

بررسی پیکره-بنیاد هم‌نگاره‌های اسمی و صفتی فارسی جهت کمک به برجسب‌گذاری صحیح اجزای کلام

الهام علایی ابوذری *

دکتری زبان‌شناسی همگانی

استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

پذیرش: ۹۶/۱۲/۰۶

دریافت: ۹۶/۰۷/۲۲

چکیده:

در تهیه سامانه‌های برجسب‌گذاری اجزای کلام در زبان فارسی، بررسی ساخت-واژی اسم‌ها و صفت‌ها از دو نظر حائز اهمیت است: ۱- اگر در یک پیکره متنی فارسی، کلمه‌ای در واژگان حضور نداشته باشد (کلمه خارج از واژگان)، نمی‌توان برجسب‌های مربوط به کلمه را بازیابی کرد. در این صورت، برجسب کلمه را تنها می‌توان با توجه به شکل کلمه (انواع پیشوندها و پسوندهایی که به کلمات متصل می‌شوند) یا بافتی که کلمه در آن ظاهر می‌شود، یا هر دو، حدس زد. ۲- زبان فارسی ظرفیت بالایی برای ساخت هم‌نگاره‌های جدید که از ساخت‌واژه فارسی نشأت می‌گیرند، را نیز دارد. بنابراین، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، به منظور تفکیک آن‌ها از هم ضروری به نظر می‌رسد، زیرا اکثر صفت‌ها در بافت‌های گوناگون، با صورت نوشتاری یکسان، می‌توانند برجسب «اسم» بگیرند. در تحقیق حاضر ساخت‌واژه اسم‌ها و صفت‌ها در فارسی بررسی شده است. نظام نوشتاری زبان فارسی نیز مورد بررسی قرار گرفته است تا از این رهگذر بتوان به شناسایی انواع هم‌نگاره‌ها در زبان فارسی پرداخت. سپس، انواع هم-نگاره‌ها در زبان فارسی مورد مطالعه قرار گرفته است و در نهایت از طریق جستجو به دو روش ماشینی و دستی، فهرست مبسوطی از هم‌نگاره‌ها از پیکره‌های «پیکره متنی زبان فارسی»، «پایگاه دادگان زبان فارسی» و «پیکره وابستگی نحوی زبان فارسی» تهیه شده

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۲۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در SCOPUS و LISA و ISC
<http://jlist.irandoc.ac.ir>
دوره XX | شماره X | صص XX-XX
۱۳XX X

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:

دورن متن:

(علایی ابوذری، زودآیند)

در فهرست منابع:

علایی ابوذری، الهام. زودآیند. بررسی پیکره-بنیاد هم‌نگاره‌های اسمی و صفتی فارسی جهت کمک به برجسب‌گذاری صحیح اجزای کلام. با پژوهشنامه پردازش و مدیریت اطلاعات. <http://Jipm.irandoc.ac.ir> (دسترسی در روزنامه/سال)

است. بررسی کلی هم‌نگاره‌ها در پیکره‌های مورد مطالعه نشان می‌دهد که بیشتر هم-نگاره‌ها، فراوانی بالایی در پیکره‌های متنی فارسی دارند و اکثر آن‌ها در اثر یکسان بودن نمود نوشتاری تکواژ یاء نکره، یاء اسم‌ساز، شناسهٔ دوم شخص مفرد، یاء صفت‌ساز و یاء متصل به گروه اسمی، ایجاد شده‌اند.

کلیدواژه‌ها: برچسب‌گذاری اجزای کلام، ساخت‌واژه‌اسم‌ها و صفت‌ها، نظام نوشتاری، هم‌نگاره‌ها

*الهام علایی ابوذر Alayi@irandoc.ac.ir

۱. مقدمه

مفهوم برچسب‌گذاری اجزای کلام^۱، مفهومی برگرفته از زبان‌شناسی است که به طبقه‌نحوی کلمات مربوط می‌شود؛ به این صورت که ویژگی‌های ترکیبی یک کلمه در بافت/ساخت نحوی مشخص می‌شود. در زبان‌شناسی پیکره‌ای، برچسب‌گذاری اجزای کلام، در واقع عمل انتساب برچسب به کلمات تشکیل‌دهندهٔ یک متن یا یک پیکره است. این برچسب‌گذاری براساس مقولهٔ آن کلمه در متن (مانند «اسم»، «فعل»، «قید»، «صفت» و غیره) صورت می‌پذیرد. کلمات در جایگاه‌های مختلف می‌توانند برچسب‌های واژگانی متفاوت داشته باشند. به عنوان مثال، کلمهٔ «آسمانی» در جملهٔ «فردا آسمانی صاف در انتظار شهروندان عزیز خواهد بود»، دارای برچسب «اسم، نکره»، در جملهٔ «او انسانی آسمانی بود»، دارای برچسب «صفت» و کلمهٔ «مردی» در جملهٔ «در راه مدرسه، مردی را دیدم»، دارای برچسب «اسم، نکره»، در جملهٔ «مثلاً تو مردی؟»، دارای برچسب «اسم + فعل اسنادی»، است. بنابراین، درصد بالایی از کلمات از نقطه‌نظر برچسب‌واژگانی، مبهم هستند و می‌توانند فارغ از بافت، بیش از یک برچسب‌واژگانی داشته باشند.

برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی^۲ از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است (Megerdooonian: ۲۰۰۴). همچنین در طراحی

^۱ Part Of Speech (POS) tagging

^۲ natural language processing

سیستم نمایه‌ساز ماشینی، یکی از بخش‌ها، طراحی زیرسیستم‌ها است که خود بخش‌های گوناگونی دارد، یکی از این بخش‌ها، زیرسیستم تحلیل واژگانی است. این زیرسیستم، متن را به واژه‌ها تفکیک می‌کند و ماهیت هر کلمه را تشخیص می‌دهد و تشخیص نوع واژه و شناسایی فعل‌ها، الفاظ و اصطلاح‌ها را در بر دارد. بنابراین سیستم برچسب‌دهی خودکار، ماهیت مقوله کلمات را مشخص می‌کند تا بتوان در مراحل بعدی از این اطلاعات در جهت استخراج کلیدواژه‌ها یا هر نوع بازایی اطلاعات از متن، استفاده کرد. تاکنون از مدل‌ها و روش‌های بسیاری برای برچسب‌گذاری در زبان‌های مختلف استفاده شده است. این روش‌ها و مدل‌ها را می‌توان به دو دسته کلی تقسیم‌بندی کرد: دسته اول، رویکردهای آماری است که از به کار بردن روش‌های احتمالاتی برای پیکره‌های برچسب‌خورده^۱ بهره می‌جویند و دسته دوم، رویکردهای غیرآماري و مبتنی بر دانش^۲ / مبتنی بر قاعده^۳ است که بر مبنای یادگیری ماشینی و دانش بشری استوار هستند، البته رویکرد سومی هم وجود دارد که در واقع تلفیقی از دو رویکرد آماری و مبتنی بر قانون است که برچسب‌گذاری تلفیقی یا دووجهی^۴ خوانده می‌شود (Megerdooimian: ۲۰۰۴).

در زبان‌هایی که ساخت‌واژه تصریفی و اشتقاقی آن‌ها پیچیده است، ظرفیت بالایی برای ساخت کلمات با اشکال جدید وجود دارد، زیرا در این زبان‌ها، تکواژها به روش‌های گوناگون به یکدیگر متصل می‌شوند و کلمات جدیدی را تولید می‌کنند که امکان دارد این کلمات حتی در پیکره‌های بزرگ وجود نداشته باشند و یا فراوانی آن‌ها بسیار اندک باشد. تکواژهای تصریفی معمولاً مفاهیم دستوری را به کلمه اضافه می‌کنند و بنابراین مقوله نحوی را تغییر نمی‌دهند، اما تکواژهای اشتقاقی کلمات جدید می‌سازند و دارای مفهومی خاص هستند یا مقوله نحوی کلمه را تغییر می‌دهند. بررسی ساخت‌واژی کلمات از دو نظر حائز اهمیت است: ۱- شناسایی برچسب‌کلماتی که در واژگان وجود ندارند و تنها می‌توان با توجه به شکل کلمه یا بافتی که کلمه در آن ظاهر می‌شود، یا هر دو، برچسب نحوی چنین کلماتی را حدس زد. ۲- کمک به رفع ابهام از برچسب نحوی هم‌نگاره‌ها^۵ در فارسی؛ بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابه

¹ tagged corpora

² knowledge-based

³ rule-based

⁴ hybrid tagging

⁵ homographs

دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به عنوان مثال، نمود نوشتاری وند
تصریفی نکره ساز / i / و وند اشتقاقی صفت‌ساز / i / یکسان است («-ی») و کلماتی مانند:
«آسمانی» و «اسلامی» ، فارغ از بافت، می‌توانند هم برچسب «اسم» داشته باشند و هم «صفت».
گویشوران فارسی به دلیل آشنایی کامل با قواعد ساخت‌واژی زبان خود، هنگام مواجهه با چنین
کلماتی که حاوی تکواژهای تصریفی و اشتقاقی هستند یا برای اولین بار با آن‌ها در متون گوناگون
فارسی مواجه می‌شوند، در فهم و پردازش آن‌ها معمولاً دچار مشکل نمی‌شوند. اما این مسئله راجع
به سامانه‌های پردازش رایانه‌ای متن، صحیح نیست. این سامانه‌ها، به خصوص سامانه‌های برچسب-
گذاری، به دلیل عدم اشراف به قواعد ساخت‌واژی زبان، در برخورد با کلمات دارای پیچیدگی-
های ساخت‌واژی، توان محدودی دارند. همچنین ویژگی‌های خاص دستوری و نگارشی زبان و
خط فارسی، دشواری‌هایی را در ذخیره و بازیابی اطلاعات در محیط رایانه‌ای نیز پدید آورده
است. رسم‌الخط فارسی نیز از یک سو به علت اختلاف نظر پدیدآورندگان متون و از سوی دیگر
پیچیدگی‌های ذاتی خود، به هنگام ذخیره، جستجو و بازیابی اطلاعات، چالش‌های متعددی را
برای طراحان و نمایه‌سازان پایگاه‌ها، کاربران و پدیدآورندگان منابع به وجود آورده است
(آخشیک و فتاحی: ۱۳۹۱)؛ این پیچیدگی‌ها مشکلاتی بسیاری را در مسیر برچسب‌گذاری رایانه-
ای اجزای واژگانی کلام ایجاد می‌کند؛ برخی از این مشکلات شامل موارد زیر است:

الف. پیچیدگی ساخت‌واژه زبان فارسی و ارتباط آن با خط فارسی: اگر چند وند در یک کلمه
ظاهر شوند، همه آن‌ها معمولاً به کلمه می‌چسبند (۲۰۰۰: Megerdooimian)؛ نشانه‌های جمع، کسره
اضافه، نشانه نکره، ضمائر ملکی و..... می‌توانند به کلمه متصل شوند. مانند: «کتابهایم» که شامل
«کتاب+ها (علامت جمع)+ + ی» < نمود نوشتاری همخوان میانجی /z/ + م (ضمیر ملکی اول
شخص مفرد)» است. این ویژگی باعث اشکال متفاوتی از کلمات با ریشه یکسان می‌گردد و این
کلمات در سامانه‌های محاسباتی، متفاوت از یکدیگر فرض می‌شوند. در مورد افعال می‌توان
گفت، از نقطه نظر ساخت‌واژی، فعل شامل بن فعل و وندهای تصریفی است؛ افعال با توجه به
شخص و شمار صرف می‌شوند و بنابراین اشکال متفاوتی از آن‌ها به وجود می‌آید. شکل یکسان
برخی از تکواژها نیز باعث ابهام در متون فارسی می‌شود. به عنوان مثال، پسوند «-ی» در کلمه
«مردی»، فارغ از بافت، می‌تواند هم نشانه نکره در نظر گرفته شود و هم شناسه دوم شخص مفرد
در یک فعل اسنادی («مرد»+ «ی» (به معنی مرد هستی)). علاوه بر این، در فارسی معمولاً واکه-

های کوتاه نمود نوشتاری ندارند و این مسئله نیز باعث ابهام در تحلیل می‌گردد. مانند: کلمه «مردم» که می‌تواند به صورت‌های /mardam/ و /mordam/ و /mardom/ تلفظ شود. این ابهام، ابهام هم-نگاره خوانده می‌شود.

پ. عبارت‌های چندکلمه‌ای: حضور افعال مرکب و کلمات مرکب اسمی مشکلی دیگر در پردازش زبانی محسوب می‌شود. به عنوان مثال، در کلمات مرکب اسمی که هسته‌آغاز هستند، یعنی هسته گروه اسمی در ابتدای عبارت ظاهر شده است، وند جمع به قسمت اول متصل می‌شود. مانند: ماشین لباس شویی + وند جمع = ماشین‌های لباس شویی.

ت. مرز کلمات: فاصله نیز یکی از عوامل ابهام در متون فارسی است. در متون گوناگون، فاصله ماهیت اختیاری دارد. این مسئله در برخی موارد باعث می‌شود کلمات مجزا یک قطعه^۱ در نظر گرفته شوند (مانند: «رفتند مردم»). همچنین وندهای تصریفی (مانند: «-تر»، «-ترین»، «-ها»، «-می») را می‌توان به سه صورت به ستاک مربوط کرد: ۱) به ستاک چسبانند (۲) به صورت مجزا (غیرچسبان) با یک فاصله از ستاک در آخر کلمه / قطعه آورد (۳) به صورت غیرچسبان ولی با استفاده از نویسه نیم‌فاصله به ستاک مربوط کرد. به عنوان مثال، تکواژ جمع «-ها» در اسم‌ها می‌تواند به چند شکل ظاهر شود؛ در مورد کلمه «کتاب»، سه شکل «کتابها»، «کتاب‌ها» و «کتاب‌ها» برای حالت جمع در خط فارسی وجود دارد. در مورد افعال نیز می‌توان به پیشنهاد تصریفی زمان استمراری (یعنی «می-») اشاره کرد. مانند، حال استمراری اول شخص مفرد بن مضارع «رو» که هنگام اضافه شدن وند تصریفی «-می» می‌تواند به سه شکل «میروم»، «می‌روم» و «می‌روم» نوشته شود. این مورد نیز عامل تأثیرگذار دیگری در برجسب‌گذاری اجزا واژگانی کلام در زبان فارسی محسوب می‌شود، زیرا چند شکل نوشتاری متفاوت از یک کلمه عملاً به عنوان کلمات متفاوت تفسیر می‌شوند؛ هر سامانه برجسب‌گذاری باید توانایی تشخیص این شکل‌ها را داشته باشد تا

¹ token

متن، رشته‌ای از قطعات نوشتاری است و قطعه، نمود نوشتاری یک تکواژ یا دنباله‌ای از تکواژهاست که معمولاً از راست یا چپ، یا هر دو، از طریق فاصله یا علائم نقطه‌گذاری از دیگر قطعات تفکیک می‌شود. به عنوان مثال، کلمه‌ای مانند «می‌روم» یک قطعه نوشتاری است که از سه واحد زبانی تشکیل شده است: «می»، «رو» و «م». دو واحد زبانی اول با نیم‌فاصله از هم متمایز شده‌اند و چون حرف واو («و») به حرف بعد از خود نمی‌چسبد، استفاده از نیم‌فاصله بین واحد دوم و سوم ضرورتی ندارد (علایی و بی‌جن‌خان: ۱۳۹۲).

برچسب دستوری صحیح به کلمه بدهد (Megerdooimian: ۲۰۰۴). لازم به ذکر است فرهنگستان زبان و ادب فارسی در باب پیوسته‌نویسی یا جدانویسی ترکیبات در زبان فارسی، سه فرض را متصور است: ۱- تدوین قواعدی برای جدانویسی همه کلمات مرکب و تعیین موارد استثنا، ۲- تدوین قواعدی برای پیوسته‌نویسی همه کلمات مرکب و تعیین موارد استثنا و ۳- تدوین قواعدی برای جدانویسی الزامی بعضی از کلمات مرکب و پیوسته‌نویسی بعضی دیگر و دادن اختیار در خصوص سایر کلمات به نویسندگان؛ فرهنگستان در تدوین و تصویب دستور خط فارسی، فرض سوم را برگزیده و تنها موارد الزامی جدانویسی و پیوسته‌نویسی را مشخص کرده است (دستور خط فارسی: ۱۳۸۸، نقل از آخشیک و فتاحی: ۱۳۹۱).

ث. قطعه‌های پیچیده^۱: منظور از قطعه‌های پیچیده، شکل‌های چندجزئی هستند که در واقع از چسبیدن مقولات دستوری (مانند: « به »، « را »، « که » و « آن ») به ستاک به وجود می‌آیند که این تکواژهای وابسته/ مقولات دستوری چسبیده به ستاک، تکواژ شبهه کلمه^۲ خوانده می‌شوند. مانند: «بدین صورت»، «بعقیده»، «اینکار».

ج. قواعد آوایی و واجی: در فارسی، شکل وندهای متصل به ستاک براساس کاراکتر پایانی ستاک تغییر می‌کند. به عنوان مثال، اگر اسم جاندار به همخوان ختم شود، تکواژ جمع «-ان/ -an/» به ستاک متصل می‌شود. مانند: «زن + ان = زنان»، اگر به واکه ختم شود، همخوان غلت /z/ با نمود نوشتاری < ی > بین ستاک و پسوند قرار می‌گیرد. مانند: گدا+ان = گدایان.

چ. مرز عبارت‌ها: مسائل گوناگونی باعث ابهام در ساختار عبارت‌های اسمی در فارسی می‌شود. تکواژهای آشکار بسیار اندکی در فارسی برای نشان دادن مرزهای عبارات اسمی وجود دارد. اغلب هیچ جزئی در خط فارسی برای متصل کردن اجزا یک عبارت اسمی وجود ندارد، زیرا تکواژ اضافه^۳، اغلب واکه کوتاهی است که در خط ظاهر نمی‌شود. علاوه بر این، از آنجایی که ترتیب پایه و غالب اجزا جمله در فارسی، فاعل مفعول فعل^۴ است، فقدان تکواژهای آشکار نشان-

¹ complex tokens

² word-like morpheme

³ Ezafe morpheme

⁴ Subject-Object-Verb (SOV)

دهنده مرزها، تشخیص محل به پایان رسیدن فاعل و شروع مفعول را سخت می‌کند و دشواری-
هایی را در پردازش متن پدید می‌آورد (۲۰۰۴: Megerdooian).

از میان مسائل مطرح شده فوق، مسائلی که از ساخت‌واژه فارسی ناشی می‌شوند از اهمیت بیشتری برخوردار هستند، زیرا ساخت‌واژه علاوه بر شکل کلمه، برچسب کلمه را نیز تحت تأثیر قرار می‌دهد. در زبان فارسی، تکواژهای تصریفی و اشتقاقی بسیاری وجود دارد که با اتصال به کلمات باعث تغییر برچسب کلمات در پیکره‌های زبانی و همچنین کاهش فراوانی کلمات و در نهایت کاهش کارایی سامانه‌های برچسب‌گذاری آماری در زبان فارسی می‌شود. بنابراین، سامانه‌های برچسب‌گذاری باید حاوی تحلیل‌گر ساخت‌واژی نیز باشند (محسنی: ۱۳۸۷).

در تهیه سامانه‌های برچسب‌گذاری در زبان فارسی (نمونه فارسی: سامانه «هضم» *sobhe/hazm*)، بررسی ساخت‌واژی اسم‌ها و صفت‌ها از دو نظر حائز اهمیت است: ۱- اگر کلمه‌ای قبلاً در پیکره متنی ظاهر نشده باشد، نمی‌توان از پیکره اطلاعات دقیقی راجع به آن کلمه به دست آورد و حتی از توزیع کلمات در پیکره نیز نمی‌توان استفاده کرد، زیرا توزیع کلمات ناشناخته کاملاً متفاوت از کلمات شناخته شده در متن است. بنابراین، اگر در یک پیکره متنی فارسی، کلمه‌ای در واژگان حضور نداشته باشد (کلمه خارج از واژگان^۱) نمی‌توان برچسب‌هایی که کلمه به آن‌ها منتسب می‌شود را بازیابی کرد. در این صورت، برچسب کلمه را تنها می‌توان با توجه به شکل کلمه یا بافتی که کلمه در آن ظاهر می‌شود، یا هر دو، حدس زد. منظور از شکل کلمه، ویژگی‌هایی مانند انواع پیشوندها و پسوندهایی که به کلمات متصل می‌شوند، است. طبق گزارشات، روش‌هایی که مبتنی بر استفاده از چنین اطلاعاتی هستند، حاوی دقت بالاتری در زبان انگلیسی هستند (بریل ۱۹۹۴b و ویشدل ۱۹۹۳، نقل از محسنی: ۱۳۸۷). به عنوان مثال، یکی از این روش‌ها، روش برچسب‌گذار زیراکس^۲ است که در آن از قوانینی براساس بخش‌های پایانی کلمات، یعنی حروف انتهایی کلمات، برای انتساب مجموعه‌ای از برچسب‌ها به کلمات خارج از واژگان، استفاده می‌کند (کاتینگ ۱۹۹۲، نقل از محسنی: ۱۳۸۷). بنابراین، در حله اول، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، فارغ از بافت، ضروری به نظر می‌رسد تا از این رهگذر بتوان اسم‌ها و صفت‌ها را در

¹ Out Of Vocabulary word (OOV word)

² Xerox

صورت دارا بودن شاخص‌های صرفی و اشتقاقی، شناسایی و سپس برچسب‌دهی کرد. ۲- زبان فارسی ظرفیت بالایی برای ساخت هم‌نگاره‌های جدید که از ساخت واژه فارسی نشأت می‌گیرند، را نیز دارد. بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابه دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به عنوان مثال، نمود نوشتاری وند تصریفی نکره‌ساز / i / و وند اشتقاقی صفت‌ساز / i / یکسان است («-ی») و کلماتی مانند: «آسمانی» و «اسلامی»، فارغ از بافت، می‌توانند هم برچسب «اسم» داشته باشند و هم «صفت». بنابراین، بررسی ساخت واژگی اسم‌ها و صفت‌ها، به منظور تفکیک آن‌ها از هم ضروری به نظر می‌رسد، زیرا اکثر صفت‌ها در بافت‌های گوناگون، با صورت نوشتاری یکسان، می‌توانند برچسب «اسم» بگیرند.

در پژوهش حاضر ابتدا به پیشینه پژوهش و بررسی مطالعات انجام شده در زمینه ساخت واژه فارسی و سامانه‌های برچسب‌دهی پرداخته می‌شود، سپس، خلاصه‌ای از بررسی ساخت واژه اسم‌ها و صفت‌ها در فارسی ارائه خواهد شد و به دنبال آن، بخش بعد، مطالعه هم‌نگاره‌ها در فارسی است و در نهایت روش جمع آوری داده‌ها و معرفی پیکره‌ها و پایگاه‌های دادگان و نتیجه‌گیری ارائه خواهد شد.

۲- پیشینه پژوهش

از آنجائی که در این تحقیق، هدف از بررسی ساخت واژگی اسم‌ها و به دنبال آن، صفت‌ها، کمک به شناسایی و برچسب‌گذاری هم‌نگاره‌های اسمی و صفتی در پیکره‌های علمی است، بررسی مطالعات و تحقیقاتی که در زمینه برچسب‌گذاری اجزای واژگانی کلام صورت گرفته است، ضروری به نظر می‌رسد. بنابراین، در این بخش علاوه بر ذکر چند نمونه از تحقیقاتی که در زمینه ساخت واژه اسم‌ها و بعضاً، صفت‌ها، صورت گرفته است، به ذکر نمونه‌هایی از تحقیقاتی که در زمینه برچسب‌گذاری اجزای واژگانی کلام انجام شده است نیز پرداخته خواهد شد.

۱-۲- پیشینه مطالعاتی که در زمینه برچسب‌گذاری اجزای واژگانی کلام انجام شده است

Hearst (۱۹۹۱) روشی دقیق و ساده برای ابهام‌زدایی از هم‌نگاره‌های اسمی معرفی می‌کند که از پیکره‌متنی بزرگی در این راستا بهره می‌گیرد. الگوریتم معرفی شده، بافت اطراف اسم هدف را چک می‌کند و معنایی را انتخاب می‌کند که بیشترین شاهد برای انتخاب آن موجود باشد. این شواهد شامل مجموعه‌ای از ویژگی‌های واژگانی، نحوی و نوشتاری است. Merialdo (۱۹۹۴) مدل مارکوفی پنهانی را معرفی می‌کند که دو کلمه قبل به عنوان history در نظر گرفته می‌شود. در این مدل دادن هر برچسبی بستگی به برچسب دو کلمه قبل دارد که به صورت زیر نمایش داده می‌شود:

$$p(t_i | w_{i-1}, t_{i-1}, w_{i-2}, t_{i-2}, \dots, w_1, t_1) = p(t_i | t_{i-1}, t_{i-2})$$

Assi & Abdolhosseini (۲۰۰۰) سامانه برچسب‌دهی نحوی را به عنوان پروژه‌ای در پژوهشگاه علوم انسانی و مطالعات فرهنگی، معرفی می‌کند که سامانه مذکور، خود، در خدمت تهیه پیکره‌ای فارسی به نام «پایگاه دادگان زبان فارسی (FLDB)» قرار خواهد گرفت و می‌توان گفت اولین تلاش برای تهیه سامانه‌ای جهت برچسب‌دهی نحوی زبان فارسی در نظر گرفته می‌شود. عاصی و حاجی عبدالحسینی به بررسی روشی می‌پردازند که Schutze (۱۹۹۵) برای برچسب‌دهی نحوی در متون انگلیسی به کار برده است و هدف آن‌ها بررسی اعمال همان روش در فارسی است. آن‌ها معتقدند که با بهره‌گیری از محاسبات دقیق‌تر، می‌توان همان روش شوتز (۱۹۹۵) را در فارسی استفاده کرد. Hu et al (۲۰۰۵) به بهبود لینگویستیکا از طریق اضافه کردن اطلاعات نحوی در کنار اطلاعات ساخت‌واژی مربوط به هر کلمه در داخل شبکه مربوطه، کمک می‌کنند. آن‌ها علاوه بر درج فهرست ستاک‌ها و وندها، فهرست دیگری که مربوط به برچسب‌های نحوی کلمات است در شبکه وارد کرده‌اند. Elhaj (۲۰۰۹) نحوه برچسب‌دهی کلمات موجود در قرآن کریم را معرفی می‌کند که خود بخشی از سامانه‌های تجزیه و تحلیل رایانه‌ای قرآن کریم می‌باشد. در واقع سامانه‌ای که وی معرفی می‌کند تلفیقی است از تجزیه و تحلیلی ساخت‌واژی و مدل مارکوفی پنهان^۱ که براساس ساختار جملات عربی است. Frodl (۲۰۱۳) برچسب‌دهی با استفاده از شبکه‌های عصبی را معرفی می‌کند. وی برای این منظور از دو روش استفاده می‌کند: ۱- استفاده از شبکه‌های پرسپترون

^۱ Hidden Markov Model (HMM)

چندلایه تکرارشونده که محتمل‌ترین برچسب نحوی برای کلمه را در متن به طور احتمالاتی در نظر می‌گیرد. ۲- در روش دوم کلمات به بردارهایی از مشخصه‌ها در فضای چندبعدی تبدیل می‌شوند. وی معتقد است روش معرفی شده ۹۴٪ برچسب‌دهی نحوی درست از متن ارائه می‌دهد. راسخ و همکاران (۲۰۱۴) سعی دارند ضمن بررسی الگوریتم‌های موجود در رفع ابهام معنایی کلمه، روش‌هایی بر اساس بررسی مفهومی و ساختاری کلمات در جهت رفع ابهام کلمات در متون ارائه گردد و به این ترتیب راهکاری در جهت رفع مشکلات ترجمه ماشینی و بهبود کیفیت آن ارائه شود. محسنی (۱۳۸۸) با در نظر گرفتن مسائل و مشکلاتی که در مسیر برچسب‌گذاری اجزا کلام در فارسی وجود دارد، ابتدا طرحی کلی برای برچسب‌گذاری خودکار با دقت بالا در زبان فارسی پیشنهاد می‌کند. سپس تحلیل ساخت‌واژی و استفاده از آن را برای پوشش دادن تعداد زیادی از برچسب‌های پیکره با حفظ دقت بالا در برچسب‌گذاری کلمات مورد بررسی دقیق‌تر قرار داده و تأثیر وجود یک تحلیل‌گر ساخت‌واژی در سطح تصریف را بر برچسب‌گذاری اجزا واژگانی کلام در زبان فارسی بررسی می‌کند. وی معتقد است نتایج به دست آمده نشان از کارآیی بسیار مناسب این روش پیشنهادی در برچسب‌گذاری دارد.

۲-۲- نمونه‌هایی از مطالعاتی که در زمینه ساخت‌واژه اسم‌ها و بعضاً، صفت‌ها در فارسی صورت گرفته است

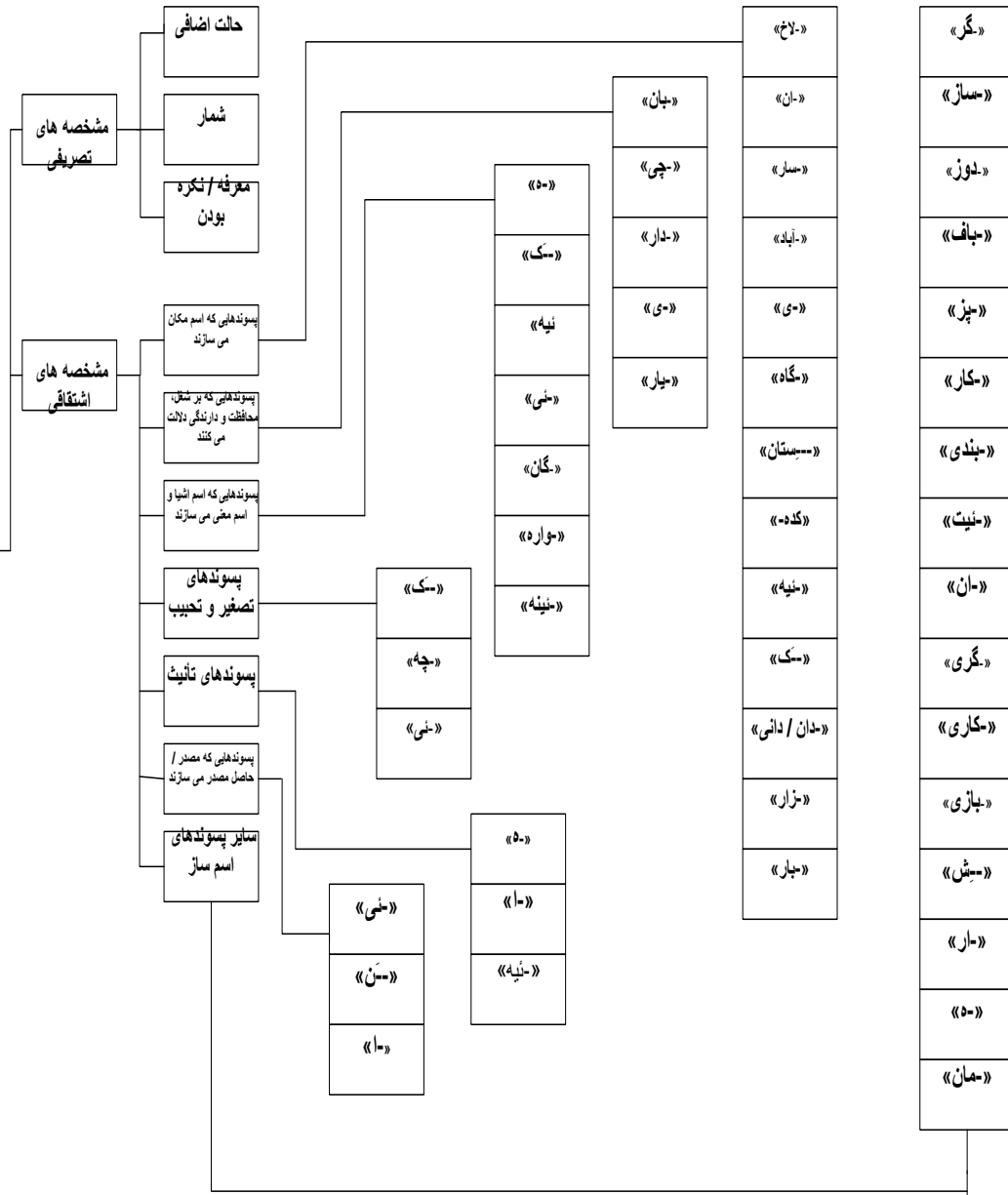
صادقی (۱۳۷۰-۱۳۷۲) طی ۱۲ مقاله، به شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر می‌پردازد. وی این گفتارها را با بحث از اشتقاق در زبان فارسی معاصر آغاز می‌کند و در صدر آن‌ها از امکانات واژه‌سازی از افعال مرکب که در سال‌های اخیر به عنوان عامل عقیم و غیرزایا بودن زبان فارسی مورد حمله بعضی از محققان قرار گرفته است، گفتگو می‌کند. صادقی (۱۳۷۰-۱۳۷۲)

همچنین طبقه‌بندی از انواع پسوندهای اسم‌ساز و صفت‌ساز ارائه می‌دهد. کشانی (۱۳۷۱) ابتدا به تعریف پسوند و نقش آن در زبان فارسی می‌پردازد، وی سپس پسوند و پسوندواره‌ها و انواع آن‌ها را در زبان فارسی به تفصیل شرح می‌دهد (نوعاً پسوندهای اشتقاقی که در ساختن انواع اسم، انواع صفت و قید به کار می‌روند به دقت بررسی می‌شوند) و در نهایت حرکت پسوندها را مورد مطالعه قرار می‌دهد. عاصی (۱۳۸۳) موضوعات اصلی در پردازش زبان طبیعی را معرفی می‌کند که شامل: حوزه خط (که خود شامل غلط‌یابی املائی و بازشناسی خودکار متن (OCR) است) و حوزه زبان (که خود مستلزم بهره‌گیری از حوزه‌های آواشناسی، ساخت‌واژه، نحو و معنی‌شناسی است). Lazard (۱۳۸۹) با رویکردی زبانشناختی و نگاهی موشکافانه، نکته‌سنج و بی‌طرفانه و با پرداختن به مسائلی که کمتر به آن‌ها توجه شده است، خصوصیات آوائی، صرفی و نحوی و نیز چگونگی شکل‌گیری واژه‌ها را در زبان فارسی معاصر تبیین می‌کند و گاه نیز به کشف ویژگی‌هایی در عناصر سازندهٔ زبان و روابط میان آن‌ها دست می‌یابد که طرح آن‌ها در نوشته‌های دستورنویسان فارسی‌زبان پیش از او سابقه نداشته است. نمونه‌هایی که از زبان فارسی معاصر در این دستور زبان ارائه شده، بر اساس آثار نویسندگانی چون صادق هدایت، صادق چوبک، علی دشتی، جمالزاده و دیگران و نیز نوشته‌های روزنامه‌ها و مجلات و سرانجام شنیده‌های مؤلف هنگام اقامتش در ایران بوده است.

بررسی پژوهش‌های انجام شده در زمینهٔ برچسب‌گذاری اجزای کلام نشان می‌دهد، مسئلهٔ برچسب‌دهی در پردازش متون حائز اهمیت است. برخی از پژوهش‌های اشاره شده در بخش ۲-۱، استفاده از ابزارهایی مانند مدل مارکوفی پنهان، تلفیقی از تجزیه و تحلیلی ساخت‌واژی و مدل مارکوفی پنهان و شبکه‌های پرسپترون چندلایه تکرار شونده را پیشنهاد می‌کنند و برخی دیگر روش‌هایی را معرفی می‌کنند که مستلزم بهره گرفتن از ابزارهای تحلیل‌کننده ساخت‌واژی و نحوی است. در این راستا و به منظور ارائهٔ بهینهٔ چنین سیستم‌های برچسب‌دهی به اجزای کلام، ناگزیریم به بررسی ساخت‌واژی اجزای کلام در فارسی بپردازیم. بنابراین، در بخش ۲-۲، نمونه‌هایی از مطالعات انجام شده در این زمینه ارائه شده است.

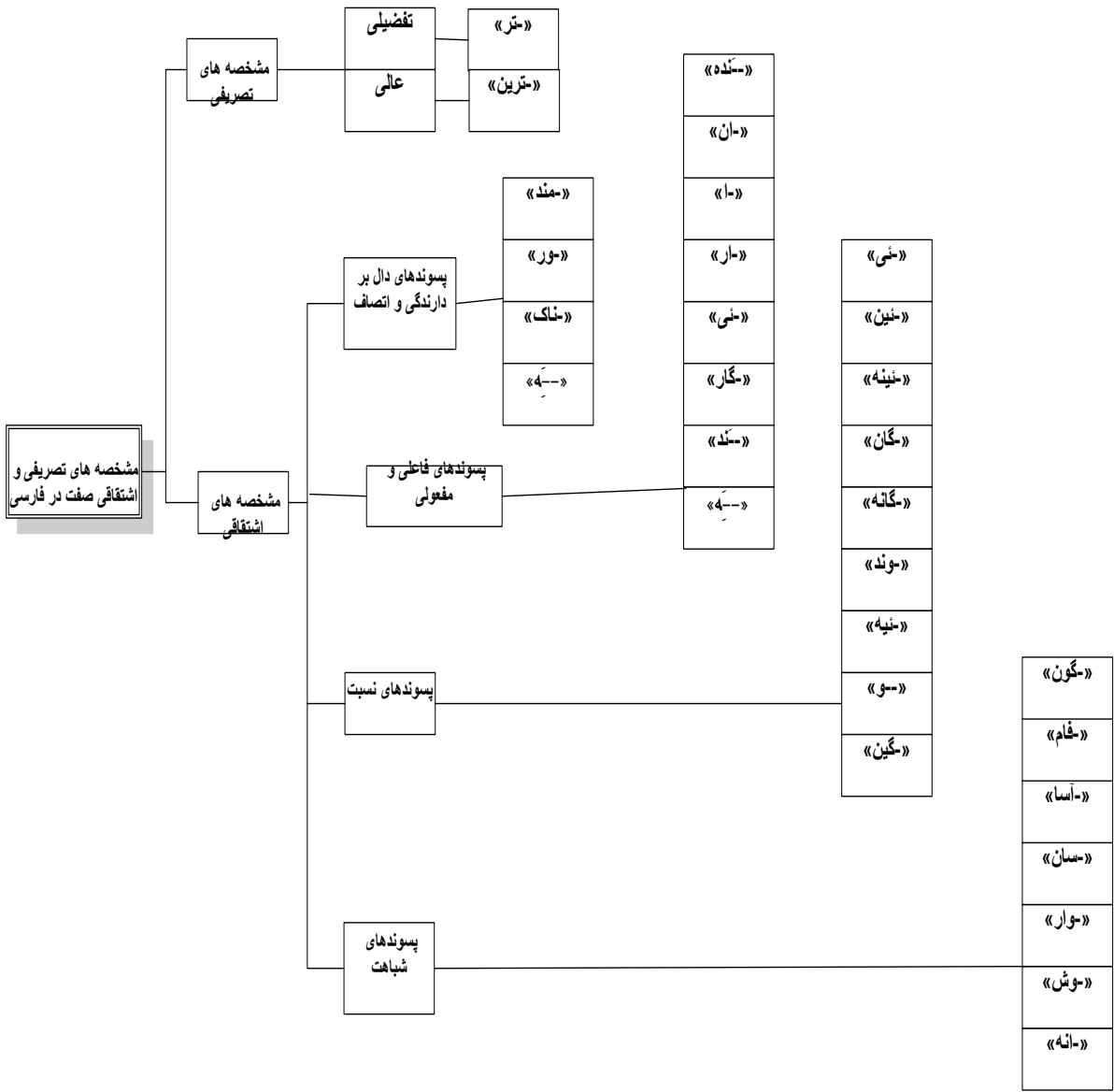
زبان فارسی دارای شمار نسبتاً زیادی از پسوندها برای اشتقاق در مقوله اسم است؛ همچنین اسم در فارسی با نشانه‌های تصریفی خاصی می‌تواند همراه شود. در زبان فارسی معاصر، بسیاری از این پسوندها چندان زایا نیستند. شکل ۱ حاصل جمع‌بندی نشانه‌های تصریفی و اشتقاقی مربوط به اسم است که مستخرج از مجموعه ۱۲ مقاله دکتر علی‌اشرف صادقی تحت عنوان « شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر ۱ تا ۱۲: ۱۳۷۲-۱۳۷۰ » (در بحث اشتقاق) و لازار: ۱۳۸۹ و قطره: ۱۳۸۶ (در بحث تصریف) است. این فهرست تنها پسوندهایی را عرضه می‌دارد که مهم‌تر و زایاتر شمرده می‌شوند، گر چه برخی از پسوندهایی که در فارسی معاصر چندان زایا نیستند نیز بررسی شده است.

مشخصه‌های تصریفی و
اشتقاقی اسم در فارسی



شکل ۱: مشخصه‌های تصریفی و اشتقاقی اسم در زبان فارسی

به منظور تشخیص کلی صفت‌ها و به منظور بررسی امکان هم‌نگاره بودن با اسم‌ها، ساخت‌واژه صفت‌ها مورد بررسی قرار گرفته است. پسوندهای صفت‌ساز از نظر معنایی به چند دسته تقسیم می‌شوند: پسوندهای دارندگی و اتصاف، پسوندهای فاعلی و مفعولی، پسوندهای نسبت، پسوندهای شباهت و پسوندهای حالت. شکل ۲، حاصل جمع‌بندی نشانه‌های تصریفی و اشتقاقی صفت در فارسی است.



شکل ۲: مشخصه‌های تصریفی و اشتقاقی صفت در فارسی

۴- بررسی هم‌نگاره‌ها در فارسی

کلماتی هم‌نگاره خوانده می‌شوند که صورت نوشتاری یکسان، اما معنای (و گاهی، تلفظ) متفاوتی دارند. در واقع این کلمات، املاء یکسان دارند و می‌توان واژه هم‌املاء را نیز در مورد آن‌ها به کار برد. هم‌نگاره‌ها در زبان‌های مختلف یکی از مهم‌ترین لایه‌های ابهام را در سطح متن ایجاد می‌کنند و منجر به چالش‌هایی در پردازش متن می‌شوند. در برخی زبان‌ها، مانند زبان فارسی، به دلیل ساختار نوشتاری خاص آن، تعداد هم‌نگاره‌ها نسبت به دیگر زبان‌ها بیشتر است. تعداد زیادی از هم‌نگاره‌ها از ساختار زبان ناشی می‌شوند، به عبارت دیگر، هم‌نگاره‌ها در هر زبان به آن زبان خاص وابسته‌اند و هر چه ساختار واژه اشتقاقی و تصریفی زبان پیچیده‌تر باشد، می‌توان انتظار داشت هم‌نگاره‌های بیشتری وجود داشته باشد و در نتیجه، ابهام‌زدایی از آن‌ها نیز مشکل‌تر است. با این اوصاف می‌توان دریافت بازشناسی هم‌نگاره‌ها نیز تا حد زیادی وابسته به ساختار زبان باشد (محسنی: ۱۳۸۷). در این بخش به بررسی نظام نوشتاری زبان فارسی پرداخته می‌شود تا از این رهگذر بتوان به شناسایی انواع هم‌نگاره‌ها در زبان فارسی پرداخت، سپس، انواع هم‌نگاره‌ها در زبان فارسی مورد مطالعه قرار خواهند گرفت، از جمله هم‌نگاره‌های اسمی.

۴-۱- برخی از مسائل مربوط به نظام نوشتاری زبان فارسی که در پردازش متن حائز اهمیت است

در نظام نوشتاری زبان فارسی، می‌توان گفت اطلاعات واجی که از طریق صورت نوشتاری کلمات به دست می‌آید کامل نیست، زیرا واکه‌های کوتاه معمولاً نمود نوشتاری ندارند (اسپروت^۱ ۱۶۳: ۲۰۰۰، نقل از علایی و بی‌جن‌خان: ۱۳۹۲). تنها تعداد انگشت‌شماری نگاره وجود دارد که برای نشان دادن واکه‌های کوتاه در خط به کار می‌روند؛ از جمله نگاره «ه»، که برای نشان دادن واکه کوتاه /e/ یا /a/ به کار می‌رود؛ مانند «به» و «نه»؛ و حرف «و» که می‌تواند نمود نوشتاری واکه کوتاه /o/ باشد. مانند «تو». فقدان واکه‌های کوتاه در نظام نوشتاری فارسی باعث ایجاد ابهام

¹ Sproat

می‌شود. به عنوان مثال، هم‌نگاره‌هایی مانند «مرد» /mard/ و «مرد» /mord/ در اثر فقدان واکه‌های کوتاه در نظام نوشتاری فارسی ایجاد شده‌اند (علایی و بی‌جن‌خان: ۱۳۹۲). همچنین استفاده نکردن از برخی از علائم زیروزبری در خط فارسی نیز می‌تواند بعضاً هم‌نگاره‌هایی را ایجاد کند. برخی از این علائم شامل: فتحه، کسره، ضمه، سکون، تنوین، مد، تشدید و همزه است.

۲-۴- طبقه‌بندی هم‌نگاره‌ها در زبان فارسی

هم‌نگارگی در زبان فارسی ناشی از بازنمایی واجی و صرفی عناصر زبانی در خط فارسی است، به این صورت که رابطه‌ای چند-به-چند و در برخی موارد، غیرنظام‌مند میان عناصر واجی و صرفی زبان فارسی و تظاهر نوشتاری آن‌ها در خط فارسی دیده می‌شود (بی‌جن‌خان و مرادزاده: ۱۳۸۳، نقل از محسنی: ۱۳۸۷). هم‌نگاره‌ها در زبان فارسی می‌توانند واژگانی باشند. منظور از هم-نگاره‌های واژگانی، هم‌نگاره‌هایی هستند که در فرهنگ‌های لغات به صورت مدخل‌های مجزا ذکر شده‌اند و حضورشان وابسته به بافت نحوی و اضافه شدن تکواژهای اشتقاقی و تصریفی است. هم‌نگاره‌های موجود در هر دو دسته، می‌توانند هم‌آوا (مانند: «دوش» و «دوش»، «شانه» و «شانه»، «شیر» و «شیر») یا غیرهم‌آوا (مانند: «کند» /kond/، «کند» /kanad/ و «کند» /konad/، «ملک» /malak/، «ملک» /melk/ و «ملک» /molk/ و «ملک» /malek/، «حسن» /hasan/ و «حسن» /hosn/ باشند. هم‌نگاره‌های فارسی را با توجه به علل پیدایش آن‌ها، می‌توان در طبقات مختلفی قرار داد که در زیر به برخی از آن‌ها اشاره می‌شود:

۱-۲-۴. هم‌نگاره‌هایی که در اثر عدم نمایش علائم زیر و زبری در خط فارسی، ایجاد شده‌اند. مانند: «مرد» /mard/ و «مرد» /mord/، «کند» /kanad/، «کند» /kond/ و «کند» /konad/.

۲-۲-۴. هم‌نگاره‌هایی که در اثر عدم تناظر یک-به-یک میان واج‌ها و نگاره‌ها در فارسی ایجاد شده‌اند. مانند: «رود» /rud/ و «رود» /ravad/، «شوم» /sum/ و «شوم» /faval/.

۲-۲-۴- هم‌نگاره‌هایی که در اثر یکسانی تظاهر واجی و نوشتاری تکواژها به وجود می‌آیند. در زبان فارسی ممکن است چندین تکواژ متفاوت، نمود نوشتاری یکسانی داشته باشند که در زیر به برخی از آن‌ها اشاره شده است (بی‌جن‌خان و مرادزاده: ۱۳۸۳، نقل از محسنی: ۱۳۸۷):

۱-۲-۲-۴- یکسان بودن نمود نوشتاری تکواژ یاء نکره، یاء اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنی یا اشیا، تصغیر و تحیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم شخص مفرد، یاء صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یاء متصل به گروه اسمی. مانند:

- الف. یاء نکره: استادی که به موقع سر کلاس حاضر می‌شود.
- ب. یاء اسم‌ساز: وی در سال ۱۳۸۸ به درجهٔ استادی رسید.
- پ. یاء شناسه دوم شخص مفرد: تو دیگر خود استادی.
- ت. یاء نسبت: مشکلات جوانی.
- ث. یاء متصل به گروه اسمی: کارگر ماهر ی که ما می‌شناسیم....

۲-۲-۲-۴- یکسانی تکواژهای ضمیر متصل سوم شخص مفرد با تکواژ اسم‌ساز. مانند:

- الف. ضمیر متصل سوم شخص مفرد: به رویش خندیدم. /rujaʃ/
- ب. تکواژ اسم‌ساز: رویش گل‌ها را بین. /ru`jeʃ/

هم‌نگاره‌ها را در دو سطح می‌توان بررسی کرد: سطح تکواژی و سطح کلمه. منظور از سطح تکواژی، تکواژهایی هستند که فارغ از پایه‌ای که به آن متصل می‌شوند، با تکواژهای دیگر هم‌نگاره هستند. به عنوان مثال، تکواژ دال بر شمار در اسم‌ها، تکواژی که اسم مکان می‌سازد و تکواژی که صفت فاعلی و مفعولی می‌سازد، همه دارای نمود نوشتاری یکسان («ی») هستند. برخی از مواردی که هم‌نگارگی را در سطح تکواژ نشان می‌دهد را می‌توان به صورت زیر دسته‌بندی کرد:

- پسوند «-ین» که هم شمار را در اسم‌ها نشان می‌دهد و هم می‌تواند در صفت‌ها، دال بر نسبت باشد.
 - پسوند «-ه / -4» که می‌تواند نشانه معرفه در گفتار، سازنده اسم اشیا و اسم معنی، تأنیث در اسم‌ها، دال بر دارندگی و اتصاف در صفت‌ها باشد یا صفت فاعلی و مفعولی بسازد.
 - پسوند «-ش» که می‌تواند هم نشانه معرفه در گفتار باشد و هم اسم بسازد.
 - پسوند «-گاه» که اسم مکان و قید زمان می‌سازد.
 - پسوند «-ئیه» که اسم مکان می‌سازد یا دال بر تأنیث است.
 - پسوند «-ک» که تصغیر و تحیب را در اسم‌ها نشان می‌دهد و یا اسم اشیا و معنی می‌سازد.
 - پسوند «-ا» که هم می‌تواند دال بر تأنیث در اسم‌ها باشد و هم مصدر، اسم مصدر، صفت فاعلی و مفعولی بسازد.
 - پسوند «-ئینه» که در صفت‌ها، دال بر نسبت است و در اسم‌ها، اسم معنی و اشیا می‌سازد.
- نوع دوم هم‌نگارگی، هم‌نگارگی در سطح کلمه است. به این صورت که، همانگونه که ذکر شد، کلماتی هم‌نگاره خوانده می‌شوند که صورت نوشتاری یکسان، اما معنای (و گاهی، تلفظ) متفاوتی دارند.

۵- روش شناسی پژوهش

پس از بررسی انواع هم‌نگاره‌ها در زبان فارسی به منظور شناخت آن‌ها در متون گوناگون، لازم است در پیکره‌های متنی گوناگون زبان فارسی به جستجوی هم‌نگاره‌ها پرداخته شود و فهرستی از آن‌ها تهیه شود تا بتوان در تحقیقات بعدی، هم‌نگاره‌هایی که فراوانی بالاتری در پیکره‌ها دارند را شناسایی کرده و اقداماتی در جهت رفع ابهام از آن‌ها در پیکره‌ها انجام داد. پیکره‌های استفاده شده برای این منظور، پیکره متنی زبان فارسی، پایگاه دادگان زبان فارسی و پیکره وابستگی نحوی زبان فارسی می‌باشد.

در مورد برخی پیکره‌ها که فایل آن‌ها موجود است، باید از روش جستجوی ماشینی جهت تهیه فهرست هم‌نگاره‌ها، استفاده می‌شد، نگارنده در این بخش از یک متخصص برنامه‌نویس کمک گرفته است. در سایر موارد از برچسب‌دهی دستی استفاده شده است. روش جستجوی ماشینی هم-نگاره‌ها در پیکره وابستگی نحوی زبان فارسی و پیکره متنی زبان فارسی، که فایل آن‌ها موجود است، به صورت زیر توضیح داده می‌شود:

جستجو برای هم‌نگاره‌ها با یک دستور در زبان برنامه‌نویسی پایتون (Python) انجام شده است؛ به این صورت که کلمات موجود در پیکره، مانند پیکره متنی زبان فارسی، و برچسب‌هایی که به آن کلمات تعلق گرفته است را بررسی می‌کند و کلماتی که دارای بیش از یک برچسب باشند را - به همراه فراوانی هر یک از برچسب‌ها - گزارش می‌کند. فرض کنید یک مجموعه داشته باشیم به نام M، برنامه با خواندن اولین کلمه در پیکره، بررسی می‌کند آیا این کلمه قبلاً در مجموعه M موجود است یا نه. اگر موجود نباشد، کلمه را به همراه برچسب آن به مجموعه M اضافه می‌کند، اگر موجود باشد، بررسی می‌کند که آیا برچسب کلمه‌ای که الان مشاهده کرده است مشابه همان برچسب قبلی است؛ اگر بود یک واحد به فراوانی برچسب قبلی اضافه می‌کند و اگر برچسب جدید بود، آن را در کنار برچسب قبلی قرار می‌دهد. عملکرد برنامه به زبان ساده - به عنوان مثال، برای پیکره بی‌جن‌خان^۱، که پیکره مبسوط‌تر آن، پیکره متنی زبان فارسی است - به صورت زیر است:

- هر سطر در پیکره متنی زبان فارسی شامل یک واژه و برچسب اجزای کلام (POS) آن است.
- یک مجموعه به اسم دیکشنری (Dictionary) داریم که در واقع، یک ساختار داده در زبان برنامه‌نویسی پایتون است؛ ساختار آن بدین شکل است که هر واژه بعلاوه برچسب‌های (POS) آن در مجموعه ذخیره می‌شود. نوع داده دیکشنری در زبان برنامه‌نویسی پایتون، به صورت فهرستی از

^۱ پیکره متنی استاندارد زبان فارسی / پیکره بی‌جن‌خان که در پژوهشکده پردازش هوشمند علامت تهیه شده است، مجموعه‌ای از متون نوشتاری و گفتاری زبان فارسی به صورت رسمی است که از منابع واقعی همچون روزنامه‌ها، سایت‌ها و مستندات از قبل تایپ شده، جمع‌آوری شده، تصحیح کرده و برچسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده است و دارای تنوع بسیار زیادی است (بی‌جن‌خان و همکاران: ۲۰۱۱).

کلیدها (key) و ارزش‌ها (value) است. در دیکشنری‌های پایتون، ترتیبی وجود ندارد و با دو مفهوم کلید (key) و ارزش (value) سر و کار داریم (oruji.org)؛ در مورد پژوهش حاضر، واژه‌ها همان key و برجسب اجزای کلام (POS) مربوط به هر واژه، value آن در نظر گرفته می‌شوند.

- برنامه هر سطر از پیکره را می‌خواند، اگر واژه قید شده در آن سطر قبلاً در Dic وجود نداشت، آن را به همراه POS آن به Dic اضافه می‌کند. اگر قبلاً در Dic وجود داشت، بررسی می‌کند که آیا POS قیدشده در این سطر قبلاً در Dic برای این واژه درج شده است یا نه، اگر درج نشده بود آن را به لیست POSهای این واژه در Dic اضافه می‌کند.

- در نهایت مجموعه Dic بررسی می‌شود و واژه‌هایی که بیش از یک POS برای آن‌ها ثبت شده است به عنوان خروجی برنامه ارائه می‌شوند.

از پایگاه دادگان زبان فارسی^۱ در تحقیق حاضر به این صورت استفاده شده است که فهرستی از کلمات پرسامد در پیکره در اختیار نگارنده قرار گرفت و از آنجائی که فهرست مذکور فاقد هر گونه برجسبی است، نگارنده، خود به صورت دستی کلماتی که می‌تواند بیش از یک برجسب داشته باشند، برجسب‌گذاری کرده است. برای این منظور، نگارنده هر کلمه موجود در فهرست مذکور را در جملات فرضی، به عنوان بافت، در نظر گرفته است تا امکان داشتن بیش از یک برجسب برای هر کلمه را با توجه به بافت بررسی کند، سپس، تنها، کلماتی که در بافت بیش از یک برجسب می‌توانسته‌اند داشته باشند را برجسب‌گذاری کرده است و هر تعداد برجسب ممکن برای هر یک از کلمات مذکور را کنار آن‌ها مشخص کرده است. وی برای کنترل و تایید صحت روایی و درستی تشخیص برجسب‌ها، از یک متخصص زبان و ادبیات فارسی کمک گرفته است.

جمع‌آوری داده‌ها در این مرحله دو کاربرد مهم داشت: ۱- نشان‌دهنده تعداد بسیار زیاد هم-نگاره‌ها در پیکره‌های متنی فارسی است که خود نشان می‌دهد در صورت رفع ابهام نکردن از هم-نگاره‌ها در متون گوناگون، پردازش متن با مسائل و مشکلاتی روبرو خواهد شد. ۲- فهرست تهیه شده از هم‌نگاره‌ها در پیکره‌های گوناگون می‌تواند در پژوهش‌های دیگر مورد استفاده قرار گیرد؛

^۱ پایگاه دادگان زبان فارسی، مجموعه‌ای نرم‌افزاری برای ذخیره، پردازش و ارائه داده‌های زبانی فارسی است. این پایگاه دربرگیرنده پیکره‌های گوناگونی از زبان فارسی است که با وجود حجمی عظیم و با گستردگی و گوناگونی‌های بسیار، دارای ساختاری بسامان و منطقی است و امکان هرگونه جستجو و دستیابی سریع به آگاهی‌های مورد نیاز را در هر زمان فراهم آورده است. پیکره‌های این پایگاه می‌توانند همواره روزآیند شود. پایگاه دادگان زبان فارسی فراگیر و متنوع است. در واقع فراتر از یک یا چند پیکره خاص است و کاربران بر پایه نیاز و هدف پژوهشی خود می‌توانند پیکره مناسب را از آن برگزینند.

به این صورت که می‌توان هم‌نگاره‌هایی که فراوانی بالاتری دارند در بافت نحوی مورد مطالعه قرار داد تا با توجه به ساخت نحوی که در آن به کار رفته‌اند، برجسب‌های درست به آن‌ها اختصاص داده شود.

قسمتی از فهرست تهیه شده در جدول ۱ آورده شده است:

جدول ۱: قسمتی از فهرست هم‌نگاره‌های تهیه شده از پیکره‌های متنی فارسی

فراوانی	کلمات	برجسب‌های هم‌نگاره‌ها
۱۰۳۲۳۶۱	به	P , ADV, N
۱۳۲۱۵۷	کرد	N , V, ADJ
۱۱۱۲۸۶	تا	P , N(NO-V)
۱۰۹۶۰۹	بر	P , N(NO-V)
۱۰۹۳۱۶	کند	V , V , ADV
۷۵۲۵۲	کرده	V , ADJ+V-PRE , ADJ-INO , N
۲۹۵۱۴	اجتماعی	N-IN , ADJ
۲۹۳۶۶	گرفته	V , ADJ , ADJ-INO
۱۷۶۷۰	شهرداری	N-IN , N , N+V-PRE, ADJ
۱۸۱۸	داوری	N-IN , N, N+V-PRE, ADJ

راهنمای برچسب‌ها:

P= حرف اضافه	ADV= قید
ADJ= صفت	N (NO-V) = جزء غیر فعلی فعل مرکب
V= فعل	V-PRE= فعل اسنادی
ADJ-INO= صفت مفعولی	N-IN= اسم نکره
N= اسم	

۶- نتیجه‌گیری

یکی از بخش‌های مهم سامانه‌های برچسب‌دهی به کلمات، برچسب‌گذاری اسم‌ها و صفت‌ها در فارسی است. این سامانه‌ها جهت برچسب‌گذاری اجزای واژگانی کلام مانند: اسم، فعل، قید، صفت و.... تهیه می‌شوند؛ برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمه ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است. با توجه به تعداد بسیار زیاد اسم‌ها در زبان فارسی و هم‌نگاره بودن بسیاری از آن‌ها با صفت‌ها، به راحتی نمی‌توان اسم‌ها را در یک پیکره متنی فارسی برچسب‌گذاری کرد. در تهیه سامانه‌های برچسب‌گذاری برای اسم‌ها و صفت‌ها در زبان فارسی، بررسی ساخت‌واژی اسم‌ها و صفت‌ها از دو نظر حائز اهمیت است: ۱- اگر کلمه‌ای قبلاً در پیکره آموزشی ظاهر نشده باشد، نمی‌توان از پیکره آموزشی اطلاعات دقیقی راجع به آن کلمه به دست آورد و حتی از توزیع کلمات در پیکره نیز نمی‌توان استفاده کرد، زیرا توزیع کلمات ناشناخته کاملاً متفاوت از کلمات شناخته شده در متن است. بنابراین، اگر در یک پیکره متنی فارسی، کلمه‌ای در واژگان حضور نداشته باشد (کلمه خارج از واژگان)، نمی‌توان برچسب‌هایی که کلمه به آن‌ها منتسب می‌شود را بازیابی کرد. در این صورت، برچسب کلمه را

تنها می‌توان با توجه به شکل کلمه یا بافتی که کلمه در آن ظاهر می‌شود، یا هر دو، حدس زد. منظور از شکل کلمه، ویژگی‌هایی مانند انواع پیشوندها و پسوندهایی که به کلمات متصل می‌شوند، است. بنابراین، در حله اول، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، فارغ از بافت، ضروری به نظر می‌رسد تا از این رهگذر بتوان اسم‌ها و صفت‌ها را در صورت دارا بودن شاخص‌های صرفی و اشتقاقی، شناسایی و سپس برجسب‌دهی کرد. ۲- زبان فارسی ظرفیت بالایی برای ساخت هم-نگاره‌های جدید که از ساخت‌واژه فارسی نشأت می‌گیرند، را نیز دارد. بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابه دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به عنوان مثال، نمود نوشتاری وند تصریفی نکره‌ساز /i/ و وند اشتقاقی صفت‌ساز /i/ یکسان است («-ی») و کلماتی مانند: «آسمانی» و «اسلامی»، فارغ از بافت، می‌توانند هم برجسب «اسم» داشته باشند و هم «صفت». بنابراین، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، به منظور تفکیک آن‌ها از هم ضروری به نظر می‌رسد، زیرا اکثر صفت‌ها در بافت‌های گوناگون، با صورت نوشتاری یکسان، می‌توانند برجسب «اسم» بگیرند.

در پژوهش حاضر، پس از مقدمه، ابتدا به پیشینه تحقیق پرداخته شد. سپس ساخت‌واژه اسم‌ها و صفت‌ها در فارسی بررسی شد تا از این رهگذر بتوان اسم‌ها و صفت‌ها را در صورت دارا بودن شاخص‌های صرفی و اشتقاقی، شناسایی و سپس برجسب‌دهی کرد. بخش چهارم، مطالعه هم-نگاره‌ها در فارسی بود که به بررسی نظام نوشتاری زبان فارسی پرداخته شد تا از این رهگذر بتوان به شناسایی انواع هم‌نگاره‌ها در زبان فارسی پرداخت، سپس، انواع هم‌نگاره‌ها در زبان فارسی مورد مطالعه قرار گرفت، از جمله هم‌نگاره‌های اسمی و صفتی. در بخش پنجم به روش جمع‌آوری داده‌ها از پیکره‌ها پرداخته شد؛ پیکره‌های استفاده شده برای این منظور، شامل پیکره متنی زبان فارسی، پایگاه دادگان زبان فارسی و پیکره وابستگی نحوی زبان فارسی می‌باشد و جستجو به دو روش جستجوی ماشینی و دستی انجام گرفت و فهرست مبسوطی از هم‌نگاره‌های به دست آمده از پیکره‌های مذکور تهیه شد نمونه‌ای از این فهرست در جدول ۱ آورده شده است. بررسی کلی هم-نگاره‌ها در پیکره‌های مورد مطالعه نشان می‌دهد که تعداد هم‌نگاره‌ها در پیکره‌ها قابل توجه است و حتی می‌توان گفت، بیشتر هم‌نگاره‌ها کلماتی هستند که فراوانی آن‌ها در پیکره‌ها زیاد است. اکثر این هم‌نگاره‌ها، به نظر می‌رسد در اثر یکسان بودن نمود نوشتاری تکواژ یاء نکره، یاء اسم‌ساز

(اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنی یا اشیا، تصغیر و تحییب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم شخص مفرد و یاء صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یاء متصل به گروه اسمی باشند. مانند:

الف. یاء نکره: کشاورزی که محصول خود را به موقع درو می‌کند.

ب. یاء اسم‌ساز: وی چند سال کشاورزی را دنبال می‌کرد.

پ. یاء شناسه دوم شخص مفرد: تو خود کشاورزی.

و یا در اثر یکسانی تکواژهای ضمیر متصل سوم شخص مفرد با تکواژ اسم‌ساز به وجود آمده‌اند. مانند:

الف. ضمیر متصل سوم شخص مفرد: به رویش خندیدم.

ب. تکواژ اسم‌ساز: رویش گل‌ها را ببین.

جمع‌آوری داده‌ها در این مرحله دو کاربرد مهم داشت: ۱- نشان‌دهندهٔ تعداد بسیار زیاد هم‌نگاره‌ها در پیکره‌های متنی فارسی است که خود نشان می‌دهد در صورت رفع ابهام نکردن از هم‌نگاره‌ها در متون گوناگون، پردازش متن با مسائل و مشکلاتی روبرو خواهد شد. ۲- فهرست تهیه شده از هم‌نگاره‌ها می‌تواند پژوهش‌های گوناگون مورد استفاده قرار گیرد؛ به این صورت که هم‌نگاره‌هایی که فراوانی بالاتری دارند در بافت نحوی مورد مطالعه قرار گیرند و با توجه به ساخت نحوی که در آن به کار رفته‌اند، برجسب‌های درست به آن‌ها تخصیص داده شود.

۷- پیشنهاد برای پژوهش‌های بیشتر

در پژوهش حاضر به بررسی ساخت‌واژی اسم‌ها و صفت‌ها در فارسی پرداخته شده است؛ همچنین پس از بررسی تعداد و انواع هم‌نگاره‌ها در پیکره‌های متنی فارسی، این نتیجه به دست آمد که تعداد زیادی از هم‌نگاره‌ها در اثر یکسان بودن نمود نوشتاری/ املای وندهای اشتقاقی و تصریفی، مخصوصاً وندهایی با نمود نوشتاری «-ی» ایجاد شده‌اند. پیشنهاد می‌شود این هم‌نگاره‌ها

که فراوانی بالایی در پیکره‌های متنی فارسی دارند، در پیکره‌ها در بافت نحوی مورد مطالعه قرار گیرند. ضمناً از آنجائی که در فارسی هم‌نگاره‌های دیگری نیز وجود دارد که در اثر عدم نمایش واژه‌های کوتاه در خط فارسی یا عدم وجود تناظر یک-به-یک میان واج‌ها و نگاره‌ها ایجاد می‌شوند، پیشنهاد می‌شود چنین هم‌نگاره‌هایی نیز در پیکره‌های متنی فارسی مورد مطالعه قرار گیرند.

فهرست منابع

- Assi, M. and Haji Abdolhosseini, M. 2000. "Grammatical tagging of a Persian corpus". *International journal of corpus linguistics*. 5(1). 69-81.
- Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from Building a Persian Written Corpus: Peykare. *Language Resources and Evaluation*, 45(2), 143-164.
- Elhaj, M. 2009. "Statistical part of speech tagger for traditional Arabic texts". *Journal of computer science*, 5 (11) 794-800.
- Frodl, M. 2013. *Part of speech tagging using Neural Networks*, Masaryk University, faculty of informatics.
- <https://github.com/sobhe/hazm>
- <https://oruji.org/python/dictionary>
- Hearst, M. A. 1991. "Noun homograph disambiguation using local context in large text corpora". *Proceedings of the 17th annual conference of the university of Waterloo Centre for the new OED and text research*. Oxford.
- Hu, Y., Matveeva, I., Goldsmith, J., and Sprague, C., 2005. "Using morphology and syntax together in unsupervised learning". In *Proceeding of the workshop on Psychocomputational Models of Human Language Acquisition*. PMHLA 05. 20-27. Stroudsburg: PA, USA. Association for computational linguistics
- Megerdooian, K. 2000. "Unification-based Persian morphology". In *Proceedings of CICLing, Centro de investigacion en computacion- IPN, Mexico*: 311-318.
- Megerdooian, K. 2004. "Developing a Persian part-of-speech tagger". *Proceedings of the 1st workshop on Persian language and computer*. 99-105.
- Merialdo, B. 1994. "Tagging English text with a probabilistic model". *Computational linguistics*. 20. 155-171.

- Schuetze, Hinrich. (1995). "Distributional Part-of-Speech Tagging," From Texts to Tags: Issues in Multilingual Language Analysis. Online Proceedings of the ACL SIDGAT Workshop. On the Internet at <http://xxx.lanl.gov/find/cmp-lg>

- آخشیک، س. و فتاحی، ر. ۱۳۹۱. «تحلیل چالش‌های پیوسته نویسی و جدا نویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی». فصلنامه کتابداری و اطلاع‌رسانی، ۱۵ (۳): ۹-۳۰.
- راسخ، ا، صدرالدینی، م. و فخر احمد، س. م. ۲۰۱۴. «ابهام‌زدایی کلمات بر اساس ویژگی‌های لغوی و معنایی کلمات، با استفاده از طبقه‌بندی بیزین». *Journal of computing and security*. شیراز: دانشکده برق و کامپیوتر، دانشگاه شیراز. ۱ (۲).
- صادقی، ع. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱)». تهران: نشر دانش. شماره ۶۴.
- صادقی، ع. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۲)». تهران: نشر دانش. شماره ۶۵.
- صادقی، ع. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۳)». تهران: نشر دانش. شماره ۶۷.
- صادقی، ع. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۴)». تهران: نشر دانش. شماره ۶۹.
- صادقی، ع. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۵)». تهران: نشر دانش. شماره ۷۰.
- صادقی، ع. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۶)». تهران: نشر دانش. شماره ۷۱.
- صادقی، ع. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۷)». تهران: نشر دانش. شماره ۷۲.

- صادقی، ع. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۸)». تهران: نشر دانش. شماره ۷۴.
- صادقی، ع. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۹)». تهران: نشر دانش. شماره ۷۵.
- صادقی، ع. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۰)». تهران: نشر دانش. شماره ۷۶.
- صادقی، ع. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۱)». تهران: نشر دانش. شماره ۷۷.
- صادقی، ع. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲)». تهران: نشر دانش. شماره ۷۹ و ۸۰.
- عاصی، م. ۱۳۸۳. «پردازش دستوری زبان فارسی با رایانه». ویژه‌نامه نامه فرهنگستان (دستور). شماره ۱.
- علایی، ا. و بی‌جن‌خان، م. ۱۳۹۲. «عمق خط فارسی». پژوهش‌های زبانی (مجله سابق دانشکده ادبیات و علوم انسانی دانشگاه تهران). دوره ۴. شماره ۱.
- قطره، ف. ۱۳۸۶. «مشخصه‌های تصریفی در زبان فارسی امروز». دستور. شماره ۳. ۵۲-۸۱.
- کشانی، خ. ۱۳۷۱. اشتقاق پسوندی در زبان فارسی امروز. تهران: مرکز نشر دانشگاهی.
- لازار، ژ. ۱۳۸۹. دستور زبان فارسی معاصر. ترجمه مهستی بحرینی و توضیحات و حواشی هرمز میلانیان. تهران: انتشارات هرمس. چاپ دوم.
- محسنی، م. و مینایی بیدگلی. ب. ۱۳۸۸. «سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی». دو فصل‌نامه پردازش علائم و داده‌ها. شماره ۲. پیاپی ۱۲.

- محسنی، م. ۱۳۸۷. سیستم برجسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی. تهران: دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر.
- مرادی، م و وزیرنژاد، ب. ۱۳۹۱. «ساخت پیکره‌های نشانه‌گذاری شده با رویکرد وب به عنوان پیکره». مجموعه مقالات دومین هم‌اندیشی زبان‌شناسی رایانشی. تهران: انجمن زبان‌شناسی ایران.
- مسعودی، ب.، قوچانی، س. و استاجی، ا. ۱۳۸۹. «یک روش بیزی برای رفع ابهام معنایی کلمات در زبان فارسی با تأکید بر ویژگی‌های محلی کلمه». اولین کنفرانس ملی محاسبات نرم و فناوری اطلاعات. ماهشهر: دانشگاه آزاد اسلامی، واحد ماهشهر.

A Corpus-based study of Persian noun and adjective homographs to help right POS tagging

General Linguistics, phd. Iranian Research Institute for Information Science and Technology (IranDoc)

Elham Alayiaboozar, Alayi@irandoc.ac.ir

Abstract:

Present research studies morphological structure of nouns and adjectives; there are two main reasons for studying them in the process of making any POS tagger system for tagging nouns: 1. If the system faces an out of vocabulary word (OOV word), one way to identify its tag would be considering its morphological structure. 2. In Persian, lots of homographs are made due to Persian complex morphology; studying morphological structure of nouns in order to distinguish them from adjectives seems to be necessary, since many adjectives, having the same orthographic forms of nouns, would be wrongly tagged as “noun” or vice versa. After studying morphological structure of nouns and adjectives in present study, Persian writing system is studied; then definition of homographs and the related classifications are presented. Finally, the study uses different famous Persian corpora (including Bijankhan, and syntactical dependency corpus (vabastegi ye nahvi) for searching for homographs (using search tools) and Data center for Persian language (Paygah e Dadegan) whose non-tagged file was available (the homographs are searched and tagged manually)) to make a list of homographs. The result of studying the mentioned list showed that the frequency of homographs, especially those which are made due to identical orthographic form of indefinite morpheme, adjective-maker morpheme and second person inflectional morpheme is high in Persian corpora which makes POS tagging difficult.

Key words: POS tagger system, morphological structure of Persian nouns and adjectives, Persian writing system, homographs



الهام علایی ابوذر، دانش‌آموخته دکترای تخصصی در رشته زبان‌شناسی همگانی در سال ۱۳۹۲ از دانشگاه تهران و کارشناسی ارشد زبان‌شناسی همگانی در سال ۱۳۸۶ از همان دانشگاه و کارشناسی زبان و ادبیات انگلیسی در سال ۱۳۸۳ از دانشگاه گیلان است. وی همکاری خود را با پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) از سال ۱۳۹۴، به عنوان عضو هیئت علمی، آغاز نموده است و هم‌اکنون استادیار پژوهشی این پژوهشگاه است. بیشتر فعالیت‌های وی تاکنون در حوزه زبان‌شناسی رایانه‌ای (سیستم‌های تبدیل متن به گفتار، نظام‌های نوشتاری، یادگیری ماشینی و پردازش زبانی)، زبان‌شناسی نظری و پیکره‌ای بوده است و در حوزه‌های ذکر شده با پژوهشگاه علوم و فناوری اطلاعات ایران همکاری می‌نماید.