایی که مفرد به حساب نمی‌آیند اما می‌توانند وند جمع نیز بپذیرند: کجا، چی، کی.

برخی ضمایر پرسشی تنها با واژه‌بست ربطی بکار می‌روند.

پرسش + (ربطی)

واژگان این ساخت استثنایی به این شرح است: چگونه، چطور، چه جور، کی، چه، که، چسان، چند، چطوری، چجوری، چندمی، کدامی، کدومی، کدامیکی، چقدری، چندی.
مثال: چگونه، چطورند، چجورین.

ساختار استثنایی دیگری نیز وجود دارد که تنها یک واژه **کو** دارد اما از ساختار واژه‌بست ربطی ویژه استفاده می‌کند.
ضمیر پرسشی کو + واژه بسط ربطی ویژه(وربطویژه)

مثال: کو، کوشش، کوشی، کوشن.

ساختار استثنایی نهایی نیز تک پرسش آیا است که به شکل مستقل استفاده می‌شود.

۷-۳-۵- صفت یا ضمیر تعجبی

صفت یا ضمیر تعجبی محدود به چند واژه محدود است که همه آنها غیر از یکی در واژگان پرسشی وجود دارد. بسته به بافت می‌توان از میان پرسشی‌ها آنها را شناسایی کرد. واژگان تعجب به این شرحند: چه، چقدر، چقد.
واژه دیگر تعجبی کلمه **عجب** است که در واژگان **جملک** قرار می‌گیرد و می‌توان با توجه به بافت آن را از جملک تمیز داد.

۸-۳-۵- ضمیر شخصی

ساخت رسمی ضمایر به صورت زیر است.

$$\left[\begin{array}{c} (\text{موصولی}) + (\text{ربطی}) \\ (\text{اضافه}) \end{array} \right] + \text{شخصی}$$

مثال: ایشان، او، تویی، منی (که)، اوست.

ساختار غیر رسمی ضمیر شخص به طریق زیر است.

$$\left[\begin{array}{c} (\text{موصولی}) + (\text{ربطی}) \\ ((\text{را}) + (\text{هم})) + (\text{که}) + \dots \end{array} \right] + \left[\begin{array}{c} (\text{رو) + (را}) \\ (\text{تمکید}) \\ (\text{تاكید}) \end{array} \right] + \left[\begin{array}{c} (\text{ربطی}) + (\text{تمکید}) \\ (\text{اعطف}) \\ (\text{اضافه}) \end{array} \right] + \text{ضمیر شخصی}$$

* را+هم مختص ضمایری است که مفرد نیستند. هر کدام به صورت تکی تکی می‌توانند برای مفرد بکار روند.
مثال: ماهایی (که)، شماهایید، منما، شمارم (را+هم).

واژه‌های ضمیر شخصی مفرد: ایشون، ایشان، اینجانب، من، شما، ما.
واژه‌های ضمیر شخصی غیر مفرد: شماها، ماها، اینجانبان، تو (استثنائاً)، او، وی، همو.

۹-۳-۵- ضمیر مشترک

ضمیر مشترک تنها از یک کلمه تشکیل می‌شود که ساختهای تصریفی رسمی و غیر رسمی زیر را می‌پذیرد.

ضمیر مشترک (خود) + (شخصی) + (وربطی)

مثال: خودم، خودشانند.

ساختار غیر رسمی ضمیر مشترک نیز به این صورت است.

$$\left[\begin{array}{c} ((ربطی) + (تاكيد)) \\ ((را) + (تاكيد)) \\ ((هم)) \\ ((عطف)) \end{array} \right] + خود + (شخصی)$$

مثال: خودشم، خودشون، خودتی.

۱۰-۳-۵- حرف اضافه

حرف اضافه آنطور که در دستور زبان فارسی تعریف می‌شود به صورت زیر ساخته می‌شود.

حرف اضافه + (شخصی) + (ربطی)

مثال: با، بدین، باکمک.

ساخت غیر رسمی حرف اضافه نیز به صورت زیر است.

$$\left[\begin{array}{c} ((ربطی) + (تاكيد)) \\ ((هم)) \end{array} \right] + حرف\ اضافه + (شخصی)$$

مثال: باهاشون، برامون.

۱۱-۳-۵- شماره

ساختار اعداد دو گونه است

شماره + (ترتیبی ۱ + (ترتیبی ۲) + (جمع)

مثال: دوم، سومین، چهارمی‌ها

ساختار دیگر اعداد به شرح زیر است.

شماره + تا + (جمع)

مثال: دوتا، سه‌تا.

۱۲-۳-۵- حرف ربط

حرف ربط هیچ ساختار تصریفی‌ای نمی‌پذیرد و بشكّل انفرادی استفاده می‌شود. حرف ربط شامل دو دسته حرف ربط ساده و حرف ربط گروهی می‌شود.

مثال: اگر، آنکه، اما، از این گذشته.

۱۳-۳-۵- مصدر

مصدرهای فعل‌های ساده و پیشوندی در این قسمت ساخته می‌شوند.

$$\left[\begin{array}{c} \text{(نکره)} \\ \text{(موصولی)} \\ \text{(شخصی)} \\ \text{(اضافه)} \end{array} \right] + \left[\begin{array}{c} \text{(ربطی)} \end{array} \right] + \text{(بن فعل ماضی} + \text{n} + \text{(جمع)} + \text{(پیشوند)} + \text{(منفی)})$$

مثال: برنگشتن، رفتن‌ها، خوابیدنشان.
ساختار غیر رسمی مصدرها به شرح زیر است.

$$\left[\begin{array}{c} \left[\begin{array}{c} \text{(تاکید)} \\ \text{(هم)} \end{array} \right] + \left[\begin{array}{c} \text{(ربطی)} \\ \text{(را)} \end{array} \right] \end{array} \right] + \left[\begin{array}{c} \text{(جمع)} + \text{(نکره/موصولی)} + \text{(شخصی)} \\ \text{(معرفه)} \end{array} \right] + \text{(بن فعل ماضی} + \text{n} + \text{(پیشوند)} + \text{(منفی)})$$

مثال: پاشدن، واينستادن، ننداختن‌ها، ورومدمشون.

۱۴-۳-۵- صفت مفعولی

ساختمان رسمی صفت مفعولی به صورت زیر است.

$$\left[\begin{array}{c} \text{(نکره)} \\ \text{(موصولی)} \\ \text{(شخصی)} \\ \text{(اضافه)} \end{array} \right] + \text{(بن فعل ماضی} + \text{ه} + \text{(جمع)} + \text{(پیشوند)} + \text{(منفی)})$$

مثال: رفته، نگفته‌هایشان، خورده‌هایت.
ساختمان غیر رسمی صفت مفعولی به شکل زیر است.

$$\left[\begin{array}{c} \left[\begin{array}{c} \left(\begin{array}{c} \text{(تاكيد)} \\ \text{(هم)} \end{array} \right) + \left(\begin{array}{c} \text{(ريطي)} \\ \text{(را)} \end{array} \right) \end{array} \right] \\ \left[\begin{array}{c} \left(\begin{array}{c} \text{(عطف)} \\ \text{(اضافه)} \end{array} \right) \end{array} \right] \end{array} \right] + \left[\begin{array}{c} \left(\begin{array}{c} \text{(شخصي)} \\ \text{(معرفه)} \end{array} \right) + \left(\begin{array}{c} \text{(جمع)} \\ \text{(نكره/موصولي)} \end{array} \right) + \left(\begin{array}{c} \text{(بن فعل ماضي)} \\ \text{(منفي)} \end{array} \right) + \left(\begin{array}{c} \text{(پيشوند)} \\ \text{هـ} \end{array} \right) \end{array} \right]$$

مثال: دررفتههاش، درومدهه، بازنداشته.

۱۵-۳-۵ - شبه جمله

شبه جمله ساختار تصریفی ندارد و به شکل مستقل استفاده می‌شود.

مثال:

رسمی: شگفتا، عجب، امان.

غیر رسمی: جون، زکی، هان.

۱۶-۳-۵ - شاخص

ساختار رسمی شاخص از ساختار رسمی اسم و ساختار غیر رسمی شاخص از ساختار غیر رسمی اسم استفاده می‌کند.

مثال: عمه‌ام، بابایت، عمه‌ها، سرهنگم، اوستاما.

۴-۵ - قواعد نگارشی

قواعد نگارشی بخش غیر فعلی نیز مانند بخش افعال عمل می‌کند. این قواعد برای کلمات غیر رسمی همانطور که در این فصل در بخش‌های مربوطه مشاهده شد با استفاده از نویسه‌های مرزنمای جدول ۸-۵ مشخص شده است.

جدول ۸-۵ مرزنماهای به کار رفته در قواعد نگارشی کلمات غیر فعلی

کارکرد	نویسه
دو تکواز را بدون هیچ فاصله‌ای به هم متصل می‌کند.	*
دو تکواز را با نیم‌فاصله به هم متصل می‌کند.	×
بین دو تکواز یک فاصله کامل قرار می‌دهد.	^
در بین تکواز منتهی به واکه ۵ و واژه‌بست شخصی قرار می‌گیرد. یکبار واکه ۵ قبل از خود را حذف می‌کند و بدون فاصله واژه‌بست شخصی را به تکواز قبل از خود می‌چسباند و یکبار دیگر بدون حذفی خود تبدیل به نیم‌فاصله می‌شود. مثل، همه+شون؛ همشون، همهشون (مختص به کلمات غیر فعلی). در بین تکواز منتهی به واکه ۵ و واژه‌بست ربطی سوم شخص مفرد می‌قرار می‌گیرد. یکبار واکه ۵ قبل از خود را حذف می‌کند و واژه‌بست را بدون هیچ فاصله‌ای به تکواز قبلی می‌چسباند و یکبار دیگر بدون حذفی خود تبدیل به نیم‌فاصله می‌شود. مثل، خسته+س؛ خستس، خسته‌س (مختص به کلمات غیر فعلی). در بین تکواز منتهی به واکه ۵ و ند جمع جانداران قرار می‌گیرد و واکه ۵ را حذف می‌کند حرف گ را جایگزین آن می‌کند. مثل، ستاره+ان؛ ستارگان، بنده+ان؛ بندگان (مختص به کلمات غیر فعلی).	☒

شیوه به کار رفتن این مرزنماها در این فصل، در جدول ۱-۵ وندهای نکرگی / موصولی، جدول ۲-۵ واژه‌بستهای شخصی، جدول ۳-۵ نگارش مختلف وند معرفه، جدول ۴-۵ حروف نمایش‌دهند تکواز اضافه، جدول ۵-۵ واژه‌بستهای ربطی، جدول ۶-۵ وندها و واژه‌بستهای محاوره و جدول ۷-۵ واژه‌بستهای ربطی ویژه مشخص شده است.

۵-۵-۵- جمع بندی

ساختار رسمی و غیر رسمی کلمات غیر فعلی در این فصل تعریف شد. عمدۀ ساختارهای رسمی مبتنی بر واژگان زایا (اسلامی و همکاران، ۱۳۸۳) تعریف شده است. برای واژگان غیر رسمی نیز این کار با استفاده از مگردویان (۲۰۰۸) و انجام پژوهش و استخراج قواعد از پیکره جمع آوری شده برای این منظور انجام پذیرفت. سعی شده است تمام استثنائات و قواعد ویژه‌ای که در فارسی غیر رسمی وجود دارد نیز شناسائی و در این فصل گنجانده شود. قواعد نگارشی کلمات غیر فعلی نیز برای کلمات فارسی معاصر در این فصل تعریف شده است.

فصل ششم

روش پیشنهادی؛ ساختمان مبدل‌ها

۱-۶- مقدمه

تقسیم کار در ساختمان تحلیل‌گر تصریفی به این صورت است که تحلیل تصریفی صرف در یک مبدل مجزا انجام می‌شود و سایر تغییرات آوایی / نگارشی برای تبدیل / نزدیک کردن کلمه به شیوه استاندارد شده در قواعد تحلیل‌گر، در مبدل‌های دیگر انجام می‌شود. دلیل این تقسیم‌بندی مدیریت ساده‌تر قواعد و ساختارهای است و استفاده اختیاری از هر یک در صورت نیاز به آن امکان پذیر می‌شود.

قاعده‌های سازنده کلمات در مبدل استاندارد، پایه و مبنای مبدل‌های دیگر نیز هست. در مبدل‌های دیگر علاوه بر قاعده‌های سازنده کلمات، قاعده‌های ثانویه برای تبدیل / نزدیک کردن شکل آوایی به کلمات به کار می‌رود.

ابزار جستجوی فوما امکان دو نوع استفاده از چند مبدل را می‌دهد؛ یکی وصل کردن خروجی مبدل اول به ورودی مبدل دوم و الی آخر که به دلیل مستقل بودن مبدل‌های تحلیل‌گر امکان استفاده از این حالت فراهم نیست. دیگری وارد کردن کلمه به مبدل اول و توقف پردازش در صورت خروجی گرفتن از آن، در غیر این صورت رفتن کلمه به مبدل بعدی و در صورت تولید خروجی توقف پردازش در این مرحله و در غیر این صورت کلمه ادامه همین فرایند تا جایی که یا مبدلی خروجی تولید کند و یا مبدلی در چرخه باقی نمانده باشد.

حالت سومی در این ابزار وجود ندارد، مگر اینکه با استفاده از کتابخانه‌های طراحی شده و یا کد شخصی بر روی ساختمان داده مبدل جستجو انجام دهیم. شیوه مطلوب در این پایان‌نامه و شکل انجام ارزیابی به این صورت است که برای هر کلمه که به تحلیل مقبولی نرسیدیم با مبدل بعدی روی کلمه تحلیل انجام دهیم. در واقع از هر مبدل به شکل مستقل استفاده می‌کنیم.

۲-۶- مبدل استاندارد

قاعده‌های سازنده کلمه رسمی و غیر رسمی، فعل و غیر فعل همه در این مبدل قرار دارند. مرزنماهای فاصله، نیم‌فاصله و اتصال در این مبدل به صورت غیر قطعی^۱ می‌توانند جایگرین یکدیگر شوند. دلیل این کار پوشش همه حالت‌های نگارش کلمات است که هم در فارسی رسمی و هم غیر رسمی به صورت متنوع شایع است. در واقع در این مبدل همه فاصله‌گذاری‌ها برای عبارت در این صورت تولید می‌شود تا امکان شناسایی همه نگارش‌های آن وجود داشته باشد (شکل ۱-۶).

^۱ Non-deterministic

در این صورت

دراين صورت

دراين صورت

در اين صورت

دراين صورت

دراين صورت

شکل ۱-۶ همه حالت‌های تولید شده در مبدل

حروف‌های ۱ و آ نیز به صورت غیر قطعی تعریف می‌شوند. در نگارش کلمات دارای این دو حرف جایگزین شدن آنها بسیار اتفاق می‌افتد. این جایجایی ممکن است در هر دو گونه رسمی و غیر رسمی رخ دهد. این مبدل برای شناسایی و تحلیل تصrifی کلمات است و می‌بایست همه حالت‌ها مختلف کلمات را تولید کند. برای تولید کلمه بر اساس قاعده سازنده، مبدل و قواعد ساده‌تری طراحی شده است تا پیچیدگی‌های جزئیات قواعد شناسایی فرایند تولید کلمه را بیش از حد سخت نکند (۷-۶).

۶-۳- مبدل هم‌صدا

در این مبدل حروف هم‌صدا جایگزین یکدیگر می‌شود. در نگارش فارسی رسمی و غیر رسمی به دلیل خطای نگرشی و یا به عمد، این حروف می‌توانند جایگزین یکدیگر شود (جدول ۱-۶).

تفاوت دیگر این مبدل با مبدل استاندارد، تشخیص و تولید قاعده اضافه^۲ برای حرف ۵ در پایان ساخت اسمی و ساخت‌های غیر فعلی مرتبت با آن است. این ساخت به نوعی خطا تلقی می‌شود، بنابراین در مبدل استاندارد قرار نمی‌گیرد.

^۲ کسره اضافه

جدول ۶-۱ حروف هم‌گروه که میتوانند جایگزین یکدیگر شوند

حروف هم‌صدا
س، ص، ث
ت، ط
ذ، ز، ض، ظ
ح، ه
غ، ق
ع، ى، ی، ئ، ۋ، أ، إ

۶-۴- مبدل آوایی

ممکن است برخی کلمات غیر رسمی خارج از واژگان باشد. قواعد این قسمت می‌تواند کلمات رسمی واژگان را به معادل غیر رسمی‌شان نزدیک / تبدیل کند. برخی کلمات غیر رسمی نیز واحد واژگانی نیستند و تنها در نگارش ممکن است متفاوت از معادل رسمی‌شان باشند، این قواعد سعی در شناسایی آنها نیز دارد. قواعد این بخش در جدول ۶-۶ آمده است.

جدول ٦-٢ قواعد آوایی

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

۶-۵- مبدل پیانی

۶-۶- مبدل تقطیع

این مبدل کلمه را به دو قطعه می‌شکند. این تقطیع کلمه تنها یک بار اتفاق نمی‌افتد بلکه به تعداد حروف کلمه اصلی، قطعه تولید می‌شود (این تعداد شامل خود کلمه اصلی نیز می‌شود). هر کدام از قطعات تولید شده اگر در تحلیل گر قابلیت تصریف داشته باشد، در خروجی ظاهر می‌شود. در کلمه‌ها و قاعده‌های 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. البته هر جفت قطعه باید قاعده سازنده کلمه را تولید کرده باشند.

برای مثال کلمه **ازمردم**، قطعه‌های شکل ۲-۶ را تولید می‌کند.

ازمردم

ا-زمردم

از-مردم

ازم-ردم

ازمر-دم

ازمرد-م

شکل ۲-۶ قطعه‌های تولید شده از **ازمردم**

این قطعه‌ها ۲۳ قاعده مختلف تولید می‌کند(جدول ۳-۶)، اما تنها جفت کلمه‌(هایی) از بین آنها می‌توانند انتخاب شود که در شکل ۲-۶ باشد. جفت‌های مجاز در کنار هم در جدول ۳-۶ آمده است.

جدول ۶-۳ کلمه و قاعده‌های بدست آمده از تقطیع

کلمه ۱	قاعده ۱	کلمه ۲	قاعده ۲
از	<تقطیع: حضاف=از+رسمی>	مردم	<تقطیع: اسماع=مردم+رسمی>
از	<تقطیع: اسماع=آر+رسمی>	مردم	<تقطیع: ف.م.س=مرد+ش۱+رسمی>
		مردم	<تقطیع: اسماع=مرد+هم>
		مردم	<تقطیع: اسماع=مرد+وشخصی۱+رسمی>
		مردم	<تقطیع: اسماع=مرد+وربطی۱+رسمی>
		مردم	<تقطیع: ف.م.س=مرد+ش۳+هم>
		مردم	<تقطیع: صمغولی=مرده+وشخصی۱>
ازم	<تقطیع: حضاف=از+وشخصی۱+رسمی>	ردم	<تقطیع: صفت=رد+وربطی۱+رسمی>
ازم	<تقطیع: اسماع=آر+هم>	ردم	<تقطیع: صفت=رد+هم>
ازم	<تقطیع: اسماع=آر+وشخصی۱+رسمی>	ردم	<تقطیع: صفت=رد+وشخصی۱+رسمی>
ازم	<تقطیع: اسماع=آر+وربطی۱+رسمی>	ردم	<تقطیع: اسماع=رد+وشخصی۱>
		زمردم	<تقطیع: اسماع=زمرد+هم>
		زمردم	<تقطیع: اسماع=زمرد+وشخصی۱+رسمی>
		زمردم	<تقطیع: اسماع=زمرد+وربطی۱+رسمی>
		دم	<تقطیع: اسماع=دم+رسمی>
		م	<تقطیع: جملک=م+رسمی>
		م	<تقطیع: هم>

۶-۷- مبدل تولید

این مبدل برای دریافت قاعده و تولید کلمه استفاده می‌شود. به چند دلیل از مبدل استاندارد برای این منظور استفاده نشده است. جزئیات زیادی در قاعده سازنده کلمات وجود دارد که در مرحله شناسایی مفید است و قاعده‌ها را از یکدیگر متمایز می‌کند اما در مرحله تولید نوشتن قاعده را دشوار می‌کند.

استفاده از فاصله، نیم‌فاصله و اتصال در قواعد تصريفی به درستی رعایت شده است، در مرحله شناسایی از آنجایی که شناسایی نیازمند در نظر گرفتن همه حالات ممکن است، این مرز نماها به صورت غیر قطعی جایگزین یکدیگر می‌شوند.

بنابراین در مرحله تولید از روی قاعده اگر از چنین مبدلی استفاده شود به ازای هر فاصله، نیم‌فاصله و اتصال سه کلمه تولید خواهد شد. بنابراین در مبدل تولید این ویژگی غیر قطعی مرز ناماها از قواعد حذف شده است.

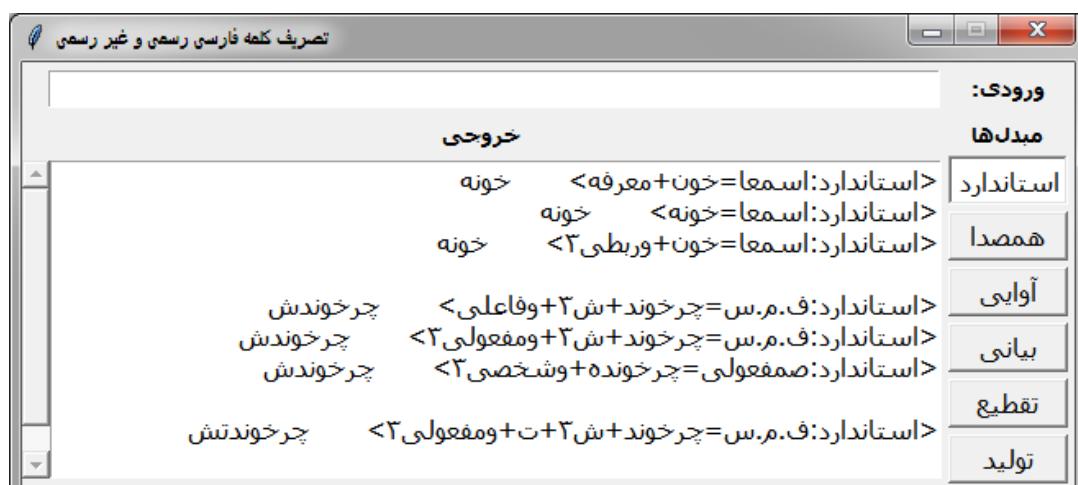
همینطور حرف‌های **ا** و **آ** که در مبدل استاندارد به صورت غیر قطعی جایگزین هم می‌شوند در اینجا از قواعد حذف می‌شوند. این حالت غیر قطعی برای این دو حرف به جای کلمه بر روی قاعده‌ها اعمال می‌شود تا اگر در قاعده‌ای این دو حرف به جای هم بکار رفته باشد مشکلی در تولید کلمه ایجاد نشود.

علاوه بر تمام نشانه‌های مختلف جمع در قاعده سازنده، قاعده‌هایی متناظر با جایگزین کلمه **جمع** بجای نشانه‌های مختلف جمع پذیرفته می‌شود.

علامت جمع مكسر (**جم**) از اسم‌های عربی‌ای که این علامت را در واژگان دارند حذف می‌شود. اعراب بعضی ریشه‌ها که برای تمایز استفاده شده است از قواعد حذف می‌شود.

۶-۸- ابزار واسط کاربر

این ابزار به زبان پایتون طراحی شده است و دارای یک واسط گرافیکی برای استفاده توسط کاربر است. همانطور که در شکل ۳-۶ مشاهده می‌شود امکان انتخاب بین مبدل‌ها وجود دارد و پس از ورود کلمه در قسمت ورودی، تصrif(های) مورد نظر در صورت موجود بودن برای کلمه تولید می‌شود و در قسمت خروجی نمایش داده می‌شود. اگر مبدل تولید انتخاب شود در ورودی می‌توان تصrif کلمه را وارد کرد و در خروجی کلمه(های) تولید شده برای آن قاعده را مشاهده کرد.



شکل ۳-۶ واسط گرافیکی کاربر برای استفاده از مبدل‌ها

۶-۹- جمع‌بندی

مبدل استاندارد را به عنوان مبدل اصلی این بخش باید در نظر گرفت. تمام قواعد فارسی رسمی و غیر رسمی در این مبدل گنجانده شده است. سایر مبدل‌های ثانویه نیز بر اساس این مبدل ساخته شده و کار می‌کند.

مبدل‌های ثانویه بیشتر بر روی کلمه‌های غیر رسمی تاثیر می‌گذراند (فصل ارزیابی) و به تحلیل تصrifی این کلمات کمک می‌کنند. این مبدل‌ها برای شناسایی تغییرات آوایی، خطای املایی بر روی کلمات رسمی (خطای کلمه‌های به هم چسبیده و خطاهای نگرشی^۳) نیز می‌توانند مفید باشند.

بسته به نوع کاربرد و دادگان می‌توان از هر کدام از این مبدل‌ها که مناسب‌تر است استفاده کرد. حتی می‌توان از همه این مبدل‌ها به شکل موازی بهره گرفت و تمام حالات ممکن را برای هر کلمه تولید کرد. بنابراین استفاده از هر یک با توجه به قابلیت‌ها و محدودیت آنها امکان‌پذیر است.

^۳ Cognitive Errors

فصل هفتم

ارزیابی

۱-۷ - مقدمه

برای ارزیابی، داده‌هایی از گونه‌های مختلف فارسی غیر رسمی انتخاب شده است؛ جملات کاملی که مجموعاً از هزار کلمه ساخته شده است. این جملات با رعایت اصول نمونه‌گیری و پیکره‌سازی از قسمت آزمون پیکره جمآوری شده برای این پژوهش انتخاب شده است (توضیحات کامل در بخش پیکره فارسی معاصر ۳-۱-۳). روش ارزیابی به این صورت است که هر کلمه توسط تحلیل گر تصrifی دریافت شده و قاعده‌های سازنده آن در خروجی تولید می‌شود. این کلمات مستقل از متن ارزیابی می‌شوند. معیارهای ارزیابی نیز شامل **فراخوانی^۱**, **درستی^۲** و **معیار اف (۱)^۳** هستند.

۲-۷ - معیارهای ارزیابی

با توجه به مستقل از متن بودن تحلیل گر، ارزیابی نیز مستقل از متن انجام می‌شود. در حالت مستقل از متن، مثبت صحیح تمام قاعده‌هایی (تاپ‌هایی) است که به ازای آنها کلمه می‌تواند در کاربردهای مختلف بکار رود.

اگر قاعده‌هایی (تاپ‌هایی) برای تمام کاربردهای ممکن کلمه غلط باشد مثبت غلط است. در صورتی که قاعده‌ای (تاپ) برای کلمه‌ای می‌بایست تولید می‌شده که تولید نشده است، منفی غلط در نظر گرفته می‌شود. هشدارهای نادرست مبنای صرفی- نحوی دارد و نه مبنای مفهومی. این بدان معنی است که به طور مثال قاعده سازنده کلمه **کتابند** که اسم+وربظی^۶ است، خطای نیست هرچند از نظر مفهومی کلمه بی‌ربطی است اما به این دلیل که از نظر صرفی- نحوی مشکلی در ساخت آن نیست، این کلمه و قاعده سازنده آن هشدار نادرست (مثبت غلط) حساب نمی‌شود. در مورد کلمات چه فعل‌ها و چه غیر فعل‌ها تا جای ممکن قاعده سازنده هر بخش به اندازه‌ای که هشدار نادرست تولید نکند، محدود شده است و از طرف دیگه هر کدام تا جایی که لازم بوده تا قاعده‌ای از دست نرود (منفی غلط) جامع است. این شاخص‌ها برای محاسبه معیارهای اصلی ارزیابی نظری **فراخوانی**, **درستی** و **معیار اف (۱)** بکار می‌رود.

$$\text{فراخوانی}^{\text{۴}} = \frac{\text{مثبت صحیح (موفقیت)}}{\text{مثبت صحیح} + \text{منفی غلط (از دست رفته)}}$$

^۱ Recall

^۲ Precision

^۳ F1 Measure

^۴ Recall

$$\text{درستی}^1 = \frac{\text{ثبت صحیح (موفقیت)}}{\text{ثبت صحیح} + \text{ثبت غلط (هشدار نادرست)}}$$

$$\text{معیار اف}(1)^2 = \frac{\text{فراخوانی} \times \text{درستی}}{\text{فراخوانی} + \text{درستی}}$$

۳-۷- ارزیابی قواعد بدست آمده از کلمات

در ارزیابی مستقل از متن، به نقش کلمات در جمله‌ها توجه نمی‌شود و کلمات یکتای این جملات استخراج می‌شود، از طرف دیگر اما به تمامی نقش‌هایی که یک کلمه می‌تواند ایفا کند و همه قاعده‌هایی (تاپ‌هایی) که یک کلمه قادر به تولید آنهاست توجه می‌شود.

گرچه کلمات یکتای آزمون برای این ارزیابی استخراج شده‌اند و تعداد این کلمات به نسبت تعداد کلمات موجود در جمله‌های آزمون کمتر است، اما باید به این نکته توجه داشت که در بررسی مستقل از متن به جای یک قاعده سازنده که مرتبت با جایگاه استفاده شده از کلمه در جمله است، باید به همه قاعده‌های سازنده با توجه به کاربردهای مختلف آن توجه کرد. از آنجایی که بخشی از کلمات به کار رفته در زبان فارسی معاصر، کلمات رسمی است، علاوه بر نوع خطاهای نوع کلمات از نظر رسمی بودن و غیر رسمی بودن نیز بررسی شده است.

جملاتی از هر زیر سیاق پیکره جمع‌آوری شده برای این پژوهش انتخاب شده است (با همان شیوه جمع‌آوری پیکره). مجموع تعداد همه کلمات ۴,۰۴۰ است که اگر کلمات یکتای آن را در نظر بگیریم ۱,۷۸۶ می‌شود. هر کدام از این کلمات یکتا می‌تواند یک یا بیشتر قاعده سازنده تولید کند و یا هیچ قاعده‌ای در تحلیل‌گر تولید نکند (کلمه خارج از واژگان باشد و یا قاعده سازنده آن در تحلیل‌گر وجود نداشته باشد). هر کدام از قاعده‌های سازنده کلمه یک تایپ از آن کلمه است که در بافت خاصی که استفاده می‌شود معنی پیدا می‌کند. باید توجه داشت که کلمات آزمون به صورت دستی تقطیع شده‌اند. در صورتی که از هر ابزار خودکاری برای تقطیع استفاده شود، به دلیل تقطیع متفاوت منجر به تغییر اعداد ارزیابی‌های این بخش می‌شود (احتمالاً معیارهای پایین‌تر).

بررسی گونه‌ای مستقل از متن خطاهای کلمات (تاپ‌های) آزمون در جدول ۱-۷ آمده است. همانطور که مشاهده می‌شود، در مبدل استاندارد از مجموع ۱,۷۸۶ کلمه یکتا، تحلیل‌گر، ۳,۷۰۴ تایپ را با موفقیت (TP) تولید کرده 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 دیگر برروی از دست رفته‌ها (FN) آزمایش شده و در صورتی که تغییری اتفاق افتد

^۱ Precision

^۲ F1 Measure

در جدول مشخص شده است. در این جدول رنگ سبز نشانه بهتر شدن شاخص و رنگ قرمز نشان دهنده بدتر شدن شاخص نسبت به مبدل قبل تر از خود است.

جدول ۱-۷ بررسی گونه‌ای خطاهای

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

کمتر از ۶۰ درصد تایپ‌ها، رسمی است. همانطور که در جدول ۱-۷ مشاهده می‌شود، بیشتر خطاهای رخ داده در بین همه تایپ‌ها (۷۱,۵ درصد)، به دلیل **کمبود واژگان** است. رتبه دوم بیشترین از دست رفته‌ها، خطاهای املائی است (۲۵ مورد، ۱۴,۵ درصد همه خطاهای)، که هیچ یک از مبدل‌ها قادر به رفع خطا و شناسایی آنها نیستند. خطاهای نقص در قاعده نیز با ۱۰ خطأ در جایگاه بعدی است. این خطاهای بیشتر مربوط به این مورد است: اجزای کلمات مرکب که جدا از سایر اجزای کلمه مرکب به تنهاًی به تحلیل‌گر تصریفی داده شده است. خطاهای بعدی خطاهای آوایی است که مبدل آوایی موفق شده از این ۵ خطأ ۳ مورد را شناسائی کند، همینطور کلمات خارج از واژگان غیر رسمی نیز ۱۲ تایپ تولید کرده است که مبدل آوایی توانسته با موفقیت (TP) شناسائی کند. بیشتر کلمات به هم چسبیده نیز توسط مبدل تقطیع شناسائی و رفع شده است. البته این مبدل در حین شناسائی **هشدارهای نادرست** نسبتاً زیادی نیز تولید می‌کند که درستی را کاهش می‌دهد. آخرین نوع خطاهای هم استفاده از **حروف هم‌صدا** و استفاده از ۵ به جای **کسره اضافه** است

که همگی آنها توسط مبدل همودا شناسایی و رفع شده است. از دست رفته‌های ادبی نیز ساختارهای ادبی‌ای است که مبدل قادر به شناسائی و تحلیل تصویری آنها نیست و مجموعاً سه خطاب تولید کرده است که به مجموع از دست رفته‌های رسمی (۱۱۲) افزوده شده است.

معیارهای ارزیابی اصلی **فراخوانی، درستی و معیار اف (۱)** تایپ‌های رسمی در جدول ۲-۷ آمده است. مبدل‌های ثانویه تاثیری در نتایج به دست آمده از مبدل استاندارد برای تایپ‌های رسمی ندارد. پایین آمدن معیار **درستی** نیز به دلیل کلی حساب کردن (بدون در نظر گرفتن رسمی یا غیر رسمی بودن تایپ) هشدارهای غلط (FP) است.

جدول ۲-۷ ارزیابی تایپ‌های رسمی

معیار	استاندارد	همودا	آوایی	بیانی	تقطیع
فراخوانی	%۹۵,۱	%۹۵,۱	%۹۵,۱	%۹۵,۱	%۹۵,۱
درستی	%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 غیر رسمی در مبدل همودا، آوایی و تقطیع بهبود می‌یابد اما درستی به دلیل افزایش هشدارهای خطاب کاهش پیدا می‌کند. این تغییرات در جدول ۳-۷ مشاهده می‌شود.

جدول ۳-۷ ارزیابی تایپ‌های غیر رسمی

معیار	استاندارد	همودا	آوایی	بیانی	تقطیع
فراخوانی	%۹۶,۲۳	%۹۶,۴۲	%۹۷,۳	%۹۷,۴۸	%۹۷,۴۸
درستی	%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، درستی و معیار اف (۱) برای تمامی کلمات مستقل از متن نیز در جدول ۴-۷ آمده است.

جدول ۴-۷ ارزیابی کل تایپ‌ها

معیار	استاندارد	همودا	آوایی	بیانی	تقطیع
فراخوانی	%۹۵,۵۶	%۹۵,۶۴	%۹۶	%۹۶,۰۸	%۹۹,۱
درستی	%۹۹,۶۵	%۹۹,۶۵	%۹۹,۴۹	%۹۹,۴۹	%۹۷,۵۷
معیار اف (۱)	%۹۷,۵۶	%۹۷,۶	%۹۷,۷۱	%۹۷,۷۱	%۹۷,۵۷

۴-۷- جمع‌بندی

با توجه به اینکه بخش زیادی از کلمات به کار رفته در جملات فارسی معاصر از کلمات رسمی تشکیل یافته است، تحلیل تصیری این کلمات نیز بخشی از پروسه پردازش زبان فارسی معاصر است. نتایج بدست آمده نشان می‌دهد که کلمات رسمی موجود تحلیل ساده‌تری دارد و در صورت وجود واژگان مناسب و کافی تحلیل تصیری آنها با موفقیت انجام می‌شود. از طرف دیگر کلماتی که غیر رسمی است، به دلیل ساختهای متنوع‌تر و تمایز در نگارش آنها چالش بیشتری ایجاد می‌کند. تا حد زیادی می‌توان با استفاده از مدل‌هایی که طراحی شده است این تنوع را پوشش داد و به دقت قابل قبولی نیز رسید. اما بخشی از کار در سطوح دیگر پردازش می‌باشد انجام شود. مثلاً تحلیل مبنی بر بافت و یا خطایابی املایی مبنی بر بافت و توجه بیشتر به قواعد آوازی در رفع و شناسایی خطاهای املایی، شناسایی و پردازش این متون را امکان‌پذیرتر می‌سازد.

فصل هشتم

نتیجه‌گیری و کارهای آتی

۱-۸ - مقدمه

تحلیل‌گر تصویفی پایین‌ترین سطح پردازش را (بعد از جداساز) بر روی متن فارسی انجام می‌دهد و اطلاعات تصویفی را به قطعه‌های متن خام اضافه می‌کند.

به جهت محدودیت زمانی و مشخص بودن مرز پژوهش همچنان قدم‌هایی می‌توان در جهت بهبود این تحلیل‌گر برداشت. از آنجاییکه این تحلیل‌گر مستقل از متن عمل می‌کند برای ابهام زدایی از قواعد نولیدی برای هر کلمه و انتخاب یک قاعده مشخص نیاز به انجام پردازشی دیگر است و سرنخ‌های بسیاری برای این امر در قواعد وجود دارد. از طرف دیگر از اطلاعات بدست آمده از تحلیل‌گر تصویفی می‌توان در سطوح بالاتر تحلیل نحوی و شناسایی گروه‌های نحوی استفاده کرد. از لحاظ نظری و زبانی نیز در طول این پژوهش پدیدارهایی در فارسی غیر رسمی مشاهده شده است که نیاز به بررسی و تحلیل زبان‌شناسانه دارد. ابزارهای مکمل و ساده‌ای بر اساس این تحلیل‌گر می‌توان ساخت که در تحلیل زبانی بکار آید. این موارد در بخش‌های زیر بیشتر بررسی شده است.

۲-۸ - بهبود تحلیل‌گر تصویفی

- با توجه به ارزیابی انجام شده، غیر از چند مورد نقص قاعده در ساختمان قواعد غیر رسمی، عمدۀ خطاهای ناشی از کمبود واژگان در تحلیل‌گر است. رده دوم خطاهای پر بسامد مربوط به تغییرات آوایی شدید و یا خطای حروف چینی است که روش‌های خطایابی و رفع خطای املایی مبتنی بر بافت (جورافسکی و مارتین، ۲۰۰۸؛ شیخ‌الاسلام و همکاران، ۲۰۱۲) با دقت بالایی می‌تواند در رفع آنها و انجام تحلیل تصویفی درست بکار آید.
- تغییرات آوایی نه فقط در سطح تصویف که در سطح اشتراق نیز اتفاق می‌افتد، گاهی هر سیلاپ کلمه که می‌تواند نقش تکواز اشتراقی یک کلمه را نیز نشان دهد، در صورت داشتن واکه پایانی با اتصال به تکواز دیگر (سیلاپ دیگر) این واکه‌پایانی تغییر می‌کند یا حذف می‌شود. با توجه به وجود تلفظ کلمات و ساخت سیلاپی هر کلمه در مدخل‌های واژگان زایا (اسلامی و همکاران، ۱۳۸۳)، اعمال این تغییرات آوایی در سطح سیلاپ تکوازها (که برخی نشان‌دهنده تکوازها هستند) امکان‌پذیر است. مثال، ایستگاه (ایست.گاه) ← ایسگاه ← ایسگا.
- در حال حاضر تنها اعدادی که به شکل حروف نگارش شوند در تحلیل‌گر شناسایی می‌شوند. برای شناسایی اعداد می‌بایست قواعدی ساخته شود که قادر به شناسایی اعداد چند رقمی، ممیزدار، اعشاری، درصد و غیره باشد. همینطور لازم است که قواعدی برای تبدیل عدد به معادل کلمه‌ای (عبارت) آنها و برعکس ساخته شود. تبدیل عدد به کلمه به خصوص در کاربردهایی نظیر متن به گفتار استفاده می‌شود.
- شناسایی عبارت‌های مربوط به ساعت و تاریخ نیز در تحلیل‌گر وجود ندارد. شناسایی این کلمات و عبارت‌ها نیز به عنوان بخشی از تحلیل تصویفی و سپس در سطوح دیگر تحلیل نیاز به پیاده‌سازی دارد.

- به صورت نظری کلمات خارج از واژگان را نیز با توجه به تصریفشان (در حالت مبتنی بر بافت روش‌های آماری گوناگونی برای شناسایی آنها می‌توان استفاده کرد) می‌توان تحلیل تصrifی کرد و آنها را با توجه به این ساختها شناسایی و تحلیل کرد. امکان ساخت قواعدی در محیط مبدل حالت متناهی برای حدس این کلمات با توجه به ساختشان وجود دارد.
- با افرودن امکان تحلیل تصrifی وندها و واژه‌بست‌هایی که از هسته کلمه جدا می‌افتد (هم در زبان رسمی و هم غیر رسمی بسیار اتفاق می‌افتد)، تحلیل گر تصrifی امکان تحلیل در پایین‌ترین سطح پردازش را می‌یابد و از ابزار جداساز^۱ بی‌نیاز می‌شود. مسیر وندها و واژه‌بست‌ها در قواعد وجود دارد و کافی است بجای الحاق به پایان کلمه، مسیرهای خالی ساخت که صاحب تصrif مربوط به آن وند و واژه‌بست باشد و در نهایت آن‌ها را تولید کند. به این صورت هر وند و واژه‌بست با تحلیل تصrifی مشخص می‌شود که به چه کلمه‌ای و در کجا کلمه (پیشوند در ابتدای کلمه و پسوند و واژه‌بست در انتهای کلمه) متصل می‌شود. تحلیل پس‌پردازشی شبیه به نوع چارت‌پارزینگ (آن، ۱۹۹۵) می‌تواند هسته‌های آنها را شناسایی کند.
- در قواعد فعلی تحلیل گر، قطعه‌های معرفه و نکره در توزیع تکمیلی هم هستند، این در حالی است که تولید کلمه‌ای که بعد از تک‌واژه معرفه، تک‌واژه نکرگی بپذیرد نیز امکان‌پذیر است. مثل، مغازه‌دارهای (که چاغ بود).
- به برخی پیشوندها در فعل‌های پیشوندی غیر رسمی می‌توان واژه‌بست شخصی افزود. مثل، برش دار. این ساختار محدود به فعل‌های پیشوندی متعدد است و شامل همه فعل‌های متعددی نیز نمی‌شود.
- حرف اضافه به در اتصال به برخی کلمات با از دست دادن و اکه ۵ به آن کلمه بدون فاصله متصل می‌شود. این ساختار برای همه قسم کلمات اتفاق نمی‌افتد و لازم است محدوده تاثیر آن شناسایی شده و به قواعد افزوده شود.

۳-۸- شناسایی دقیق تنها قاعده سازنده کلمه

در مرحله شناسایی قاعده دقیق از بین قاعده‌های تولیدی برای کلمه، جدای از روش‌های آماری و مبتنی بر بافت متنوع که همگی در این مورد قابل اجرا هستند، از اطلاعات افزوده‌ای که نسبت به تصrif رسمی در تصrif غیر رسمی وجود دارد می‌توان کمک گرفت. کلمات غیر رسمی نسبت به کلمات رسمی وند و واژه‌بست‌های بیشتری به همراه دارند که حاوی اطلاعات صرفی-نحوی بیشتری است. به طور مثال موارد زیر می‌تواند در شناخت قاعده کلمه و یا حداقل محدودتر کردن گزینه‌های ممکن، انتخاب قاعده صحیح را ساده‌تر سازد.

- فعل امری جمع (دوم شخص و اول شخص) منطبق بر معادل التزامیش است، به همین دلیل در قواعد تحلیل گر، فعل امری جمع وجود ندارد. یکی از وجود تمایز این دو فعل (بدون توجه به بافت) امکان استفاده فعل التزامی از

^۱ Tokenization

واژه‌بست هم است. فعل امری نمی‌تواند چنین واژه‌بستی بپذیرد. مثال، التزامی: اگه بش بگیدم فایده نداره! امری:

بش بگید! (دامنه تاثیر: فارسی غیر رسمی)

- صفت‌ها و اسم‌ها می‌توانند در نقش قید (مشترک) ظاهر شوند و به این دلیل که تنها با توجه به بافت می‌توان چنین نقشی را تشخیص داد امکان برچسب زنی آنها مستقل از متن فراهم نیست. از طرف دیگر اگر اسم و صفت از ساخت تصrifی جمع به بعد (در ساختمان اسمی اولین وند تصrifی جمع و در ساختمان صفت‌ها دومین وند تصrifی علامت جمع است) در ساختمان خود استفاده کنند دیگر قطعاً نمی‌توانند در جمله نقش قید داشته باشند. مثال، صفت در نقش قید: او خوب حرف می‌زند. صفت قابل تمیز (مستقل از متن): این کار خوبه! (دامنه تاثیر: فارسی رسمی و غیر رسمی)
- تک واژه نکره و موصولی بدون در نظر گرفتن بافت در فارسی رسمی قابل تمیز نیست، اما در فارسی غیر رسمی هر کدام ساختمان تصrifی خاص خود را دارد که اگر از آن استفاده کند در بخش زیادی از این ساختمان از یکدیگر متمایز هستند. در اینجا به ذکر یک مثال اکتفا می‌کنیم (بخش اسم، ساختمان مانع اسم، فصل پنجم). اسم دارای تک واژه موصولی هرگز تک واژه تاکید نمی‌پذیرد اما اسم دارای تک واژه نکره می‌تواند تک واژه تاکید نیز بپذیرد. مثال، موصولی: کتاباییرم که خرید نخوند. نکره: عجب جونورایین! (دامنه تاثیر: فارسی غیر رسمی)
- ماضی نقلی سوم شخص مفرد که بسامد بالایی دارد معمولاً بدون فعل کمکی است در زبان غیر رسمی استفاده می‌شود. این فعل بدون هیچ تصrifی در ظاهر هیچ تفاوتی با صفت مفعولی ندارد. اما اگر این دو کلمه از تصrif‌های ساخت‌هایشان استفاده کنند ممکن است بتوان در سطح کلمه (مستقل از متن) آنها را از هم تمیز داد. ماضی نقلی برخلاف صفت مفعولی می‌تواند ت واسطه بپذیرد. صفت مفعولی برخلاف ماضی نقلی می‌تواند وند جمع بپذیرد. مثال، ماضی نقلی: خورده‌تشون! صفت مفعولی: خورده‌هاشون.
- صفت و ضمیر اشاره (همینطور پرسشی و مبهم) را در فارسی رسمی غالباً از روی بافت می‌توان از هم تشخیص داد مگر اینکه تک واژه تصrifی در ساخت آن استفاده شده باشد که در این صورت ضمیر است (این مورد بسامد پایینی دارد). در فارسی غیر رسمی هم اگر از تک واژه تصrifی استفاده نشود راهی غیر از توجه به بافت برای تمیز صفت از ضمیر وجود ندارد. اما ضمایر از ساخت تصrifی، بسیار استفاده می‌کنند. (دامنه تاثیر: فارسی رسمی و غیر رسمی)

۴-۸- استفاده از اطلاعات تصrifی برای شناسایی سازه‌های نحوی

گروه‌های فعلی را با توجه به تصrif‌های بدست آمده برای هر قطعه می‌توان شناسایی کرد. از این جمله ساخت‌های، می‌توان به فعل‌های ساده که به عنوان واحدهای چند قطعه‌ای می‌شناسیم، اشاره کرد. گرچه ساختمان این افعال در قواعد تصrifی گنجانده شده است اما از آنجا که جداساز قادر به شناسایی این گروه‌ها نیست عملاً این ساخت‌ها بلا استفاده است و از مبدل

نهایی کنار گذاشته شده است. راه شناسایی آنها در پس‌پردازش و با داشتن اطلاعات تصریفی هر عضو آن فراهم می‌شود. برای مثال این ساخت‌ها را به این صورت می‌توان شناسایی کرد:

- ماضی التزامی = (منفی) + صفت مفعولی (بدون تصریف) + مضارع ساده باش (غیر منفی)
- مستقبل = (منفی) + مضارع ساده خواه + ماضی ساده (غیر منفی)
- ماضی بعید = (منفی) + صفت مفعولی (بدون تصریف) + ماضی ساده بود (غیر منفی)
- ماضی بعد = (منفی) + صفت مفعولی (بدون تصریف) + ماضی نقلی بود

انواع فعل‌های چند قطعه‌ای مجھول را نیز می‌توان به صورت جدول ۱-۸ شناسایی کرد.

جدول ۱-۸ نحوه شناسایی فعل مجھول در پس‌پردازش

بن فعل مضارع شو	ریشه فعل اصلی	فعل چند قطعه‌ای مجھول
مضارع ساده شو		مضارع ساده
مضارع التزامی شو		مضارع التزامی
مضارع اخباری شو		مضارع اخباری
ماضی ساده شد		ماضی ساده
ماضی نقلی شد		ماضی نقلی
ماضی استمراری شد	= صفت مفعولی فعل اصلی +	ماضی استمراری
ماضی بعید شد		ماضی بعید
ماضی نقلی مستمر شد		ماضی نقلی مستمر
ماضی التزامی شد		ماضی التزامی
ماضی بعید شد		ماضی بعید
مستقبل شد		مستقبل

گروه‌های نحوی دیگر نیز به همین شکل می‌تواند شناسایی شود.

۵-۸- پژوهش‌های زبانی

از مقایسه ساختمان کلمه رسمی و غیر رسمی در هر دو بخش فعل و غیر فعل می‌توان مشاهده کرد که ساختمان غیر رسمی دارای اجزای ساختمان بیشتری است. این موضوع در حوزه ارزیابی آماری نیز تایید می‌شود (جدول ۲-۸).

جدول ۲-۸ بررسی تعداد تکوازهای ساختمان تصریف

غیر رسمی	رسمی		
۱۵۳۱	۲۱۷۴	همه کلمات	
۱۴۸۰	۱۰۷۶	کلمات دارای تصریف	
۵۱	۱۱۰۰	کلمات بدون تصریف	
۱,۵	۰,۵۹	به کل کلمات	نسبت تعداد
۱,۵۵	۱,۱۹	به کلماتی که تصریفی‌اند	تکوازهای ساختمان تصریفی

میانگین نسبت تعداد تکوازهای تصریفی به کلمه غیر رسمی نزدیک به ۳ برابر رسمی است (۱,۵ در مقابل ۰,۵۹). این نشان می‌دهد که کلمات غیر رسمی از تکوازهای بیشتری استفاده می‌کنند. همینطور کلمات دارای تکواز تصریفی غیر رسمی بیشتر از کلمات دارای تکواز تصریفی رسمی دارای تکواز تصریفی است (۱,۵۵ در مقابل ۱,۱۹).

۶-۸- ابزارهایی که می‌توان مبتنی بر تحلیل‌گر تصریفی ساخت

- ساخت ابزار تقطیع تکوازها که برای برخی تحلیل‌های آماری ساخت‌وازی استفاده می‌شود توسط این ابزار امکان‌پذیر است. تنها کافیست در قواعد مبدل به جای اتصال تکوازها، آنها را از هم جدا و در خروجی چاپ کرد.
- ریشه‌یاب یکی از ابزارهای کاربردی در مهندسی متن به خصوص بازیابی اطلاعات است. افردون یک دستورالعمل ساده برای استخراج ریشه کلمات از قواعد خروجی تحلیل‌گر کار بسیار ساده‌ای است. قطعه بعد از مساوی و قبل از علامت بعلاوه، ریشه کلمه است. از بین قواعد خروجی تحلیل‌گر برای یکی کلمه، ریشه‌های یکتا را به عنوان خروجی انتخاب می‌کنیم. مثال: ریشه قاعده‌های تولید شده <استاندارد:صفت=خوشحال+وربطی^۳> و <استاندارد:اسمعام=قلب+وشخصی۲+رسمی> کلمه‌های خوشحال و قلب است که بین علامتهای مساوی و بعلاوه قرار دارد.

- با ساخت یک جدول ساده نگاشت ریشه کلمات رسمی به غیر رسمی برای فعل و غیر فعل می‌توان یک دستورالعمل ساده برای تبدیل کلمه غیر رسمی به رسمی و یا بر عکس ساخت. برای این کار دستورالعمل ریشه معادل را در قاعده

تولیدی تحلیل گر جایگزین می‌کند و به پایان قاعده، قطعه +رسمی را اضافه می‌کند و این قاعده را به مبدل تولیدی می‌دهد تا کلمه رسمی معادل ساخته شود. البته مواردی وجود دارد که باید در نظر داشت. اول اینکه همه ساختهای غیر رسمی معادل رسمی ندارند، برخی با یک عبارت می‌توانند جایگزین شوند مثل کتابشونو (ببر) ← کتابشان+را. برخی دیگر معادل دقیقی ندارند مثل می‌زنم! دوم اینکه قاعده‌های رسمی و غیر رسمی همیشه نسبت یک به یک ندارند و ممکن است برخی ساختهای غیر رسمی علی رغم وجود داشتن تک‌تک سازه‌های آن در رسمی، کل قاعده در تصریف رسمی پذیرفتی نباشد مثل چرخوندشون! ← (آنها را) چرخاند.

مثال: کلمه **شیطونه** قاعده <استاندارد: اسماعام=شیطون+وربٹی^۳> را تولید می‌کند. حال ریشه **شیطون** را که بین علامت مساوی و بعلاوه قرار دارد، بر می‌داریم و با معادل رسمی‌اش که **شیطان** است جایگزین می‌کنیم. قطعه +رسمی را نیز به انتهای قاعده اضافه می‌کنیم و به ورودی مبدل تولید می‌دهیم. قاعده <استاندارد: اسماعام=شیطان+وربٹی^{۳+رسمی}> کلمه **شیطان** است و **شیطان** است را تولید می‌کند.

مراجع

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

اسلامی، محرم و علیزاده لجمیری، صدیقه. ۱۳۸۸. «ساختار تصrifی کلمه در زبان فارسی» در دانشگاه تبریز. نشریه زبان و ادب فارسی. شماره ۲۱۱.

اسلامی، محرم، شریفی آتشگاه، مسعود و علیزاده لمجیری، صدیقه و زندی، طاهره. ۱۳۸۳. «واژگان زایای زبان فارسی» در اولین کارگاه پژوهشی زبان فارسی و رایانه.

احمدی گیوی، حسن و انوری، حسن. ۱۳۹۱. «دستور زبان فارسی». ویرایش چهارم. انتشارات فاطمی.

جعفری تازه‌جانی، سمیه و بحرانی، محمد. ۱۳۹۲. «بررسی روند تغییرات در تبدیل فرم رسمی افعال به فرم محاوره‌ای و ارائه یک تحلیل‌گر صرفی برای افعال محاوره‌ای» در اولین هم‌اندیشی زبان فارسی و اینترنت، تهران.

دبیر سیاقی، محمد. ۱۳۹۰. «لغت نامه فارسی بزرگ». نسخه لوح فشرده. موسسه لغت نامه دهخدا.

شقاقی، ویدا. ۱۳۹۴. «واژه‌بست چیست؟ آیا در زبان فارسی چنین مفهومی کاربرد دارد؟» در سومین کنفرانس زبان‌شناسی. مجموعه مقالات دانشگاه علامه طباطبائی. شماره ۸۳.

صادقی، علی‌اشraf و زندی مقدم، زهرا. ۱۳۸۵. «فرهنگ املایی زبان فارسی» فرهنگستان زبان و ادب فارسی.

روحانیان، مجتبی، غفاری، کامران و وزیرنژاد، بهرام. ۱۳۹۳. «طراحی تحلیل‌گر ساختواری برای زبان پارسی، با استفاده از ساختار تراکندر و روش دوستطحی» در سومین همایش ملی زبان‌شناسی رایانشی، دانشگاه صنعتی شرف.

کاشفی، امید، نصری، میترا، کنعانی، کامیار. ۱۳۸۹. «خطایابی املایی خودکار در زبان فارسی». دبیرخانه شورای عالی اطلاع رسانی معین، محمد و امیریان، منیژه. ۱۳۸۲. «فرهنگ فارسی». موسسه انتشارات امیر کبیر.

مواجی، وحید، اسلامی، محرم، بهرام. ۱۳۹۰. «پارس مورف: تحلیل‌گر صرفی زبان فارسی». نشریه پردازش علائم و داده‌ها. شماره ۱۵.

Allen, J. (1995). *Natural language understanding*. Pearson.

Amtrup, Jan Willers, Hamid Mansouri Rad, Karine Megerdoomian, and Rémi Zajac. *Persian-English Machine Translation: An Overview of the Shiraz Project*. Computing Research Laboratory, New Mexico State University, 2000.

Antworth, Evan L. "PC-KIMMO: A Two-Level Processor for Morphological Analysis," 1991.

Biber, Douglas. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8, no. 4 (1993): 243–57.

Biber, Douglas. "Using Register-Diversified Corpora for General Language Studies." *Computational Linguistics* 19, no. 2 (1993): 219–41.

- Bijankhan, Mahmood, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. “Lessons from Building a Persian Written Corpus: Peykare.” *Language Resources and Evaluation* 45, no. 2 (May 2011): 143–64. <https://doi.org/10.1007/s10579-010-9132-x>.
- 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.
- Ingason, Anton Karl, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. “A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI).” In *Advances in Natural Language Processing*, 205–16. Springer, 2008.
- Jurafsky, Daniel, and James H Martin. *Speech and Language Processing*, 2nd Edition. Prentice Hall, 2008.
- Karine Megerdoomian. “Analysis of Farsi Weblogs,” August 2008.
- Karine Megerdoomian. (2006). Extending a Persian Morphological Analyzer to Blogs.
- Karine Megerdoomian. “Persian Computational Morphology: A Unification-Based Approach.” Computing Research Laboratory, New Mexico State University, 2000.
- Kashefi, O, M Nasri, and K Kanani. “Towards Automatic Persian Spell Checking.” *Tehran, Iran: SCICT*, 2010.
- Kenneth R Beesley, and Lauri Karttunen. *Finite State Morphology (Xerox)*. CSLI Studies in Studies in Computational Linguistics. Stanford, Calif, 2003.
- Krovetz, Robert. “Viewing Morphology as an Inference Process.” In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 191–202. ACM, 1993.
- Kukich, Karen. “Spelling Correction for the Telecommunications Network for the Deaf.” *Communications of the ACM* 35, no. 5 (1992): 80–90.
- Lazard, Gilbert. “Grammaire Du Persan Contemporain.” *Bibliothèque Iranienne*, no. 61 (2006).
- Manning, Christopher D, and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Vol. 999. MIT Press, 1999.
- McEnery, Tony, and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2011.
- Porter, Martin F. “An Algorithm for Suffix Stripping.” *Program* 14, no. 3 (1980): 130–37.
- Roark, Brian, and Richard William Sproat. *Computational Approaches to Morphology and Syntax*. Oxford University Press Oxford, 2007.
- Sagot, Benoît, and Géraldine Walther. “A Morphological Lexicon for the Persian Language.” In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC’10)*, 2010.
- Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. In LREC. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/809_Paper.pdf
- Sharifloo, Amir Azim, and Mehrnoush Shamsfard. “A Bottom Up Approach to Persian Stemming.” In *IJCNLP*, 583–88, 2008.
- Sheykholeslam, M. H., Minaei-Bidgoli, B., & Juzi, H. (2012). A Framework for Spelling Correction in Persian Language Using Noisy Channel Model. In *LREC* (pp. 706–710).
- Yarowsky, David, and Richard Wicentowski. “Minimally Supervised Morphological Analysis by Multimodal Alignment.” In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 207–16. Association for Computational Linguistics, 2000.

برچسب‌های تحلیل‌گر

اسم خاص نام فامیلی	اسمفامیلی
صفت	صفت
قید	قید
مصدر	مصدر
صفت مفعولی	صمفعولی
صفت شمارشی	شماره
حرف اضافه	ضافه
حرف اضافه گروهی	ضافگ
ضمیر شخصی	شخصی
ضمیر مشترک	مشترک
صفت یا ضمیر مبهم	مبهم
صفت یا ضمیر اشاره	اشاره
صفت یا ضمیر پرسش	پرسش
شیوه جمله	جملک
شاخص	شاخص
حرف ربط	ربط
حرف ربط گروهی	ربطگ

برچسب	مفهوم
ف.و	فعل وجهی
اسناد	اسناد
ف.ح.س	فعل مضارع (حال) ساده
ف.م.س	فعل ماضی ساده
ف.م.ن	فعل ماضی نقلی
ف.م.ا	فعل ماضی استمراری
ف.م.ن.م	فعل ماضی نقلی مستمر
ف.ح.ا	فعل مضارع (حال) اخباری
التزامی	مضارع التزامی
امری	امری
ف.م.ب*	فعل ماضی بعید
ف.م.ال*	فعل ماضی التزامی
ف.م.ا.ب*	فعل ماضی ابد
ف.آ.*	فعل مستقبل (آینده)
اسمعام	اسم عام
اسمجا	اسم خاص مکان
اسمشخص	اسم خاص شخص

* گرچه این ساختارها در قواعد وجود دارد اما به دلیل چند قطعه‌ای (Multi Tokens Unit) بودن از تحلیل‌گر نهایی کنار گذاشته شده است. فایل آن به طور جداگانه وجود دارد و می‌توان از آن استفاده کرد (MTvbs.lexc و MTchain.lexc).

قطعات به کار رفته در قاعده‌ها

محدوده کاربرد	مفهوم	قطعه
غیر فعل	تکواز جمع غیر رسمی ا	جا
غیر فعل	تکواز جمع ها	جها
غیر فعل	تکواز جمع جانداران	جان
غیر فعل	تکواز جمع عربی ات	جات
غیر فعل	تکواز جمع عربی جات	جاجات
غیر فعل	تکواز جمع عربی بن	جين
غیر فعل	تکواز جمع عربی ون	جون
فعل و غیر فعل	تکواز هم	هم
فعل و غیر فعل	تکواز را (رو، و)	را
فعل و غیر فعل	تکواز و واژه و	عطف
غیر فعل	تکواز نکره یا موصولی	نم
غیر فعل	تکواز موصولی	موصولی
غیر فعل	تکواز نکره	نکره
غیر فعل	واژه‌بست شخصی	وشخصی
فعل و غیر فعل	واژه‌بست ربطی	وربطی
-۵-۲-۸ فعل	واژه‌بست ربطی ویژه	ربطویژه
غیر فعل	تکواز معرفه	معرفه
فعل و غیر فعل	تکواز تاکید ها و ا	تاکید
غیر فعل	تکواز کسره اضافه	اضافه
فعل	شناسه ماضی و مضارع	ش
فعل	واژه‌بست مفعولی فعل متعدد	ومفعولی
-۴-۳-۲ فعل	حرف واسط ت	ت
فعل و غیر فعل	کلمه رسمی، کلمه‌ای که در قاعده تولیدی آن این قطعه نباشد، غیر رسمی محسوب می‌گردد.	رسمی

واژه‌نامه فارسی به انگلیسی

Subject	فاعل	FSA	اتوماتای حالت متناهی
Modal	فعل وجهی	Finite State Automata	اتوماتای حالت متناهی
Recall	فرآخوانی	Indicative	اخباری
MUT	قطعه چند واحدی	Upgrade	ارتقاء
Multi Tokens Unit	قطعه چند واحدی	Subjunctive	التزامی
Auxiliary	کمکی	Imperative	امری
Double Compound	ماضی ابعد	Unsupervised	بی‌نظرارت
Past Progressive	ماضی استمراری	Transformation	تبديل
Past Subjunctive	ماضی التزامی	Analytic	تحليلی
Pluperfect	ماضی بعید	Synthetic	ترکیبی
Simple Past	ماضی ساده	Random	تصادفی
Imperfect	ماضی مستمر	Segmentation	تقطیع
Compound Imperfect	ماضی نقلی مستمر	Tokenize	جداسازی کردن
Present Perfect	ماضی نقلی	Tokenization	جداسازی
FST	مبدل حالت متناهی	Tokenizer	جداساز
Finite State Transducer	مبدل حالت متناهی	Lookup Table	جدول جستجو
Future	مستقبل	Annotation	حاشیه نویسی
Simple Present	مضارع ساده	Precision	درستی
Present Progressive	مضارع مستمر	Script	دستورالعمل
F Measure	معیار اف	Copula	ربطی
Object	مفهول	Stemmer	ریشه‌یاب
Heuristic	مکاشفه‌ای	Regular Expression	زبان قاعده‌مند
First Order Logic	منطقه مرتبه اول	Formal Language	زبان مقید
Markedness	نشان‌داری	Lexical Chain	زنگیره واژگانی
Marked	نشان‌دار	Feature Structure	ساختار ویژگی
Flag Diacritic	نشانه فرامتنی	Bias	سوء گیری
Supervised	نظارتی	Recognition	شناسایی
Representative	نماینده‌گی	Past Inflection	شناسه ماضی
Sampling	نمونه گیری	Present Inflection	شناسه مضارع
Sample	نمونه	Morphosyntactic	صرفی- نحوی
Character	نویسه	Formalism	صورت‌بندی
		Non-Deterministic	غیر قطعی

Semi-Supervised	نیمه-نظرارتی
MTU	واحد چند قطعه ای
Multi Units Token	واحد چند قطعه ای
Critic	واژه‌بست

واژه‌نامه انگلیسی به فارسی

Morphosyntactic	صرفی- نحوی	Analytic	تحلیلی
Multi Tokens Unit	قطعه چند واحدی	Annotation	حاشیه نویسی
Multi Units Token	واحد چند قطعه ای	Auxiliary	کمکی
Non-Deterministic	غیر قطعی	Bias	سوء گیری
Object	مفهول	Character	نویسه
Past Inflection	شناسه ماضی	Critic	واژه‌بست
Past Progressive	ماضی استمراری	Compound Imperfect	ماضی نقلی مستمر
Past Subjunctive	ماضی التزامی	Copula	ربطی
Pluperfect	ماضی بعید	Double Compound	ماضی ابعد
Precision	درستی	F Measure	معیار اف
Present Inflection	شناسه مضارع	FSA	اتوماتای حالت متناهی
Present Perfect	ماضی نقلی	FST	مبدل حالت متناهی
Present Progressive	مضارع مستمر	Feature Structure	ساختار ویژگی
Random	تصادفی	Finite State Automata	اتوماتای حالت متناهی
Recall	فراخوانی	Finite State Transducer	مبدل حالت متناهی
Recognition	شناسایی	First Order Logic	منطقه مرتبه اول
Regular Expression	زبان قاعده‌مند	Flag Diacritic	نشانه فرامتنی
Representative	نمایندگی	Formal Language	زبان مقید
Sample	نمونه	Formalism	صورت‌بندی
Sampling	نمونه گیری	Future	مستقبل
Script	دستورالعمل	Heuristic	مکاشفه‌ای
Segmentation	تقاطیع	Imperative	امری
Semi-Supervised	نیمه-نظرارتی	Imperfect	ماضی مستمر
Simple Past	ماضی ساده	Indicative	خبری
Simple Present	مضارع ساده	Lexical Chain	زنگیره واژگانی
Stemmer	ریشه‌یاب	Lookup Table	جدول جستجو
Subject	فاعل	MTU	واحد چند قطعه ای
Subjunctive	التزاماً	MUT	قطعه چند واحدی
Supervised	نظرارتی	Markedness	نشان‌داری
Synthetic	ترکیبی	Marked	نشان‌دار
Tokenization	جداسازی	Modal	فعل وچهی
Tokenizer	جداساز		

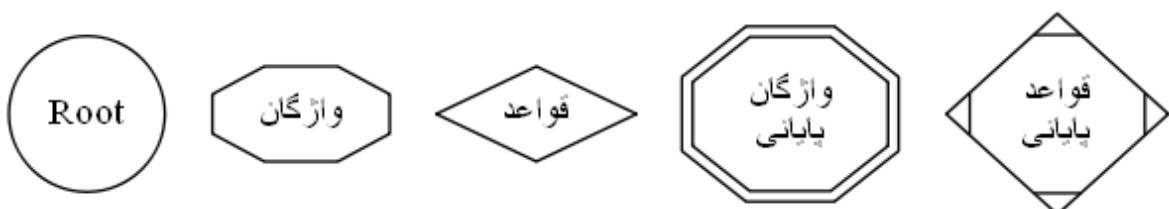
Tokenize	جداسازی کردن
Transformation	تبدیل
Unsupervised	بی‌نظرارت
Upgrade	ارتقاء

پیوست یک

پیاده‌سازی فعل

۱-۱- مقدمه

برای سادگی نمایش روابط لایه‌های درون زیرساخت که واژگان و قواعد را ترکیب کرده کلمات را می‌سازد، از نمودارهای مفهومی‌ای استفاده شده است که روابط این لایه‌ها را نمایش می‌دهد. نمونه‌ای از این نمودارها را می‌توان در شکل ۱-۵ و شکل ۱-۶ مشاهده نمود. نمودار از راست به چپ است و هر گره، رشته و برچسبی را به گره بعد منتقل می‌کند. این انتقال به گره بعدی تا جایی ادامه پیدا می‌کند که کلمه و قاعده تصریفی آن کامل شود. قاعده و رشته‌ای که هر گره منتقل می‌کند تا جایی که امکان داشته بر روی کمان‌ها نمایش داده شده است. هر گره هم بسته به محتوای خود شکل ویژه‌ای دارد که در شکل ۱-۱ و بیشگی هریک توضیح داده شده است.



شکل ۱-۱ گره‌های بکار رفته در نمودارهای مفهومی

ساختار کل هر فعل‌ها از نظر پیاده سازی شبیه به هم است. بنابراین به جای توضیح پیاده سازی تک‌تک ساخته‌ها، قالب کلی توضیح داده می‌شود، سپس موارد ویژه‌ای که در قسمت‌های مختلف وجود دارد، بیان می‌گردد.

۲-۱- فعل‌های تک قطعه‌ای و چند قطعه‌ای

فعل‌های ساده تک قطعه‌ای در فایل verbs.iverbs.lexc گنجانده شده‌اند، این در حالی است که فعل‌های چند قطعه‌ای در فایل MTvbs.lexc قرار دارند. علت این تقسیم‌بندی به این دلیل است که اگر بنا بود در تحلیل تصریفی تنها به واحدهای تکی اکتفا شود و مبدل سبک‌تر و سریع‌تری برای شناسایی کلمه‌های یک متن ساخته شود (که قطعه‌بندی آن بسیار ساده‌تر از شناسایی گروه‌های فعلی است)، به راحتی بخش وارد کردن فایل آن در فواید را حذف کنیم (شکل ۲-۱). در خط آخر کافی است که MTverbs را حذف کنیم تا تمام فعل‌های چند قطعه‌ای از مبدل حذف شود.

```

read lexc lexc/iverb.lexc
define iLexVerb;

read lexc lexc/MTchain.lexc
define MTchain;

regex iLexVerb | MTverbs;

```

شکل ۲-۱ نحوه بارگذاری و جداسازی فعلهای تک و چند قطعه‌ای

در مورد فعلهای پیشوندی هم همین جداسازی صورت گرفته است. فعلهای تک قطعه‌ای در فایل ichain.lexc و فعلهای چند قطعه‌ای در فایل MTchain.lexc قرار دارد. حذف و یا بارگذاری آن درست مانند فعلهای ساده است.

۳-۱- تفاوت فعلهای رسمی و غیر رسمی

از آنجایی که فعلهای غیر رسمی از بن فعلهای رسمی هم استفاده می‌کند و شناسه‌های رسمی هم در ساختهای غیر رسمی بکار می‌رود، هر دوی این ساختها در کنار هم تعریف شده‌اند تا از منابع یکدیگر بهر ببرند. از لحاظ نظری تفاوت فعل رسمی از غیر رسمی کاملاً روشن است و در فصل فعلهای توضیح داده شده است. در پیادهسازی نیز در قواعد فعلهای رسمی —مثل سایر کلمات رسمی— قطعه+رسمی در قاعده فعل گنجانده شده است. فعلهای غیر رسمی چنین قطعه‌ای در قاعده خود ندارند.

این تمایز در پیادهسازی به اینصورت انجام شده است که بن‌های رسمی دارای نشانه فرامتنی^۱ می‌شوند و به این صورت از بن‌های غیر رسمی جدا می‌شوند. اگر این بن‌ها در ساختهایی هم که ساختاری غیر رسمی دارند، استفاده شوند، این نشانه‌های فرامتنی خود را از دست می‌دهند و به این صورت تبدیل به المانی غیر رسمی می‌شوند. نشانه‌های فرامتنی تنها در مرحله ساخت وجود دارند و در خروجی مبدل، همراه کلمه و یا قاعده کلمه دیده نمی‌شوند. به طور مثال (شکل ۳-۱) برای فعل گذشته ساده، بن‌های رسمی و غیر رسمی، هر دو در گره SimpleP قرار دارد. بن‌های رسمی به گره past انتقال داده می‌شوند و در آنجا نشانه فرامتنی P.FORMAL.ON@ به همه آنها افزوده می‌شود و از آنجا به گره ipast منتقل می‌شوند. از طرف دیگر بن‌های غیر رسمی، مستقیم به گره ipast انتقال داده می‌شوند.

^۱ Flag Diacritic

```

LEXICON iSimpleP
افشوند ipast;
پاشوند ipast;
شوروند ipast;
شکوند ipast;
...
آفرید past;
آکند past;
افراشت past;
افروخت past;
...

LEXICON past
@P.FORMAL.ON@:@P.FORMAL.ON@ ipast;

LEXICON ipast
+م*:۱ش iObjectClitic;
+م*:۲ش iObjectClitic;
+۳:۰ش iObjectClitic;
+م*:۴ش iObjectClitic;
@C.FORMAL@+۵ش:@C.FORMAL@ ین*iObjectClitic;
+۶ش:iObjectClitic;
@C.FORMAL@+۷ش:@C.FORMAL@ ن*iObjectClitic;
+۸ش:iObjectClitic

```

شکل ۳-۱ تمایز فعل رسمی و غیر رسمی

هر دو جنس بن فعل (رسمی و غیر رسمی) در گره ipast شناسه‌های ساخت فعل ماضی را می‌گیرند. شناسه‌هایی که مربوط به ساختهای غیر رسمی‌اند، با استفاده از نشانه @C.FORMAL@ نشانه فرامتنی قبلی را که به بن‌های رسمی اضافه شده بود (@P.FORMAL.ON@)، پاک می‌کنند. ساختهایی هم که ساخت غیر رسمی ندارند، تغییری در نشانه‌های فرامتنی ایجاد نمی‌کنند.

تمام ساختهای این گره به گره iObjectClitic منتقل می‌شود (شکل ۳-۱). در این گره اگر واژه‌بست مفعولی به ساختی اضافه شود، حتماً با استفاده از نشانه @C.FORMAL@ نشانه گذاری فرامتنی رسمی آن ساخت حذف می‌شود، و فعل تبدیل به یک ساخت غیر رسمی می‌شود (فعل رسمی واژه‌بست مفعولی نمی‌پذیرد). ساختی که واژه‌بست نمی‌پذیرد از مسیر تهی ":" عبور کرده و بدون تغییر در نشانه فرامتنی احتمالی‌ای که به همراه دارد به گره بعدی (iStress) می‌رود.

```

LEXICON iObjectCitic
@C.FORMAL@+۱: و مفعولی * م iStress;
@C.FORMAL@+۲: و مفعولی * ت iStress;
@C.FORMAL@+۳: و مفعولی * ش iStress;
@C.FORMAL@+۴: و مفعولی * مون iStress;
@C.FORMAL@+۵: و مفعولی * مان iStress;
@C.FORMAL@+۶: و مفعولی * تون iStress;
@C.FORMAL@+۷: و مفعولی * تان iStress;
@C.FORMAL@+۸: و مفعولی * شون iStress;
@C.FORMAL@+۹: و مفعولی * شان iStress;
@C.FORMAL@+۱۰: و مفعولی * هم iStress;
:

LEXICON iStress
@D.FORMAL@:@D.FORMAL@ #;
@R.FORMAL@ +:=@R.FORMAL@ #;
+۱*: # تاکید;
+۲*: ^ ها;
+۳*: ^ عطف;
+۴*: * و;
+۵*: هم #;

```

شکل ۱-۴ تمایز فعل رسمی و غیر رسمی

در گره پایانی *iStrees* اگر یک از واژه‌بسته‌های محاوره (تاکید، عطف و هم) به ساخت اضافه شود، فعل به مرحله پایانی می‌رسد و تولید می‌شود. در این حالت فعل یک فعل غیر رسمی است (واژه‌بست محاوره فعل غیر رسمی می‌سازد). در قاعده سازنده فعل هم قطعه **+رسمی** درج نمی‌شود. اگر فعل از واژه‌بسته‌های این گره نپذیرد، دو مسیر برای ادامه برایش باقی می‌ماند.

مسیر اول که نشانه **@P.FORMAL.ON@** دارد، راه را برای ساختهایی که نشانه فرامتنی **@P.FORMAL.ON@** دارد، می‌بینند. این مسیر برای ساختهایی است که این نشانه فرامتنی را ندارند (یا از ابتدا نداشته‌اند یا در مسیر ساخت از دست داده‌اند). بنابراین در این مسیر فعل‌های غیر رسمی ساخته می‌شوند و در قاعده سازنده انها نیز قطعه **+رسمی** درج نمی‌شود.

مسیر دیگر دارای نشانه **@R.FORMAL@** است که برای ساختهایی است که حتی نشانه فرامتنی را همراه خود دارند. یعنی ساختهای برای عبور از این مسیر نیازمند این نشانه فرامتنی هستند، در غیر این صورت نمی‌توانند عبور کنند. ساختی که از بن رسمی تشکیل شده و در مسیرش تا رسیدن به این گره، عنصر غیر رسمی‌ای به آن اضافه نشده (بنابراین هنوز نشانه فرامتنی **@P.FORMAL.ON@** را دارد) می‌تواند از این مسیر عبور کند. در نتیجه قطعه **+رسمی** به قاعده سازنده ساختی که از این قسمت عبور می‌کند و به پایان می‌رسد، افزوده می‌شود.

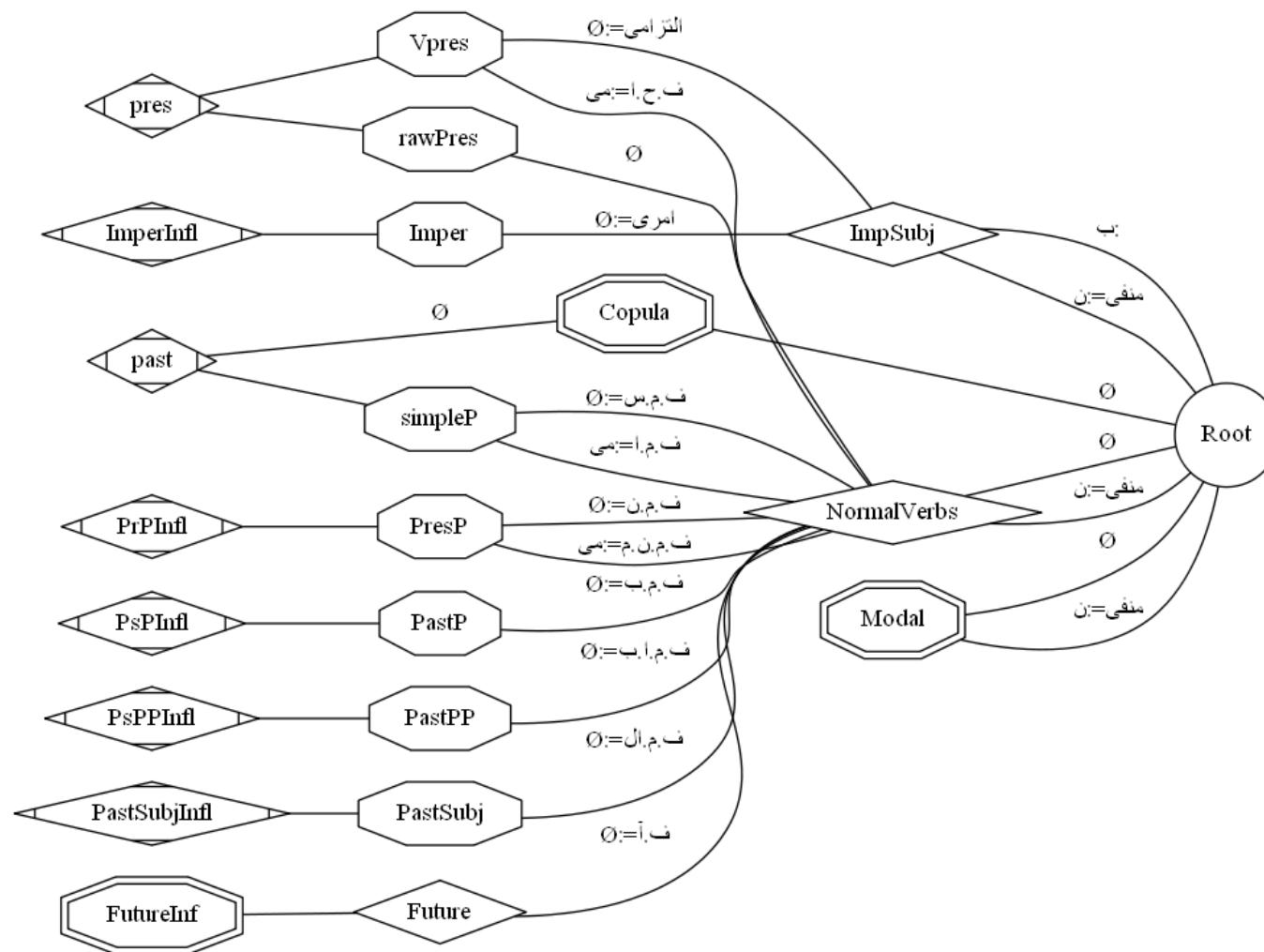
۱-۴- فعل‌های ساده

ساختار کلی لایه‌های زیرساخت فعل‌های رسمی آنطور که در صورت‌بندی صرف مبدل زیراکس^۱ ساخته می‌شود در شکل ۱-۵ آمده است. البته این ساختار تا حدودی مفهومی است و مقداری از جزئیاتی که در پیاده‌سازی واقعی آمده در آن درج نشده است. در عوض جزئیات هر بخش، طوری که ساختار را کامل و روشن توضیح دهد در ادامه آمده است.

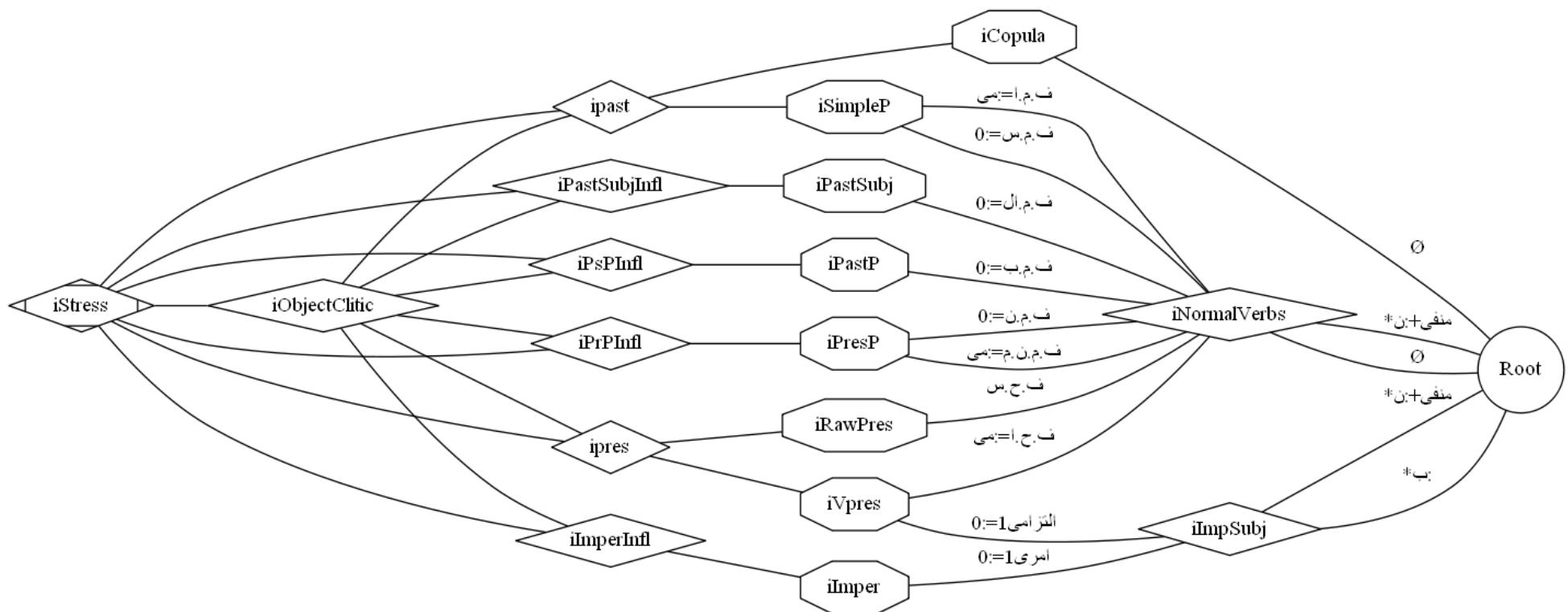
ساختار مفهومی کلی فعل‌های ساده غیر رسمی هم در شکل ۱-۶ آمده است.

فعل‌های رسمی هم در مسیر ساخت فعل‌های غیر رسمی ساخته می‌شوند، اما تغییرات مروط به فعل‌های رسمی همانی است که در گره‌های شکل ۱-۵ وجود دارد. اگر تغییر بعد از آن در این فعل‌ها رخ دهد، آنها را تبدیل به فعل غیر رسمی می‌کند.

^۱Xerox Finite State Morphology Formalism



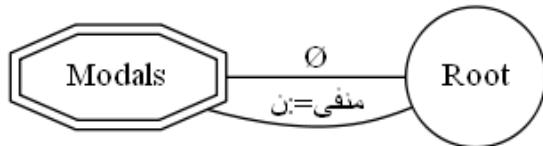
شکل ۱-۵ ساختار لایه‌های سازنده فعل رسمی ساده



شکل ۱-۶ ساختار لایه‌های سازنده فعل ساده غیر رسمی

فعل و جهی

این ساخت یکی از ساده‌ترین ساختهای است. در ابتدا از ریشه، وند تهی و نفی را می‌پذیرد و با ساختن فعل‌های و جهی گوناگون در همان لایه به اتمام می‌رسد. این ساخت در شکل ۷-۱ و پیاده‌سازی آن در شکل ۸-۱ قابل ملاحظه است.



شکل ۷-۱ لایه‌های زیرساخت فعل و جهی

LEXICON Root

```

Modals;
* منفی+ن Modals;
# ف. و=توان: بتوان
# منفی+ف. و=توان: نتوان
...
Aux;
* منفی+ن Aux;
...

```

LEXICON Modals

```

; # ف. و=میبایست+رسمی: می×بایست
; # ف. و=میبایستی+رسمی: می×بایستی
; # ف. و=بایست+رسمی: بایست
; # ف. و=بایستی+رسمی: بایستی
; # ف. و=باید+رسمی: باید
; # ف. و=میباید+رسمی: می×باید
; # ف. و=میتوان+رسمی: می×توان

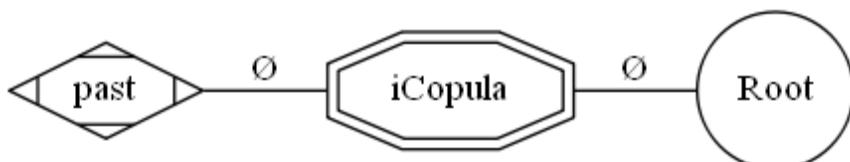
```

شکل ۸-۱ کد ساخت فعل و جهی

ساخت بتوان و نتوان آنهم به علت پذیرفتن پیشوندهای مجزا، به شکل انفرادی در ریشه قرار می‌گیرد.

اسناد

ساخت اسناد هم ساده و محدود به سه گره است (شکل ۹-۱). برخی از کلمات در گره اسناد و برخی در گره بعدی که شناسه فعل گذشته است ساخته می‌شوند. به همین دلیل هر دو گره، گره پایانی هستند (شکل ۱۰-۱).



شکل ۹-۱ لایه‌های زیرساخت اسناد

```

LEXICON Root
...
Copula;

LEXICON iCopula
@P.FORMAL.ON@ت=اسناد هست:@P.FORMAL.ON@ت ipast;
@P.FORMAL.ON@منفی+اسناد=هست:@P.FORMAL.ON@ت نیست ipast;
# اسناد=است+رسمی: است
# اسناد=ست+رسمی: ست

LEXICON past
+ م*:۱ش#;
+ ی*:۲ش#;
+ ۳ش: #;
+ یم*:۴ش#;
+ ید*:۵ش#;
+ ند*:۶ش#;

```

شکل ۱۰-۱ کد ساخت اسناد

فعل لازم و متعددی

در ساختهای غیر رسمی فعلهای متعددی می‌توانند واژه‌بست مفعولی و سپس واژه‌بستهای محاوره بپذیرند. در حالی که فعلهای لازم تنها می‌توانند واژه‌بستهای محاوره و فاعلی بپذیرند. در برخی ساختهای گذشته فعلهای لازم هم مانند فعلهای متعددی می‌توانند واژه‌بست فاعلی بپذیرند.

از همان گره ابتدایی بن‌های رسمی لازم و متعددی و بن‌های غیر رسمی لازم و متعددی به گره‌های جداگانه هدایت می‌شود (شکل ۱۱-۱). گره‌های غیر رسمی در ابتدای نام آنها حرف *n* قرار دارد، مثل *ipast* در مقابل گره رسمی *.past*. گره فعلهای لازم در انتهای نام آنها *InTrans* به معنای لازم قرار دارد، مثل *ipastInTrans* در مقابل *ipast* که گره فعلهای متعددی است.

```

LEXICON iSimpleP
بن های غیر رسمی متعدد!
آگا هوند ipast;
افشوند ipast;
پاشوند ipast;
...
تونست ipastInTrans; بن های غیر رسمی لازم!
در موند ipastInTrans;
دو بید ipastInTrans;
...
آزمود past; بن های رسمی متعدد!
آغازید past;
آغشت past;
...
اندیشید pastInTrans; بن های رسمی لازم!
انگارید pastInTrans;
انگاشت pastInTrans;
...

LEXICON ipastInTrans
+م*:۱ش iStress;
+م*:۲ش iStress;
+۳:۰ش iStress;
@C.FORMAL@+۳ش:@C.FORMAL@*ش iStress;
+م*:۴ش iStress;
@C.FORMAL@+۵ش:@C.FORMAL@*ین iStress;
+د*:۵ش iStress;
@C.FORMAL@+۶ش:@C.FORMAL@*ن iStress;
+ش*:۷ن iStress;

LEXICON pastInTrans
@P.FORMAL.ON@:@P.FORMAL.ON@ ipastInTrans;

LEXICON ipast
+م*:۱ش iObjectClitic;
+م*:۲ش iObjectClitic;
+۳:۰ش iObjectClitic;
@C.FORMAL@+۳ش:@C.FORMAL@*ش iStress;
+م*:۴ش iObjectClitic;
@C.FORMAL@+۵ش:@C.FORMAL@*ین iObjectClitic;
+د*:۵ش iObjectClitic;
@C.FORMAL@+۶ش:@C.FORMAL@*ن iObjectClitic;
+ش*:۷ن iObjectClitic;

LEXICON past
@P.FORMAL.ON@:@P.FORMAL.ON@ ipast;

```

شکل ۱۱-۱ جداسازی فعل‌های لازم و متعدد

بن های رسمی بعد از گرفتن نشانه فرامتنی رسمی بودن (@@P.FORMAL.ON@) به گره فعل غیر رسمی منتقل می‌شود. البته این انتقال با تفکیک فعل لازم و متعدد صورت می‌پذیرد. فعل‌های لازم، مستقیم به گره واژه‌بست‌های محاوره (iStress) منتقل می‌شود، و فعل‌های لازم ابتدا به گره واژه‌بست مفعولی و سپس به گره واژه‌بست محاوره منتقل می‌شوند. هر دو فعل لازم و متعدد در ساخت سوم شخص مفرد می‌توانند واژه‌بست فاعلی بپذیرند، بنابراین در این ساخت‌ها بعد از گرفتن این واژه‌بست (حتی اگر بن فعل متعدد باشد)، به گره واژه‌بست‌های محاوره منتقل می‌شود.

شناسه‌های متفاوت مضارع

بن‌های مضارع بر خلاف بن‌های ماضی که همواره منتهی به همخوان هستند، به چهار گروه منتهی می‌شود. گروه اول که همان همخوان‌ها هستند و شناسه‌های عادی می‌پذیرند. دسته‌های دیگر به واکه‌های **واو**, **آ** و **ی** منتهی می‌شود. در قسمت اتصال شناسه‌های مضارع هر کدام از این دسته‌ها، یک گره مجزا هستند که با مسیر متفاوت، به بن‌های مربوط به خود متصل می‌شوند (شکل ۱۲-۱).

```

برای بن‌های منتهی به همخوان !
+م*:۱ش iStress;
+ی*:۲ش iStress;
+یم*:۴ش iStress;
@C.FORMAL@+۵ش:@C.FORMAL@* ين iStress;
...
LEXICON iALEFpresInTrans !
+یم:۱ش iStress;
@C.FORMAL@+۱ش:@C.FORMAL@* م iStress;
+ی*:۲ش iStress;
@C.FORMAL@+۴ش:@C.FORMAL@* يم iStress;
@C.FORMAL@+۵ش:@C.FORMAL@* ين iStress;
@C.FORMAL@+۶ش:@C.FORMAL@* يين iStress;
...
LEXICON iVAVpresInTrans
+یم:۱ش iStress;
+ی*:۲ش iStress;
+یم*:۴ش iStress;
@C.FORMAL@+۵ش:@C.FORMAL@* يين*i iStress;
+ش:۵یید iStress;
...
LEXICON iVpres
تون ipresInTrans;
رنج ipresInTrans;
را زا iALEFpres;
ا ALEFpresInTrans;
بو VAVpres; !!smelling
جو VAVpres; ! searching (vowel)
جو pres; ! biting (consonant)
...

```

شکل ۱۲-۱ کد شناسه‌های متفاوت مضارع

هر بن فعل با توجه به حرف منتهی به آن به گره مربوط به خود منتقل می‌شود و شناسه مناسب می‌پذیرد. مثلاً بن **جو** به معنای جستجو کردن منتهی به واکه **واو** است و به بخش VAVpres منتقل می‌شود. اما بن **جو** به معنای جویدن منتهی به همخوان **واو** است و به گره شناسه‌های همخوان منتقل می‌شود.

غیر از ساخت‌های مضارع التزامی و اخباری که شناسه‌های متفاوتی برای همخوان و واکه پایانی بن خود دارند، فعل‌های امری نیز برای این بن‌ها شناسه‌های متفاوتی دارند.

فعل امری و مضارع التزامی

تفاوت فعل امری با فعل مضارع التزامی یکی در ساخت دوم شخص مفرد امری است که مابازایی در ساخت مضارع التزامی ندارد و دیگری در عدم پذیرش واژه‌بست محاوره هم است. یعنی فعل امری برخلاف فعل مضارع التزامی این واژه‌بست را نمی‌پذیرد.

دو ساخت دیگر فعل امری، یعنی اول شخص جمع و دوم شخص جمع، همان ساخت‌های معادل مضارع التزامی هستند. برای حالتی که در نظر داشته باشیم، قاعده اضافی برای کلمات (مستقل از بافت) تولید نشود، می‌توانیم راه تولید فعل‌های امری غیر از مفرد را ببندیم. یعنی تنها فعل امری مفرد تولید شود.

```
LEXICON iImperInfl
: iObjectClitic;
+ يَدْ : شَهْ ObjectImperNoImperGate;
@C.FORMAL@+هَنْ:@C.FORMAL@ يَنْ ObjectImperNoImperGate;
+ يَمْ : شَهْ ObjectImperNoImperGate;

LEXICON ObjectImperNoImperGate
@R.IMPERATIVE@:@R.IMPERATIVE@ iObjectClitic;
!@D.IMPERATIVE@:@D.IMPERATIVE@ iObjectClitic;
```

شکل ۱۳-۱ کد محدود کردن فعل امری به ساخت مفرد

همانطور که در شکل ۱۳-۱ دیده می‌شود، فعل مفرد بدون هیچ محدودیتی مستقیم به گره واژه‌بست مفعولی منتقل می‌شود. ساخت‌های جمع به گره ObjectImperNoImperGate منتقل می‌شوند. در این گره با توجه به نیاز می‌توان یکی از مسیرها را حذف و دیگری را باز گذاشت. مثلاً اگر خط دارای نشانه فرامتنی @R.IMPERATIVE@ باز باشد و خط دیگر حذف شود، فعل امری جمع نیز تولید می‌شود. اما اگر خط دارای نشانه فرامتنی @D.IMPERATIVE@ باز باشد و خط دیگر حذف شود، فعل امری جمع از مبدل حذف می‌شود و تنها فعل‌های امری مفرد باقی می‌مانند و برای شناسایی فعل‌های امری جمع باید در پس‌پردازش نسبت به بافت، آنها را از فعل‌های مضارع التزامی تمیز داد.

فعل خوا

به علت محدودیت این بن فعل در پیوستن به شناسه‌های مضارع و امری، این ساخت استثناءً به شکل مجزا پیاده‌سازی شده است.

```

LEXICON iNormalVerbs
@P.IMPERATIVE.ON@:@P.IMPERATIVE.ON@ exceptionalImp;
: exceptionPres;
...

LEXICON exceptionPres
اَلْتَزَامِيٌّ=خُواهُ: بُخُوا presException;
مَنْفَى+الْتَزَامِيٌّ=خُواهُ: نُخُوا presException;
فَحْشَةً=خُواهُ: مُخُوا presException;
مَنْفَى+فَحْشَةً=خُواهُ: نُمَخَّوا presException;

LEXICON exceptionalImp
اَمْرِيٌّ=خُواهُ: بُخُوا iALEFImperInfl;
مَنْفَى+اَمْرِيٌّ=خُواهُ: نُخُوا iALEFImperInfl;
...

LEXICON presException
+ م*: ۱ ش iObjectClitic;
+ ش*: ۲ ي iObjectClitic;
+ ش*: ۳ ي iObjectClitic;
+ د*: ۴ ش و فَا عَلَى iStress;
+ ي*: ۵ ش iObjectClitic;
+ ين*: ۶ ش iObjectClitic;
+ يد*: ۷ ش iObjectClitic;
+ ن*: ۸ ش iObjectClitic;

```

شکل ۱۴-۱ کد پیاده سازی فعل خوا

استثنایات امری

برخی بن‌های مضارع که بدون پیشوند ب فعل التزامی / امری می‌سازند جدا از فعل‌های امری دیگر ساخته می‌شوند (شکل ۱۵-۱). البته شناسه‌ها در همان گره فعل‌های امری به آنها افزوده می‌شود.

```

LEXICON iNormalVerbs
@P.IMPERATIVE.ON@:@P.IMPERATIVE.ON@ exceptionalImp;
...

LEXICON exceptionalImp
@P.FORMAL.ON@: اَمْرِيٌّ=بَاشِ iImperInfl;
@P.FORMAL.ON@: اَمْرِيٌّ=دَارِ iImperInfl;
@P.FORMAL.ON@: اَمْرِيٌّ=كَنِ iImperInfl;
@P.FORMAL.ON@: اَمْرِيٌّ=شَوِ iImperInfl;

```

شکل ۱۵-۱ کد ساخت فعل‌های امری بدون پیشوند ب

فعل مضارع ساده

این ساخت هم بن‌های محدودی دارد که در شکل ۱۶-۱ مشخص شده است.

```

LEXICON iRawPres
    دار pres;
    باش presInTrans;
    کن pres;
    ش ipres;
    @P.LONEFRMSTEM.ON@ه@P.LONEFRMSTEM.ON@ pres;
    @P.LONEFRMSTEM.ON@خواه@P.LONEFRMSTEM.ON@ pres;
    @P.LONEFRMSTEM.ON@شو@P.LONEFRMSTEM.ON@ presInTrans;

```

شکل ۱۶-۱ کد ساخت فعل مضارع ساده

محدودیت برخی بن‌های غیر رسمی

برخی بن‌های غیر رسمی فعلِ سادهِ مضارع و پیشوندیِ مضارع در ساخت امریِ مفرد و سوم شخصِ مفردِ مضارع (بعضی ساخت‌ها) استفاده نمی‌شوند (جدول ۸-۴ و جدول ۹-۴). به این بن‌ها نشانه‌های فرامتنی افزوده شده است (شکل ۱۷-۱).

```

LEXICON iVpres
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ ipresInTrans;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ ipresInTrans;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ ipres;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ ipres;
...
LEXICON iImper
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ iImperInflInTrans;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ iImperInflInTrans;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ iImperInfl;
@P.LONEINFSTEM.ON@{@P.LONEINFSTEM.ON@ iImperInfl;
...
LEXICON iChainVpresSub
@R.PA@{@R.PA@ NoInformalInflection;
@R.DAR@{@R.DAR@ NoInformalInflection;
@R.VAR@{@R.VAR@ NoInformalInflection;
@R.VA@{@R.VA@ NoInformalInflection;
@R.FORU@{@R.FORU@ NoInformalInflection;
...
LEXICON NoInformalInflection
@P.LONEINFSTEM.ON@:@P.LONEINFSTEM.ON@ ipresInTrans;

LEXICON iChainImper
@R.DAR@{@R.DAR@ ImpersecondIREG;
@R.VAR@{@R.VAR@ ImpersecondIREG;
@R.VA@{@R.VA@ ImpersecondIREG;
@R.FORU@{@R.FORU@ ImpersecondIREG;

LEXICON ImpersecondIREG
@P.LONEINFSTEM.ON@:@P.LONEINFSTEM.ON@ iImperInflInTrans;

```

شکل ۱۷-۱ نشانه گذاری فرمتی برای بن‌های غیررسمی دارای محدودیت

بن‌های پیشوندی از آنجایی که یک نشانه فرامتنی در ابتداء می‌پذیرند، به ناچار می‌بایست نشانه فرامتنی دیگر را (@P.LONEINFSTEM.ON@) در گرهای دیگر به آنها الصاق کرد (ImpersecondIREG) و (NoInformalInflection).

بعد از نشانه گذاری این بن‌ها، در مرحله تصریف با استفاده از نشانه فرامتنی @D.LONEINFSTEM@ راه ساخت آن را در مسیر فعل امری مفرد و بعضی ساختهای مضارع سوم شخص مفرد، می‌بندیم (شکل ۱۸-۱).

```

LEXICON iImperInflInTrans
@D.LONEINFSTEM@:@D.LONEINFSTEM@ iImperInflInTransFinal;
+يد*:يـ ColloqImperNoImperGate;
@C.FORMAL@+ـــ: @C.FORMAL@ يـــ ColloqImperNoImperGate;
+يم*:ـــ ColloqImperNoImperGate;

LEXICON ipresInTrans
@C.FORMAL@+ـــ: @C.FORMAL@*ـــ iStressThirdPerson;
@D.LONEINFSTEM@+ـــ: @D.LONEINFSTEM@*ـــ iStress;
...

```

شکل ۱۸-۱ کد بستن مسیر ساخت‌های غیر مجاز برای برخی بن‌های غیر رسمی

نشانه فرامتنی @D.LONEINFSTEM@ باعث می‌شود فعل امری نظیر بُر، بُش و فعل سوم شخص مفرد مضارع نظیر بُگ، بُشد، می‌رد، بُدد تولید نشود.

محدودیت برخی بن‌های رسمی

در ساخت‌های مضارع، برخی بن‌های رسمی در ساخت سوم شخص مفرد، شناسه غیر رسمی را نمی‌پذیرند (شکل ۱۹-۱). به این بن‌ها، نشانه فرامتنی @P.LONEFRMSTEM.ON@ افزوده می‌شود. از طرف دیگر در گره ObjectCiticThirdPerson با استفاده از نشانه فرامتنی @D.LONEFRMSTEM@ مسیر برای این بن‌ها بسته شده است. در گره افزودن شناسه به بن‌های مضارع، مسیرهایی که ساخت سوم شخص محاوره غیر رسمی دارند، به این گره هدایت می‌شوند تا در صورت وجود این قبیل بن‌های رسمی، مسیرشان مسدود شود.

```

LEXICON iVpres
@P.LONEFRMSTEM.ON@ پژوه@P.LONEFRMSTEM.ON@ pres;
@P.LONEFRMSTEM.ON@ نزرا@P.LONEFRMSTEM.ON@ pres;
@P.LONEFRMSTEM.ON@ جه@P.LONEFRMSTEM.ON@ presInTrans;
@P.LONEFRMSTEM.ON@ رون@P.LONEFRMSTEM.ON@ presInTrans;! Going
@P.LONEFRMSTEM.ON@ دده@P.LONEFRMSTEM.ON@ pres;
@P.LONEFRMSTEM.ON@ خواه@P.LONEFRMSTEM.ON@ pres;
@P.LONEFRMSTEM.ON@ شو@P.LONEFRMSTEM.ON@ presInTrans;
...
LEXICON iObjectCliticThirdPerson
@D.LONEFRMSTEM@:@D.LONEFRMSTEM@ iObjectClitic;

LEXICON presInTrans
@C.FORMAL@+ڙ: @C.FORMAL@* ه iObjectCliticThirdPerson;
@D.LONEINFSTEM@+ڙ: @D.LONEINFSTEM@* د iObjectClitic;
@P.TINTER.ON@+ت: ه* ڙ@P.TINTER.ON@ iObjectCliticThirdPerson;
...

```

شکل ۱۹-۱ کد محدود سازی برخی بن‌های رسمی

حرف ت واسط

در ساختهای سوم شخصی متعددی که واژه‌بست مفعولی می‌پذیرند، ممکن است برای تداخل نکردن شناسه و واژه‌بست مفعولی حرف واسط ت استفاده شود. از آنجایی که بیشتر ساختهای در گره افزودن شناسه با نشانه‌های فرامتنی دیگر اشغال شده، به ناچار گره‌ای مستقل برای ت واسط به نام Tintv طراحی شده است (شکل ۲۰-۱).

```

LEXICON Tintv
@P.TINTER.ON@+ت:@P.TINTER.ON@* ت iObjectClitic;

LEXICON ipast
@C.FORMAL@+ڙ: @C.FORMAL@ Tintv;
...

LEXICON iALEFpres
@C.FORMAL@+ڙ*: ڙ@C.FORMAL@ Tintv;
...

LEXICON iObjectClitic
@C.FORMAL@+۱: @C.FORMAL@* م iStress;
@C.FORMAL@+۲: @C.FORMAL@* ت iStress;
@D.TINTER@:@D.TINTER@ iStress;

```

شکل ۲۰-۱ کد پیاده‌سازی ت واسط

پس از افزودن شدن حرف ت به فعل (پس از شناسه) در گره Tintv، فعل به گره واژه‌بست مفعولی می‌رود. از آنجایی که این حرف زمانی اضافه می‌شود که واژه‌بستی در کار باشد، پس حتماً باید در گره واژه‌بست مفعولی، واژه‌بستی به فعل اضافه شود. برای اینکه فعل بدون دریافت واژه‌بست مفعولی از این قسمت خارج نشود، در قسمت تهی از نشانه

فرامتنی @D.TINTER استفاده شده است، تا مسیر حرکت بدون واژه‌بست مسدود شود. بنابراین فعل با دریافت یکی از واژه‌بست‌های دیگر از این گره خارج شده و به گره واژه‌بست‌های محاوره می‌رود (iStress).

فعل‌های ناقص

فعل‌های ناقص در ساخت‌های مضارع به این دلیل که همواره پیشوند می‌پذیرند (ب در مضارع التزامی و امری، می در مضارع اخباری و نفی)، در همه این ساخت‌ها استفاده می‌شوند. بنابراین در مضارع این فعل‌ها ناقص نیستند و به شکل کامل در همه ساخت‌ها صرف می‌شوند. بن‌های این فعل‌ها در بخش واژگان این فعل‌ها مانند سایر بن‌ها قرار دارد.

در فعل‌های ماضی اما ساخت‌هایی وجود دارد که در آنها هیچ پیشوندی استفاده نمی‌شود. برای تمایز بخشی به ساخت‌هایی که دارای پیشوند هستند، در ابتدای این ساخت‌ها نشانه فرامتنی @P.SPECINFL.ON استفاده شده است (شکل ۲۱-۱). در قسمت افزوده شدن بن به ساختمان فعل نیز این بن‌های ناقص با نشانه فرامتنی @R.SPECINFL.ON تنها به ساخت‌های اضافه می‌شوند که نشانه فرامتنی @P.SPECINFL.ON را دارند. بنابراین این بن فعل‌ها تنها در ساخت‌های دارای پیشوند استفاده می‌شوند. در قسمت افزودن بن‌ها نیز، قرار دادن شکل اصلی بن فعل (مثل گذاشت: ذاشت) باعث می‌شود در قاعده سازنده فعل، بن اصلی و نه بن تغییر یافته قرار بگیرد.

```

LEXICON iNormalVerbs
· := ف.م. iSimpleP;
· := ف.م.ن. iPrestP;
@P.SPECINFL.ON@ . =:@P.SPECINFL.ON@ می × iSimpleP;
@P.SPECINFL.ON@ .م.ن. =:@P.SPECINFL.ON@ می × iPrestP;
...

LEXICON iSimpleP
@R.SPECINFL@ گذاشت: ذاشت; @R.SPECINFL@ ipast;
@R.SPECINFL@ نشست: شست; @R.SPECINFL@ ipastInTrans;
@R.SPECINFL@ شوند: شوند; @R.SPECINFL@ ipastInTrans;
@R.SPECINFL@ انداخت: نداخت; @R.SPECINFL@ ipast;
@R.SPECINFL@ افتاد: فتاد; @R.SPECINFL@ ipastInTrans;
@R.SPECINFL@ افتاد: وفتاد; @R.SPECINFL@ ipastInTrans;
@R.SPECINFL@ اورد: ورد; @R.SPECINFL@ ipast;
@R.SPECINFL@ اومد: ومد; @R.SPECINFL@ ipastInTrans;
انداخت past;
آراست past;
...

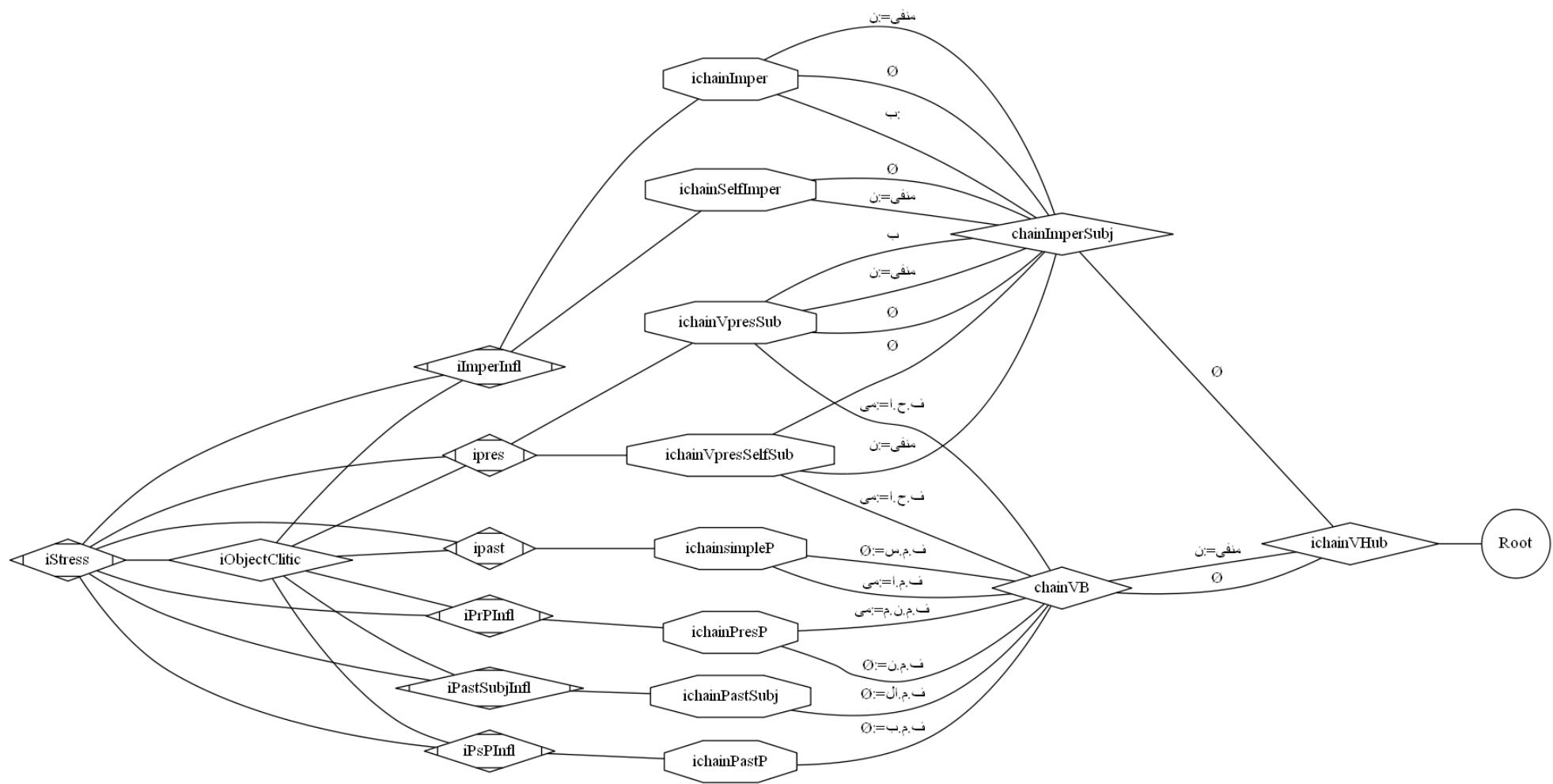
LEXICON iPrestP
@R.SPECINFL@ گذاشت: ذاشت; @R.SPECINFL@ iPrPInfl;
@R.SPECINFL@ نشست: شست; @R.SPECINFL@ iPrPInflInTrans;
@R.SPECINFL@ شوند: شوند; @R.SPECINFL@ iPrPInflInTrans;
@R.SPECINFL@ انداخت: نداخت; @R.SPECINFL@ iPrPInfl;
@R.SPECINFL@ افتاد: فتاد; @R.SPECINFL@ iPrPInflInTrans;
@R.SPECINFL@ افتاد: وفتاد; @R.SPECINFL@ iPrPInflInTrans;
@R.SPECINFL@ اورد: ورد; @R.SPECINFL@ iPrPInfl;
@R.SPECINFL@ اومد: ومد; @R.SPECINFL@ iPrPInflInTrans;
آغشت PrPInfl;
آگاهاند PrPInfl;
...

```

شکل ۲۱-۱ کد پیاده‌سازی فعل‌های ناقص

۱-۵- فعل پیشوندی

ساختمان کلی این فعل‌ها مانند فعل‌های ساده است. ساختار کلی و مفهومی این فعل‌ها در شکل ۳۲-۱ آمده است. تفاوت آنها در پیشوندی است که از گره ریشه شروع می‌شود و با الصاق نشانه فرامتنی مناسب برای هر پیشوند در سمت بن فعل، پیشوند را به بن مربوط متصل می‌کند. بعضی موارد دیگر ممکن است به دلیل وجود نشانه‌های فرامتنی در جای گره‌های پیشوندی، باعث شود پیاده‌سازی آنها کمی متفاوت‌تر از فعل‌های ساده صورت پذیرد.



شکل ۱ ۲۲- لایه‌های زیرساخت فعل پیشوندی غیر رسمی

اتصال پیشوند به بن فعل

به هر پیشوند متناسب با نام خود، نشانه‌ای فرامتنی اضافه می‌شود. مثلاً به پیشوند **در** که با چند بن متفاوت می‌تواند فعل پیشوندی بسازد، نشانه **@P.DAR.ON@** الصاق می‌شود. از طرف دیگر بن فعل پیشوندی‌ای که با این پیشوند، فعل می‌سازد نشانه **@R.DAR@** دارد که باعث می‌شود تنها با این پیشوند ترکیب شود (شکل ۲۳-۱).

```

LEXICON Root
@P.VAR.ON@پ.ور*: @P.VAR.ON@پ.ور * iChainVHub;
@P.VA.ON@پ.وا*: @P.VA.ON@پ.وا * iChainVHub;
@P.DAR.ON@پ.در*: @P.DAR.ON@پ.در * iChainVHub;
@P.BAR.ON@پ.بر*: @P.BAR.ON@پ.بر * iChainVHub;
...

LEXICON iChainVHub
iChainVB;
...

LEXICON iChainVB
می: ا.خ.ف*iChainVpresSub;
م.م.ف*iChainSimpleP;
...

LEXICON iChainVpressSub
@R.DAR@ار@R.DAR@ ipres;
@R.DAR@ار@R.DAR@ iALEFpresInTrans;
@R.BAZ@رسون@R.BAZ@ ipres;
@R.BAZ@خون@R.BAZ@ ipres;
@R.PA@پ@R.PA@ NoInformalInflection;
@R.DAR@ر@R.DAR@ NoInformalInflection;
...

```

شکل ۲۳-۱ کد اتصال پیشوند به بن فعل پیشوندی

بنابراین پیشوند **در**، فعل‌های **دار**، **dra** و دررفتن را می‌سازد.

محدودیت برخی بن‌های غیر رسمی پیشوندی

مانند فعل‌های ساده در این بخش نیز بن‌هایی هستند که در سوم شخص مفرد مضارع، شناسه رسمی نمی‌پذیرند. از آنجایی که هر مسیر فعل‌های پیشوندی نشانه فرامتنی قرار دارد، به ناچار نشانه فرامتنی برای علامت گذاری این بن‌ها در گره دیگری قرار داده شده است (شکل ۲۴-۱).

```
LEXICON NoInformalInflection
@P.LONEINFSTEM.ON@:@P.LONEINFSTEM.ON@ ipresInTrans;
```

```
LEXICON iChainVpressSub
@R.PA@_R.PA@ NoInformalInflection;
@R.DAR@_R.DAR@ NoInformalInflection;
@R.VAR@_R.VAR@ NoInformalInflection;
@R.VA@_R.VA@ NoInformalInflection;
@R.FORU@_R.FORU@ NoInformalInflection;
...
```

```
LEXICON ipresInTrans
@D.LONEINFSTEM@+۲۰:@D.LONEINFSTEM@* د iStress;
...
```

شکل ۲۴-۱ کد محدود کردن ساخت بعضی بن‌های غیر رسمی پیشوندی

محدودیت برخی بن‌های رسمی پیشوندی

برای بن‌های رسمی نیز مانند فعل‌های ساده محدودیت‌هایی وجود دارد که مانند قسمت قبل به دلیل ازدیاد نشانه‌های فرامتنی باعث شده، نشانه‌های جدید در گره‌های مستقل قرا داده شود (شکل ۲۵-۱).

```
LEXICON PresNotForInformalThirdPerson
@P.LONEFRMSTEM.ON@:@P.LONEFRMSTEM.ON@ presInTrans;
```

```
LEXICON iChainVpressSub
@R.DAR@_R.DAR@ PresNotForInformalThirdPerson;
@R.VAR@_R.VAR@ PresNotForInformalThirdPerson;
@R.VA@_R.VA@ PresNotForInformalThirdPerson;
@R.FORU@_R.FORU@ PresNotForInformalThirdPerson;
```

```
LEXICON iStressThirdPerson
@D.LONEFRMSTEM@:@D.LONEFRMSTEM@ iStress;
```

```
LEXICON ipresInTrans
@C.FORMAL@+۳۰:@C.FORMAL@* د iStressThirdPerson;
...
```

شکل ۲۵-۱ کد محدود کردن ساخت بعضی بن‌های رسمی پیشوندی

استثنای فعل پیشوندی امری

برخی فعل‌های پیشوندی غیر رسمی در ساخت امری مانند فعل‌های ساده از بن رسمی استفاده می‌کنند. از آنجایی که ساخت پیشوندی آنها معادل رسمی ندارد بنابراین این ساخت مفرد را که به شکل رسمی ساخته می‌شود می‌بایست جداگانه پیاده کرد (شکل ۲۶-۱).

LEXICON iChainVHub

```
...
@R.PA@ : امری=شو+مفرد@ ;
@R.PA@ : امری= بشو+مفرد@ ;
@R.PA@ : منفی+امری=نشو+مفرد@ ;
@R.VAY@ : امری=ستا+مفرد@ ;
@R.VAY@ : منفی+امری=نستا+مفرد@ ;
@R.VAY@ : امری=سا+مفرد@ سا# ;
@R.VAY@ : منفی+امری=نسا+مفرد@ نسا# ;
```

شکل ۲۶-۱ کد ساخت استثنای امری پیشوندی

فعل‌های ناقص پیشوندی

فعل‌های ناقص پیشوندی متفاوت از فعل‌های ناقص ساده پیاده می‌شوند. از لحاظ رفتار هم همانطور که در جدول ۱۳-۴ و جدول ۱۴-۴ ملاحظه می‌شود، اندکی متفاوت از فعل‌های ساده هستند.

فعل‌های مضارع التزامی و امری پیشوندی دو دسته هستند؛ یا پیشوند التزامی ب را به صورت اختیاری می‌پذیرند و یا اصلاً نمی‌پذیرند. از طرف دیگر فعل پیشوندی غیر رسمی بازگذاشتن در این ساخت‌ها حتماً باید پیشوند ب بگیرد. بنابراین ساخت‌های مضارع و امری آن جداگانه تعریف می‌شود (شکل ۲۷-۱). شناسه‌ها نیز در همان گره‌های مربوط به امری و مضارع به فعل افزوده می‌شود.

```

LEXICON presZar
@R.BAZ@ارذ@R.BAZ@ ipres;
LEXICON ImperZar
@R.BAZ@ارذ@R.BAZ@ iImperInfl;

LEXICON iChainImpSubj
@P.IMPERATIVE.ON@=ب امری * @P.IMPERATIVE.ON@ ImperZar;
@P.IMPERATIVE.ON@=ن منفی + امری * @P.IMPERATIVE.ON@ ImperZar;
@B=التزامی presZar;
@N=منفی + التزامی presZar;
@F.H.=می presZar;
@X=منفی + ف.ح.ا presZar;
...

```

شکل ۲۷-۱ کد ساخت فعل ناقص باز-ذار

فعل گذشته بازگذاشتن نیز به این دلیل که تنها فعل ناقصی است که تصریف اینگونه دارد، همه ساختهایش جداگانه ایجاد می‌شود (شکل ۲۸-۱).

```

LEXICON iChainVHub
@R.BAZ@=منفی+ف.م.=ذاشت ipast;
@R.BAZ@=منفی+ف.م.=ذاشت ipast;
@R.BAZ@=منفی+ف.م.=ذاشت ipast;
@R.BAZ@=منفی+ف.م.=ذاشت iPrPInfl;
@R.BAZ@=منفی+ف.م.=ذاشت iPrPInfl;
@R.BAZ@=منفی+ف.م.=ذاشت iPrPInfl;
@R.BAZ@=منفی+ف.م.=ذاشت iPSPInfl;
@R.BAZ@=منفی+ف.م.=ذاشت iPastSubjInfl;
...

```

شکل ۲۸-۱ کد ساخت فعل ناقص باز-گذاشت

فعالهای دیگر پیشوندی ناقص، مثل سایر بن‌ها به واژگان بن‌ها اضافه می‌شوند (شکل ۲۹-۱). ساختهای غیر مجاز آنها بسیار اندک و کاملاً نشان‌دار^۱ است طوری که به راحتی با یک قاعده بازنویسی فوما می‌توان آنها را از ساخت خارج ساخت (شکل ۳۰-۱). این قاعده هر رشته‌ای را که بخشی از آن با محتوای این قاعده همخوانی دارد، حذف می‌کند.

^۱ Marked

```

LEXICON iChainVpressSub
@R.DAR@فت@R.DAR@ ipresInTrans;
...
LEXICON iChainSimpleP
@R.DAR@افتاد:@R.DAR@ ipastInTrans;
@R.DAR@ورد:@R.DAR@ ipast;
@R.DAR@ومد:@R.DAR@ ipastInTrans;
...

```

شکل ۱-۲۹ فعال‌های پیشوندی ناقصی که مثل فعل‌های کامل تعریف می‌شوند

```
define VBsoftRefine ~$ [ [DAL RE] BOUNDS [ (VAV) FE TE] ];
```

شکل ۱-۳۰ قاعده بازنویسی برای حذف ساختهای غیرمجاز فعل‌های ناقص پیشوندی

۱-۶- قواعد نگارشی

قواعد نگارشی مرحله پایانی برای تبدیل زیر ساخت به روساخت در مبدل حالت متناهی است (دنیل جورافسکی و جیمز مارتین، ۲۰۰۸). قواعد نگارشی افعال برای فعل‌های ساده و پیشوندی یکسان اعمال می‌شود.

تغییرات واژه‌بست هم در انتهای بخش افعال

واژه‌بست هم بعد از حرف ه باید فاصله بگیرد و به واژه هم تبدیل شود.

```
define VBhamRefine "&" -> ["^" HEH] || HEH _ MIM BOUNDS;
```

شکل ۱-۳۱ قاعده تبدیل واژه‌بست هم به واژه

حرفی میانجی در آغاز

حرفی بعد از پیشوندهای التزامی ب و نفی ن در صورتی که بن فعل با ۱ شروع شده باشد، قرار می‌گیرد. همینطور برشی فعل‌های ناقص نیز با اینکه با ۱ شروع نمی‌شوند اما می‌پذیرند (شکل ۱-۳۲). این استثناءها در شرط قاعده گنجاده شده‌اند.

```

define VBiBehRefine "*" -> [YEH "×"] || BOUNDS BEH _  

[[ALEFCNTX \[ .#. | YEH ]] | [VAV MIM DAL] | [(VAV) FE TE] |  

[VAV RE DAL] | [(VAV) FE TE AA DAL]];  

define VBiNehRefine "*" -> [YEH "*"] || BOUNDS NOON _  

[[ALEFCNTX \[ .#. | YEH ]] | [VAV MIM DAL] | [(VAV) FE TE] |  

[VAV RE DAL] | [(VAV) FE TE AA DAL]];

```

شکل ۱-۳۲ قواعد افزودنی میانجی برای پیشوند التزامی و نفی

قسمت [[ALEFCNTX \[.#. | YEH]] برای استثناء کردن فعل **ایستادن** در ساخت التزامی و امری است که میانجی نمی‌پذیرد.

فاصله دادن ه از حرف بعدی

اگر حرف ه قبل از مرز واژه، تکواز، وند و واژه‌بست قرار بگیرد، در صورتی که حرف بعدی به آن بچسبد، می‌بایست آن حرف با فاصله از ه جدا شود. این کار توسط قاعده شکل ۱-۳۳ انجام می‌شود.

```
define VBhehSpace "*" -> "×" || HEH _ \[.#.];
```

شکل ۱-۳۳ قاعده فاصله دادن حرف ه از حرف بعدی

تبديل نشانه‌های مرزنا

در لایه‌های میانی مبدل متناهی از نشانه‌هایی برای مشخص کردن، نیم‌فاصله، فاصله، اتصال و حتی مرز اختصاصی برخی واژه‌بست‌ها استفاده شده است. در لایه آخر تبدیل به روساخت همه این مرزنماها باید به مابازای واقعی خودشان در نوشتار زبان فارسی تبدیل شوند. قواعد این قسمت این کار را انجام می‌دهند.

```

define NACHASB [ " | "ا" | "د" | "ذ" | "ز" | "ژ" | "و" | "هـ" | "ـه" ];
define CONCAT 0;
define SPACE " ";

define putConcat [ "*" | "&" | "¤" | "Æ" | "©" ] -> CONCAT;
define iPutSpace [ "^" ] -> SPACE;
define refineZWNJ ZWNJ -> CONCAT || NACHASB _ ?*;
define iExtraSpace SPACE+ -> SPACE;

```

شکل ۱-۳۴ قواعد تبدیل نشانه‌های مرزنا

قاعده refineZWNJ، نیم‌فاصله را در صورتی که بعد از حروف نچسب فارسی قرار گرفته باشد حذف می‌کند.

پیوست دو

پیاده‌سازی کلمه غیر فعلی

۱-۲ - مقدمه

پیاده‌سازی اصلی این بخش مربوط به ساختار اسم‌هاست. ساختهای دیگر این بخش به صورت کلی یا جزئی از این ساختار استفاده می‌کنند. این ساختار به طور کامل در فصل پنجم تعریف شده است. ابتدا پیاده‌سازی این ساختار به طور کامل توضیح داده می‌شود، سپس سایر ساختارهای این بخش که از آن در ساخت خود استفاده می‌کنند شرح داده می‌شود.

تمام قسم کلمه‌های این بخش غیر از **حرف اضافه** و **حرف اضافه گروهی** که در فایل `iharf.lexc` قرار دارد، به صورت کامل در فایل `inoun.lexc` قرار می‌گیرد.

۲-۲ - موارد عمومی گره‌های بخش غیر فعلی

در این قسمت مواردی که مربوط به همه ساختهای قرار گرفته است. نشانه‌گذاری فرامتنی واژگان براساس حرف-واکه پایانی آنها، شیوه تمایز بخشیدن به قواعد کلمات رسمی و غیر رسمی و طراحی گره‌های واکه آگاه، طوری که برای تک واژه منتهی به واکه تک واژه مناسب انتخاب کند همگی در این بخش قرار دارد.

علامت زدن حرف پایانی واژگان

برای قواعد نگارشی که منجر به تغییر واکه پایانی تک واژه مستقل در هنگام اتصال و یا اضافه شدن واج میانجی می‌گردد، از نشانه‌های فرامتنی^۱ لکس استفاده شده است. منبع واژگان رسمی برای این منظور واژگان زایا (اسلامی و همکاران، ۱۳۸۳) است که برای هر مدخل واژگانی تلفظ آن را نیز به همراه دارد و به راحتی می‌توان کلمات منتهی به حرف-واکه را شناسایی کرد. در مورد واژگان غیر رسمی نیز آنها یی که منتهی به حرف-واکه‌اند مشخص شده است. تمام واژگان بخش غیر فعلی به این ترتیب علامت گذاری شده‌اند.

شیوه علامت‌گذاری این حرف-واکه‌ها (۵، ۱، و) طبق صورت‌بندی لکس مطابق شکل ۱-۲ است.

```
LEXICON ComNN
@P.VOWEL.ALEF@ آب و هو vowelHaComNN;
@P.VOWEL.YEH@ آب و هو اشناسی vowelHaComNN;
@P.VOWEL.VAV@ آب و جا رو vowelHaComNN;
@P.VOWEL.HEH@ آب اکسیژنه vowelHaComNN;
...
```

شکل ۱-۲ نحوه علامت گذاری حرف-واکه پایانی واژگان

رسمی و غیر رسمی

همه واژگان رسمی با نشانه فرامتنی به عنوان واژه رسمی مشخص شده‌اند. واژگان غیر رسمی نشانه‌ای نمی‌پذیرند و همان نداشتن نشانه فرامتنی رسمی مشخص کننده غیر رسمی بودنشان است. در مسیر حرکت بین گره‌ها برای ساختن

^۱ Flag Diacritics

کامل کلمه، در صورتی که از مسیر ساخت غیر رسمی عبور کند، این نشانه حذف شده و آن کلمه تبدیل به ساختی غیر رسمی می‌شود.

در پایان گره FINPOINT برای ساخت رسمی قطعه +رسمی را در قاعده تولید می‌کند و برای غیر رسمی تغییری ایجاد نمی‌کند. بنابراین ساخت رسمی از غیر رسمی متمایز می‌شود.

```
LEXICON FINPOINT
@R.FORMAL@+رسمی:@R.FORMAL@0 #;
@D.FORMAL@:@D.FORMAL@ #;
```

شکل ۲-۲ گره پایانی غیر فعل

واکه آگاه

گره‌های واکه آگاه نسبت به حرف پایانی تکواز قبلي حساسند و با توجه به همخوان و واکه بودن حرف پایانی و حتی با توجه به نوع واکه پایانی (۱، ۵، ۵) آن ساخت مناسبی را ارائه می‌دهند تا از نظر قواعد نگارشی هماهنگ باشند. غیر از دو گره shakhs1 و rabti1 که همواره منتهی به همخوان و یا واکه ی هستند و نیازی به در نظر داشتن حالت‌های مختلف در آنها نیست، بقیه گره‌ها همه‌گی واکه آگاه هستند. مثالی از این نوع ساخت‌ها را می‌توان در مشاهده کرد.

```
LEXICON nam
@R.VOWEL.ALEF@+@R.VOWEL.ALEF@* يَ@R.VOWEL.rabtil; نَمْ
@R.VOWEL.VAV@+@R.VOWEL.VAV@* يِ@R.VOWEL.rabtil;
@R.VOWEL.HEH@+@R.VOWEL.HEH@× اَيِ@R.VOWEL.rabtil;
@R.VOWEL.YEH@+@R.VOWEL.YEH@× اِيَ@R.VOWEL.rabtil;
@D.VOWEL@+@D.VOWEL@* يَ@R.VOWEL.rabtil;
```

شکل ۳-۲ گره‌ای واکه آگاه که در تناسب با حرف پایانی تکواز قبلي تکواز جدید را به آن اضافه می‌کند

در صورتی که تکوازی به پایان یک تکواز واکه آگاه متصل شود، نشانه فرامتنی آن حذف شده و بسته به تکواز جدید، نشانه فرامتنی جدید (در صورتی که منتهی به حرف-واکه باشد) می‌گیرد و یا اضافه نمی‌شود (در صورتی که حرف پایانی همخوان باشد). همانطور که در شکل ۴-۲ مشاهده می‌شود، خط دوم مسیر تکواز منتهی به واکه ۵ است که برای فعال شدن این مسیر نیاز به وجود نشانه فرامتنی واکه ۵ در تکواز قبلي است و به دلیل افزودن تکواز ه به کلمه فقط نشانه فرامتنی واکه را از آن حذف می‌کند، اما در خط سوم نیاز به تکوازی داریم که منتهی به واکه و باشد و بعد از حذف آن نشانه واکه ی را به آن اضافه می‌کنیم زیرا تکواز ای به کلمه افزوده شده است.

-
1. LEXICON rabtil
 2. @R.VOWEL.HEH@@C.VOWEL@+۱: وربطی @R.VOWEL.HEH@@C.VOWEL@× م
 FINPOINT;
 3. @R.VOWEL.VAV@@C.VOWEL@@P.VOWEL.YEH@+۲: وربطی @R.VOWEL.VAV@@
 C.VOWEL@@P.VOWEL.YEH@* ا FINPOINT;
-

شکل ۲-۴ تغییر نشانه فرامتنی به دلیل تغییر واکه پایانی

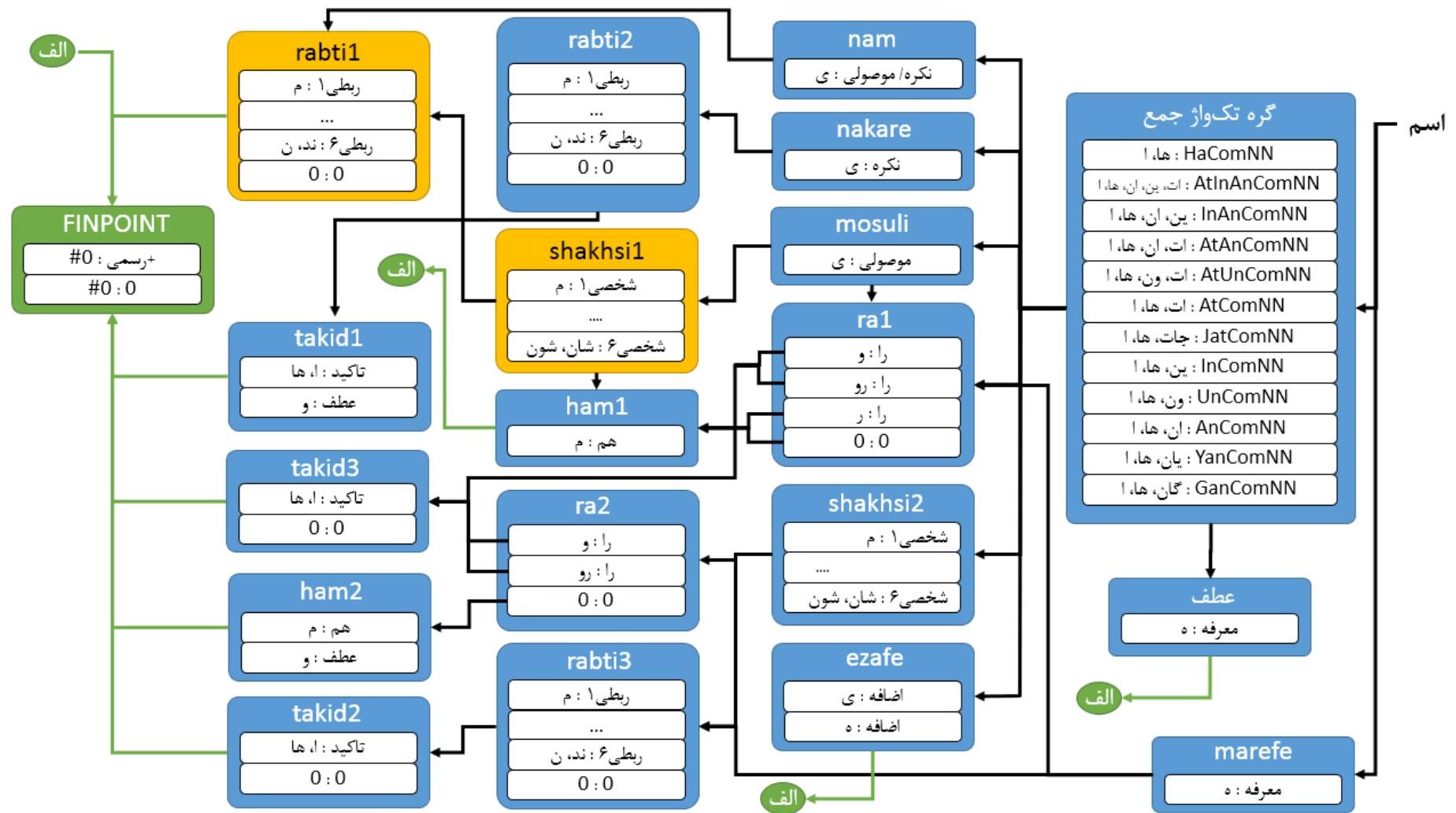
۳-۲ اسم

ساختار اسامی که در فصل کلمات غیر فعلی کامل توضیح داده شده است، مجدداً در اینجا برای درک بهتر نمودار پیاده سازی شده‌اش قرار می‌گیرد. ساختار پیاده سازی شده آن در شکل ۵-۲ قرار دارد. این اجزا در ساختار سایر قسم کلمه‌های این بخش ممکن است به کار رود.

ساختمان‌های مختلف به کار رفته در اسم در شکل ۵-۲ به صورت گویا مشخص است که چه کاری انجام می‌دهد. ساختمان جمع که به شکل خلاصه در تصویر آمده گره‌های متعددی دارد. هر کدام از این گره‌ها تکوازهای جمع مختلفی را پوشش می‌دهند. کلماتی که امکان جمع بسته شدن با تکواز یا تکوازهایی خاص را دارند، هر کدام در این قسمت به گره مناسب انتقال پیدا می‌کنند.

$$\left[
 \begin{array}{c}
 \left(\left(\text{تاكيد} \right) + \left(\text{ربطى} \right) \right) \\
 \left(\left(\text{را} \right) + \left(\text{تاكيد} \right) \right) \\
 \left(\text{هم} \right) \\
 \left(\text{عطف} \right)
 \end{array} \right] + \text{شخصى} \\
 \\
 \left(\left(\text{ربطى} \right) \right] + \left(\text{شخصى} \right) \left[\left(\text{موصولى} \right) + \left(\text{جمع} \right) \right] \\
 \left(\left(\text{هم} \right) \right] + \left(\left(\text{را} \right) + \left(\text{هم} \right) \right) \\
 \left(\left(\text{تعطف} \right) \right] + \left(\left(\text{ربطى} \right) \right] + \left(\text{نكره} \right) \\
 \\
 \left[\left(\text{هم} \right) \right] + \left(\left(\text{را} \right) \right. \\
 \left. \left(\text{تاكيد} \right) \right] + \left(\left(\text{ربطى} \right) \right. \\
 \left. \left(\text{تعطف} \right) \right] + \left(\left(\text{اضافه} \right) \right) \\
 \\
 \left(\left(\text{ربطى} \right) + \left(\text{تاكيد} \right) \right] \\
 \left[\left(\text{هم} \right) \right] + \left(\left(\text{را} \right) \right. \\
 \left. \left(\text{تاكيد} \right) \right] + \left(\text{معرفه} \right)
 \end{array} \right] + \text{اسم}$$

* ساخت را+هم (که با هم واژه‌بست و هم را می‌سازد) در دو حالت تولید می‌شود؛ اول در حالتی که کلمه وند معرفه بپذیرد، در این صورت اگر این وند در نگارش هم نوشته نشود، وجود آن به عنوان جذب تکیه کلمه مشهود است، دوم در حالتی که وند جمع به اسم مفرد افزوده شده باشد. در واقع این ساخت ظاهرا تنها محدود به اسم جمع و معرفه است. هر کدام به صورت تکی تکی می‌توانند برای مفرد بکار روند.



شکل ۲-۵ ساختار تمام گره‌های ساختمان اسم

۴-۲- سایر قسم کلمه‌ها

همانطور که اشاره شد سایر قسم کلمه‌های این بخش از اجزای ساختمان اسم‌ها استفاده می‌کنند. تکرار تک تک آنها در این بخش ضرورتی ندارد زیرا در فصل کلمه‌های غیر فعل برای هر کدام ساختمان سازنده کلمه به طور کامل شرح داده شده است و مشخص است که هر کدام در پیاده سازی از کدام گره شکل ۵-۲ استفاده می‌کنند.

موردی که در این ساختهای باید توجه داشت گاهی محدودتر بودن ساخت در آنهاست. به این صورت که مثلاً حرف اضافه از ساختمان `ra2` استفاده می‌کند، اما تکواز `را` و عطف `را` که در این ساختمان است، نمی‌پذیرد. برای حل این مشکل از نشانه‌های فرامتنی استفاده شده است. از ابتدای ساختار **حروف اضافه نشانه فرامتنی** `@P.PREP.ON@` به آن اضافه شده و با افزودن نشانه `@D.PREP@` در مسیر `را` و عطف در گره‌های `ra2` و `ham2` این دو مسیر برای **حروف اضافه مسدود** شده است (شکل ۶-۲). چنین نشانه گذاری‌ای برای سایر ساختهای این بخش که محدودتر از گره به کار رفته در بخش اسم‌ها هستند انجام شده است. با مقایسه ساختمان هر قسمت با ساختمان اصلی‌ای که در اسم به کار رفته می‌توان تمایز آنها و نیاز به استفاده از چنین نشانه گذاری فرامتنی‌ای را مشاهده کرد.

```

LEXICON Root
@P.PREP.ON@:=@P.PREP.ON@ prep;
...

LEXICON ra2
@D.PREP@+:@D.PREP@* takid3;
@D.PREP@+@D.PREP@* takid3;
0:0 ham2;
@D.PREP@+@D.PREP@x takid3;
@D.PREP+@D.PREP@x takid3;

LEXICON prep
ب FINPOINT;
@P.VOWEL.HEH@ و اسه prepHub; !VOWEL_HEH
@P.VOWEL.ALEF@ ها+@P.VOWEL.ALEF@ ها shakhsii2;
@P.VOWEL.ALEF@ برا prepHub; !VOWEL_ALEF
@P.VOWEL.HEH@ به prepHubFRM;
@P.VOWEL.ALEF@ اند@ prepHubFRM;
@P.VOWEL.ALEF@ از@ prepHubFRM;

```

شکل ۶-۲ ایجاد محدودیت در مسیر بای ساختهایی که محدودیت بیشتری دارند

صفت مفعولی و مصدر (садه و پیشوندی) هم در این بخش ساخته می‌شود که با در نظر داشتن ساخت افعال صورت می‌گیرد.

۵-۲- قواعد نگارشی

برای ساختمان مصدر و صفت مفعولی که در بخش غیر فعلی ساخته می‌شود از قاعده شکل ۷-۲ استفاده می‌شود تا تکواز نفی به درستی به کلمه افزوده شود و در مواردی که نیاز به تکواج میانجی‌ی بود، این تکواج استفاده شود.

```
define NNiNehRefine ["¤"] -> [YEH "*"] ||
  NOON _ [[ALEFCNTX \[YEH | .#.]] | [VAV MIM DAL] |
    [(VAV) FE TE] | [VAV RE DAL] |
    [(VAV) FE TE AA DAL]; [
```

شكل ۷-۲ قاعده نگارشی افزودن تکواز نفی به مصدر و صفت‌های مفعولی

در مواردی که واکه‌های پایانی در هنگام اتصال به تکواز جمع جانداران (مثل پرندگان)، تکواز شخصی (مثل همسون) و تکواز ربطی سوم شخص مفرد (مثل پرنس) حذف و یا با تکواجی دیگر جایگزین می‌شود از قواعد شکل ۸-۲ استفاده می‌شود.

```
define elimHeh [ HEH "¤" ] -> "@";
define makeBoth "@" (->) [ HEH "x" ];
define joinBoth; "*" <- "@"
```

شكل ۸-۲ قواعد بازنویسی حذف یا تغییر واکه‌های در هنگام اتصال به تکوازی دیگر

در واقع تکواز منتهی به واکه‌های ۵ برای هر سه حالت اتصال به جمع جانداران، تکواز شخصی و تکواز ربطی به صورت اختیاری حذف می‌شود؛ به این معنی که یکبار حذف می‌شود و یک بار سر جایش باقی می‌ماند.

پیوست سه

پیاده‌سازی مبدل‌ها

۱-۳- مقدمه

مبدل‌ها با استفاده از قسمت دستورالعمل مقید یا زبان قاعده‌مند ساخته می‌شوند. مبنای همه مبدل‌ها همان قواعد تصریفی و واژگان است که در صورت بندی لکس پیاده کرده‌ایم. در این قسمت فایل‌های لکس حاوی واژگان و قواعد را بارگذاری می‌کنیم و قواعد نگارشی و احتمالاً تغییرات آوایی مربوط به محاوره را نیز در ساخت‌ها اعمال می‌کنیم. تمام مبدل‌ها در فایل farsi.foma تعریف شده است.

۲-۳- مبدل استاندارد

فایل‌های لکسی که حاوی قواعد تصریفی فعل و غیر فعل است، در این مبدل بارگذاری می‌شود. در شکل ۱-۳ نحوه بارگذاری این فایل‌ها دیده می‌شود.

```

read lexc lexc/harf.lexc
define LexHarf;

read lexc lexc/MTvbs.lexc
define MTverbs;
read lexc lexc/MTchain.lexc
define MTchain;

read lexc lexc/iverb.lexc
define iLexVerb;

read lexc lexc/inoun.lexc
define iLexNoun;

read lexc lexc/ichain.lexc
define iChain;
```

شکل ۱-۳ نحوه بارگذاری فایل‌های لکس

البته فایل‌های MTchain.lexc و MTvbs.lexc حاوی فعلهای چند قطعه‌ای ساده و پیشوندی (ماضی بعید، ماضی ابعد، ماضی التزامی و مستقبل) است که در مبدل استاندارد استفاده نشده است، اما امکان به کارگیری آنها وجود دارد. قواعد بارگذاری می‌شود و قواعد نگارشی هر قسمت اعمال می‌شود (شکل ۲-۳). فعل‌ها قاعده‌های خاص خود را دارند و غیر فعلی‌ها نیز قواعد خاص خود را. قواعد هر کدام به صورت جداگانه اعمال می‌شود.

```

define NN iLexNoun | LexHarf .o.
    NNI NehRefine .o.
    elimHeh .o.
    makeBoth .o.
    joinBoth;
define iVERBS iChain | iLexVerb .o.
    VBi NehRefine .o.
    VBi BehRefine .o.
    VBhamRefine .o.
    VBoftRefine .o.
    VBhehSpace;

```

شکل ۲-۳ اعمال قواعد بازنویسی فعل و غیر فعل

سپس همه کلمات به صورت یکپارچه درمی‌آیند (شکل ۳-۳). حالت غیر قطعی برای آ و ا نیز در این قسمت به کلمات اعمال می‌شود.

```

define ALEF "ا";
define AA "آ";
define aPhone [AA | ALEF] -> [AA | ALEF];
define WORDS NN | iVERBS .o. aPhone;

```

شکل ۳-۳ تولید همه کلمات

در نهایت با استفاده از قاعده FST مبدل نهایی استاندارد ساخته می‌شود (شکل ۴-۳). متغیر WORDS که در شکل ۳-۳ توضیح داده شد، حاوی تمام کلمات و قاعده‌های متناظر آنهاست. این متغیر سپس با قاعده brdHandler ترکیب می‌شود که مرزنماهای موجود در قواعد لکس را جایگزین نویسه‌های اصلیشان می‌کند. * به اتصال، ^ به فاصله کامل و × به نیم فاصله تبدیل می‌شود. همینطور در بعضی موارد که ممکن است چند نویسه نیم‌فاصله تولید شود، تکرار آن را حذف می‌کند. قاعده WORDS ترکیب شده و حالت غیر قطعی را برای سه نویسه مرزنماهی اتصال، فاصله و نیم‌فاصله ایجاد می‌کند. قاعده refineZWNJ نیم‌فاصله‌ای که بعد از حروف نچسب قرار می‌گیرد حذف می‌کند زیرا این حروف به حرف بعد از خود در هیچ صورتی نمی‌چسبند و نیازی به نیم‌فاصله نیست (قاعده NACHASB).

در نهایت قاعده debugFST را تولید می‌کند. این قاعده با افزودن علامت‌های <> به دو طرف قاعده سازنده کلمه و علامت‌گذاری مبدل با نامی مشخص که در آرگومان دوم قاعده مشخص می‌شود (در اینجا کلمه استاندارد)، خروجی‌های مبدل را علامت‌گذاری می‌کند تا مشخص باشد قاعده تولید شده خروجی کدام مبدل است. مثلاً کلمه کتاب‌ها قاعده <استاندارد:اسمعام=کتاب+تاکید> را تولید می‌کند که مشخص است خروجی مبدل استاندارد است.

```

define brdHandler putConcat .o.
    extraZWNJ .o.
    putZWNJ .o.
    iPutSpace;
define debugFST(FST, num)
    [FST.i .o. [+ @-> [< num ":" ] ... [>] ].i ;
define NACHASB
    [ "ء" | "۰" | "۱" | "۲" | "۳" | "۴" | "۵" | "۶" | "۷" | "۸" | "۹" ];
define BRD [SPACE | ZWNJ | 0];
define fborderNonDet BRD -> BRD;
define refineZWNJ ZWNJ -> CONCAT || NACHASB _ ?*;
define STANDARD [ALEF SIN TE ALEF NOON DAL ALEF RE DAL];

#####STANDARD-1-FST#####
define FST1 WORDS .o.
    brdHandler .o.
    fborderNonDet .o.
    refineZWNJ;
regex debugFST(FST1, STANDARD);
#####

```

شکل ۳-۴ قاعده سازنده مبدل استاندارد و قواعد کمکی دیگر

۳-۳- سایر مبدل‌های شناسایی

غیر از مبدل **تولید** که برای تولید کلمه برای قاعده دریافتی طراحی شده و مبدل **تقطیع** که برای جداسازی هر دو قطعه به هم چسبیده (چه رسمی و چه غیر رسمی) به کار می‌رود، سایر مبدل‌ها همگی برای کم کردن نویز کلمات غیر رسمی در نسبت با استانداردهایی است که در قواعد مبدل استاندارد تعریف شده است. کلماتی که خود واحد واژگانی نیستند و به دلیل تغییر نگارشی تغییر یافته‌اند، خطاهای نگرشی از نوع آوایی و استفاده از نویسه‌های هم‌صدا، همگی در این بخش به کار رفته است تا علاوه بر قواعد سازنده کلمات استاندارد بتواند سایر تغییرات را هم پوشش بدهد.

مبدل هم‌صدا

این مبدل دارای یک قاعده بیشتر از مبدل **استاندارد** است (شکل ۳-۵، قاعده homoPhones) و علاوه بر آن در قواعد لکس در بخش کلمات غیر فعلی، تکواز ۵ را نیز به عنوان کسره اضافه در ساختار غیر فعلی در نظر می‌گیرد (فایل inoun.lexc بخش گره ezafe خطوط ۴۱۹ و ۴۲۰).

قاعده homoPhones حاوی مجموعه‌ای از قاعده‌های تغییر نویسه‌ها با هم‌صداها‌یشان است که در شکل ۳-۵ همه آنها تعریف شده است.

قطعه نشان دهنده این مبدل که در همه قاعده‌های خروجی آن به کار می‌رود **هم‌صدا** است. خروجی این مبدل قاعده‌های رسمی و غیر رسمی است.

```

#####
#Homo Phones#####
define zPhone [ZAL | ZAD | ZA | ZE] ->
    [ZAL | ZAD | ZA | ZE];
define tPhone [TA | TE] -> [TA | TE];
define sPhone [SAD | SE | SIN] -> [SAD | SE | SIN];
define ghePhone [GHEIN | GHAF] -> [GHEIN | GHAF];
define hePhone [HEH | HE] -> [HEH | HE];
define ePhone HAMZE (->) [YEH | EIN];
define homoPhones zPhone .o.
    tPhone .o.
    sPhone .o.
    ghePhone .o.
    hePhone .o.
    ePhone;
#####
define HOMOPHONE [ HEH MIM SAD DAL ALEF ];
#####
#HOMOPHONE-1-FST#####
define FST2 WORDS .o.
    brdHandler .o.
    fborderNonDet .o.
    refineZWNJ .o.
    homoPhones;
regex debugFST(FST2, HOMOPHONE);
#####

```

شکل ۳-۵ قاعده سازنده مبدل همسداو قواعد کمکی

مبدل آوایی

بدنه اصلی این مبدل هم مبدل استاندارد است (شکل ۳-۶)، اما علاوه بر آن دو دسته قاعده اضافه و DELIB و finalEliminate دارد که تغییرات آوایی را تولید می‌کنند. قاعده اول برای جابجایی نوسه‌هایی است که گاه (به عمد) جایگزین یکدیگر می‌شوند. قاعده دوم تغییرات آوایی را با توجه به بافت نویسه‌ها اعمال می‌کند.

در مرحله بعد به دلیل تغییرات آوایی شدید کلمه رسمی به غیر رسمی تبدیل می‌شود اما قاعده سازنده آن همچنان حاوی قطعه +رسمی است که با استفاده از قاعده elimFormal حذف می‌شود. این قاعده با معکوس کردن مبدل FST3 برچسب و قاعده را دامنه تغییر قاعده قرار می‌دهد و قطعه نشان‌گر رسمی بودن قاعده را حذف می‌کند.

در نهایت مبدل آوایی با برچسب آوایی در همه قاعده‌های خروجی آن ساخته می‌شود. خروجی این مبدل فقط قاعده‌های غیر رسمی است.

```

define AVAEE [AA VAV ALEF YEH YEH];
define fin1 HEH -> 0 || ?* ALEFCNTX _ .#.;
define fin2 TE -> 0 || ?* [SIN | SHIN | FE] _ BOUNDS+;
define fin3 RE -> 0 || ?* [KAF | GHAF DAL | GAF ALEF] _ .#.;
define fin4 DAL -> 0 || NOON _ .#.;
define UN ALEF -> VAV || _ NOON;
define khaPhone [KHE VAV ALEF] -> [KHE ALEF]; #خاستن
define nabMab NOON -> MIM || _ BEH;
define chCHE HEH -> 0 || .#. CHE _;

define TTT [TE | DAL] -> [TE | DAL];
define JJJ [ZHE | JIM] -> [ZHE|JIM];
define AAA ALEF -> [EIN] || .#. _;
define ALAL AA -> [EIN ALEF] || .#. _;
define KARGAR [KAF | GAF] -> [KAF | GAF] || _ RE; #one-way
define SHECHE [SHIN | CHE] -> [SHIN | CHE]; #one-way

define DELIB [TTT | JJJ | AAA | ALAL | KARGAR | SHECHE];
define finalEliminate [fin1 | fin2 | fin3 | fin4 | UN |
    khaPhone | nabMab | chCHE];
define elitRules [finalEliminate | DELIB];
#####AVAEE-1-FST#####
define FST3 WORDS .o.
    brdHandler .o.
    elitRules .o.
    fborderNonDet .o.
    refineZWNJ;
define elimFormal ["+" RE SIN MIM YEH] -> 0;
define finalAvaee FST3.i .o. elimFormal;
regex debugFST(finalAvaee.i, AVAEE);
#####

```

شکل ۳-۶ قاعده سازنده مبدل آوایی و قواعد کمکی

مبدل بیانی

این مبدل یک قاعده بیشتر از مبدل استاندارد دارد (شکل ۳-۷). قاعده wordStress تکرار یک حرف بیش از دوبار برای تمام حروف محاسبه می‌کند (۳۳ حرف، h1 تا h33) و در صورت موجود بودن چنین ساختاری آن را از کلمه حذف کرده و سپس تحلیل تصربیفی آن را انجام می‌دهد.

گرچه این ساختار در گونه محاوره اتفاق می‌افتد اما ساختار کلماتی که رسمی است دست نمی‌خورد و ممکن است کلمه‌ای که ساخت رسمی دارد اینطور نگارش شود. نام مبدل که در قاعده می‌آید موضوع را روشن می‌کند و بنابراین نیازی به حذف قطعه **+رسمی** از قاعده سازنده کلمه نیستیم.

```

define h1 ALEF+ -> 0 || ALEF _ ;
define h2 AA+ -> 0 || AA _ ;
define h3 BEH+ -> 0 || BEH _ ;
define h4 PEH+ -> 0 || PEH _ ;
...
define h31 VAV+ -> 0 || VAV _ ;
define h32 HEH+ -> 0 || HEH _ ;
define h33 YEH+ -> 0 || YEH _ ;

define wordStress h1 .o. h2 .o.
    h3 .o. h4 .o.
    ...
    h31 .o. h32 .o. h33;
#####
#define WORD-STRESS#####
define FST4 WORDS .o.
    brdHandler .o.
    fborderNonDet.i .o.
    refineZWNJ .o.
    wordStress.i;
regex debugFST(FST4, EXPRESSIVE);
#####

```

شکل ۷-۳ قاعده سازنده مبدل آوایی و قواعد کمکی

مبدل تقطیع

این مبدل هم یک قاعده بیشتر از مبدل استاندارد دارد (`splitJointWord`). این قاعده قطعه را به دو قسمت می‌شکند. این شکستن در مرز هر دو نویسه کنار هم در آن قطعه انجام می‌شود. بنابراین به اندازه تعداد حروف یک قطعه، قطعه‌های جدید تولید می‌شود و سپس سعی می‌شود هر کدام از این قطعه‌ها تحلیل تصربیفی شود (شکل ۸-۳). خروجی این مبدل قاعده‌های رسمی و غیر رسمی است.

```

define splitJointWord1 ?* (->) 0 || ?+ _ .#.;
define splitJointWord2 ?* (->) 0 || .#. _ ?+;
define splitJointWord splitJointWord1 | splitJointWord2;
#####$PLITTER#####
define FST5 WORDS .o.
    brdHandler .o.
    fborderNonDet.i .o.
    refineZWNJ .o.
    splitJointWord.i;
    regex debugFST(FST5, $PLITTER);
#####

```

شکل ۸-۳ قاعده سازنده مبدل تقطیع و قواعد کمکی

۴-۳- مبدل تولید

مبدل تولید برای دریافت قاعده و تولید کلمه استفاده می‌شود. در این مبدل برای ساده‌تر شدن تولید، قواعد متناظر کلمات، ساده‌تر شده است. تمام قطعه‌های نماینده جمع‌های مختلف به قطعه **+جمع** تبدیل شده است و علامت جمع مکسر نیز حذف گردیده است (شکل ۹-۳، قاعده‌های PLsignElim و ARMokasarFlagElim).

بعضی قواعد برای ایجاد تمایز بین قاعده دو همنویسه، از اعراب استفاده کرده‌اند که در اینجا حذف می‌شود. حالت غیر قطعی آ و آ توسط قاعده rawWORDS از کلمات حذف می‌شود. حالت غیر قطعی آ و آ به قواعد افزوده می‌شود، بنابراین در یک قاعده بین آ و آ تفاوتی وجود ندارد.

در نهایت مبدل به صورت معکوس ساخته می‌شود (FST6).

ورودی این مبدل قاعده‌های رسمی و غیر رسمی و خروجی آن کلمات رسمی و غیر رسمی است.

```

define rawWORDS NN | iVERBS;
define ARMokasarFlagElim [ "(" JIM MIM ")" ] (->) 0;
define JaElim [ JIM ALEF ] (->) [JIM MIM EIN] || "+" _ ;
define JhaElim [ JIM HEH ALEF ] (->)
    [JIM MIM EIN] || "+" _ ;
define JanElim [ JIM ALEF NOON ] (->)
    [JIM MIM EIN] || "+" _ ;
define JatElim [ JIM ALEF TE ] (->)
    [JIM MIM EIN] || "+" _ ;
define JinElim [ JIM YEH NOON ] (->)
    [JIM MIM EIN] || "+" _ ;
define JunElim [ JIM VAV NOON ] (->)
    [JIM MIM EIN] || "+" _ ;
define JjatElim [ JIM JIM ALEF TE ] (->)
    [JIM MIM EIN] || "+" _ ;
define JganEilm [ JIM GAF ALEF NOON ] (->)
    [JIM MIM EIN] || "+" _ ;
define PLsignElim JaElim .o. JhaElim .o.
    JanElim .o. JatElim .o. JinElim .o.
    JunElim .o. JjatElim .o. ganEilm;
define EERAB [ " " | " " | " " ];
define diaElim EERAB (->) 0;
define FST6 [rawWORDS .o. brdHandler].i .o.
    ARMokasarFlagElim .o. PLsignElim .o.
    diaElim .o. aPhone;
regex FST6;

```

شکل ۹-۳ قاعده‌های سازنده مبدل تولید

Abstract:

Inflectional analysis is one of the most important parts in Natural Language Processing. Analyzing informal Persian is one of the least explored areas in Persian language processing. Furthermore, since informal Persian consists of informal and formal words, an inflectional analyzer for informal Persian should have the ability to handle formal words too. Contemporary Persian consists of both formal and informal register; the former being used in formal documents, books, magazines and etc., and the latter in conversations, social networks and messaging of Persians.

The rapid growth of the informal register of Persian in social media and, generally speaking, in the Web, as well as the use of speech in various technology tools, necessitates the processing of this register. On the other hand, the researches and tools developed for Persian text processing are mostly focused on formal Persian. Therefore, exploring this field and making a tool for inflectional analysis is essential for other levels of language processing and engineering.

With covering both formal and informal Persian words, this research gains the credit for the first inflectional analyzer of contemporary Persian words. The linguistic resources for the informal register of Persian like lexicon, morphological and orthographic besides the informal corpus which gathered for this research and the evaluation data set are provided.

The recall and precision for this analyzer on a data set of 1786 unique words are 95.56% and 99.65% respectively. These numbers have improved by half to one percent using alternative FSTs which cover the phonological variations of informal words.

Keywords:

Inflectional Analyzer, Inflectional Analysis, NLP, Computational Linguistics, Informal Persian, Contemporary Persian, Formal Persian.



**University of Tehran
Faculty of New Sciences and Technologies
Department of Interdisciplinary Technology**

Title:

An inflectional analyzer for contemporary Persian

By:
Davood Heidarpour

Supervisors:
Dr. Mostafa Salehi

Dr. Mahmoud Bijankhan

Advisor
Dr. Hadi Veisi

A Thesis Submitted to the Graduate Office in Fulfillment
of Requirements for the Degree of Master of Science
in Computational Linguistics

February 2018

