A Corpus-based Study of Persian Noun and Adjective Homographs to help Correct POS Tagging

Elham Alayiaboozar
PhD of General Linguistics; Iranian Research Institute for Information Science and Technology (IranDoc);
Email: Alayi@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 14, Oct. 2017 **Accepted: 25, Feb. 2018**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 2 | pp. 897-922
Winter 2019

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 vic e 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.

Keywords: POS Tagger System, Morphological Structure of Persian Nouns and Adjectives, Persian Writing System, Homographs

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

الهام علایی ابوذر

دکتری زبان‌شناسی همگانی؛ استادیار؛
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
Alayi@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۶/۰۷/۲۲ | پذیرش: ۱۳۹۶/۱۲/۰۶ | مقاله برای اصلاح به مدت شش روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳٤ | شمارهٔ ۲ | ص‌ص ۸۹۷-۹۲۲
زمستان ۱۳۹۷

چکیده: در تهیهٔ سامانه‌های برچسب‌گذاری اجزای کلام در زبان فارسی، بررسی ساخت‌واژی اسم‌ها و صفت‌ها از دو نظر حائز اهمیت است: ۱. اگر در یک پیکرهٔ متنی فارسی کلمه‌ای در واژگان حضور نداشته باشد (کلمهٔ خارج از واژگان)، نمی‌توان برچسب‌های مربوط به کلمه را بازیابی کرد. در این صورت، برچسب 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 ساخت‌واژهٔ فارسی نشأت می‌گیرند، نیز دارد. بنابراین، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، به‌منظور تفکیک آن‌ها از هم ضروری به نظر می‌رسد، زیرا اکثر صفت‌ها در بافت‌های گوناگون، با صورت نوشتاری یکسان می‌توانند برچسب «اسم» بگیرند. در تحقیق حاضر ساخت‌واژهٔ اسم‌ها و صفت‌ها در فارسی بررسی شده است. نظام نوشتاری زبان فارسی نیز مورد بررسی قرار گرفته تا از این رهگذر بتوان به شناسایی انواع هم‌نگاره‌ها در زبان فارسی پرداخت. سپس، انواع هم‌نگاره‌ها در زبان فارسی مورد مطالعه قرار گرفته و در نهایت، از طریق جست‌وجو به دو روش ماشینی و دستی، فهرست مبسوطی از هم‌نگاره‌ها از پیکره‌های «پیکرهٔ متنی زبان فارسی»، «پایگاه دادگان زبان فارسی» و «پیکرهٔ وابستگی نحوی زبان فارسی» تهیه شده است. بررسی کلی هم‌نگاره‌ها در پیکره‌های مورد مطالعه نشان می‌دهد که بیشتر هم‌نگاره‌ها، فراوانی بالایی در پیکره‌های متنی فارسی دارند و اکثر آن‌ها در اثر یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز، شناسهٔ دوم شخص مفرد، یای صفت‌ساز و یای متصل به گروه اسمی ایجاد شده‌اند.

کلیدواژه‌ها: برچسب‌گذاری اجزای کلام، ساخت‌واژهٔ اسم‌ها و صفت‌ها، نظام نوشتاری، هم‌نگاره‌ها

۱. مقدمه

مفهوم برچسب‌گذاری اجزای کلام[1]، مفهومی برگرفته از زبان‌شناسی است که به طبقهٔ نحوی 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 بر اساس مقولهٔ آن کلمه در متن (مانند «اسم»، «فعل»، «قید»، «صفت»، و غیره) صورت می‌پذیرد. کلمات در جایگاه‌های مختلف می‌توانند برچسب‌های واژگانی متفاوت داشته باشند. به‌عنوان مثال، کلمهٔ «آسمانی» در جملهٔ «فردا آسمانی صاف در انتظار شهروندان عزیز خواهد بود»، دارای برچسب «اسم، نکره»، ولی در جملهٔ «او انسانی آسمانی بود»، دارای برچسب «صفت» است. همچنین، کلمهٔ «مردی» در جملهٔ «در راه مدرسه مردی را دیدم»، دارای برچسب «اسم، نکره»، ولی در جملهٔ «مثلاً تو مردی؟»، دارای برچسب «اسم+ فعل اسنادی» است. بنابراین، درصد بالایی از کلمات از نقطه‌نظر برچسب واژگانی، مبهم هستند و می‌توانند فارغ از بافت، بیش از یک برچسب واژگانی داشته باشند.

برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی[2] از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جست‌وجو و کمک به مدل‌های آماری است (Megerdoomian 2004). همچنین، در طراحی سیستم نمایه‌ساز ماشینی، یکی از بخش‌ها، طراحی زیرسیستم‌هاست 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 دربردارد. بنابراین، سیستم برچسب‌دهی خودکار، ماهیت مقولهٔ کلمات را مشخص می‌کند تا بتوان در مراحل بعدی از این اطلاعات در جهت استخراج کلیدواژه‌ها یا هر نوع بازیابی اطلاعات از متن استفاده کرد. تاکنون از مدل‌ها و روش‌های بسیاری برای برچسب‌گذاری

1. part of speech (POS) tagging
2. natural language processing

در زبان‌های مختلف استفاده شده است. این روش‌ها و مدل‌ها را می‌توان به دو دستهٔ کلی تقسیم‌بندی کرد: دستهٔ اول، رویکردهای آماری است که از به‌کار بردن روش‌های احتمالاتی برای پیکره‌های برچسب خورده[1] بهره می‌جویند و دستهٔ دوم، رویکردهای غیرآماری و مبتنی بر دانش[2]/ مبتنی بر قاعده[3] 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[4] خوانده می‌شود (Megerdoomian 2004).

در زبان‌هایی که ساخت‌واژهٔ 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، مقولهٔ نحوی را تغییر نمی‌دهند. اما تکواژهای اشتقاقی کلمات جدید می‌سازند و دارای مفهومی خاص هستند یا مقولهٔ نحوی کلمه را تغییر می‌دهند. بررسی ساخت‌واژی کلمات از دو نظر حائز اهمیت است: ۱. شناسایی برچسب کلماتی که در واژگان وجود ندارند و برچسب نحوی چنین کلماتی را تنها می‌توان با توجه به شکل کلمه یا بافتی که کلمه در آن ظاهر می‌شود، یا هر دو، حدس زد. ۲. کمک به رفع ابهام از برچسب نحوی هم‌نگاره‌ها[5] در فارسی. بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابهی دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به‌عنوان 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». گویشوران فارسی به‌دلیل آشنایی کامل با قواعد ساخت‌واژی زبان خود، هنگام مواجهه با چنین کلماتی که حاوی تکواژهای تصریفی و اشتقاقی هستند یا برای اولین‌بار 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.

1. tagged corpora
2. knowledge-based
3. rule-based
4. hybrid tagging
5. homographs

این سامانه‌ها، به‌خصوص سامانه‌های برچسب‌گذاری، به‌دلیل عدم اشراف به قواعد ساخت‌واژی زبان، در برخورد با کلمات دارای پیچیدگی‌های 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 یک‌سو به‌علت اختلاف نظر پدیدآورندگان متون و از سوی دیگر، پیچیدگی‌های ذاتی خود، به‌هنگام ذخیره، جست‌وجو و بازیابی اطلاعات، چالش‌های متعددی را برای طراحان و نمایه‌سازان پایگاه‌ها، کاربران و پدیدآورندگان منابع به‌وجود آورده است (آخشیک و فتاحی ۱۳۹۱). این پیچیدگی‌ها مشکلات بسیاری را در مسیر برچسب‌گذاری رایانه‌ای اجزای واژگانی کلام ایجاد می‌کنند. برخی از این مشکلات شامل موارد زیر است:

الف. پیچیدگی ساخت‌واژهٔ زبان فارسی و ارتباط آن با خط فارسی: اگر چند وند در یک کلمه ظاهر شوند، همهٔ آن‌ها معمولاً به کلمه می‌چسبند (Megerdoomian 2000). نشانه‌های جمع، کسرهٔ اضافه، نشانهٔ نکره، ضمایر ملکی و غیره می‌توانند به کلمه متصل شوند؛ مانند: «کتابهایم» که شامل «کتاب + ها (علامت جمع) + < ی > (نمود نوشتاری همخوان میانجی /j/) + م (ضمیر ملکی اول شخص مفرد)» است. این ویژگی باعث اشکال متفاوتی از کلمات با ریشهٔ یکسان می‌گردد و این کلمات در سامانه‌های محاسباتی، متفاوت از یکدیگر فرض می‌شوند. در مورد افعال می‌توان گفت از نقطه‌نظر ساخت‌واژی، فعل شامل بن فعل و وندهای تصریفی است. افعال با توجه به شخص و شمار صرف می‌شوند و بنابراین، اشکال متفاوتی از آن‌ها به‌وجود می‌آید. شکل یکسان برخی از تکواژها نیز باعث ابهام در متون فارسی می‌شود. به‌عنوان مثال، پسوند «-ی» در کلمهٔ «مردی»، فارغ از بافت، می‌تواند هم نشانهٔ نکره در نظر گرفته شود و هم شناسهٔ دوم شخص مفرد در یک فعل اسنادی («مرد» + «ی» (به معنی مرد هستی)). علاوه بر این، در فارسی معمولاً واکه‌های کوتاه نمود نوشتاری ندارند و این مسئله نیز باعث ابهام در تحلیل می‌گردد؛ مانند: کلمهٔ «مردم» که می‌تواند به‌صورت‌های /mardam/ و /mordam/ و /mardom/ تلفظ شود. این ابهام، ابهام هم‌نگاره خوانده می‌شود.

ب. عبارت‌های چندکلمه‌ای: حضور افعال مرکب و کلمات مرکب اسمی مشکلی دیگر در پردازش زبانی محسوب می‌شود. به‌عنوان مثال، در کلمات مرکب اسمی که هستهٔ آغاز

هستند، یعنی هستهٔ گروه اسمی در ابتدای عبارت ظاهر شده است، وند جمع به قسمت اول متصل می‌شود. مانند: ماشین لباس‌شویی + وند جمع = ماشین‌های لباس‌شویی.

ج. 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 که کلمات مجزا یک قطعه[1] در نظر گرفته شوند (مانند: «رفتندمردم»). همچنین، وندهای تصریفی (مانند: «-تر»، «-ترین»، «-ها»، «-می») را می‌توان به سه صورت به ستاک مربوط کرد: ۱) به ستاک چسباند، ۲) به‌صورت مجزا (غیرچسبان) با یک فاصله از ستاک در آخر کلمه/ قطعه آورد، و ۳) به‌صورت غیرچسبان ولی با استفاده از نویسهٔ نیم‌فاصله به ستاک مربوط کرد. به‌عنوان مثال، تکواژ جمع «-ها» در اسم‌ها می‌تواند به چند شکل ظاهر شود. در مورد کلمهٔ «کتاب»، سه شکل «کتابها»، «کتاب‌ها» و «کتاب ها» برای حالت جمع در خط فارسی وجود دارد. در مورد افعال نیز می‌توان به پیشوند تصریفی زمان استمراری (یعنی «می-») اشاره کرد. مانند، حال استمراری اول شخص مفرد بن مضارع « رو» که هنگام اضافه‌شدن وند تصریفی «-می» می‌تواند به سه شکل «میروم»، «می‌روم» و «می روم» نوشته شود. این مورد نیز عامل تأثیرگذار دیگری در برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی محسوب می‌شود، زیرا چند شکل نوشتاری متفاوت از یک کلمه عملاً به‌عنوان کلمات متفاوت تفسیر می‌شوند. هر سامانهٔ برچسب‌گذاری باید توانایی تشخیص این شکل‌ها را داشته باشد تا برچسب دستوری صحیح به کلمه بدهد (Megerdoomian 2004). لازم به ذکر است که «فرهنگستان زبان و ادب فارسی» در باب پیوسته‌نویسی یا جدانویسی ترکیبات در زبان فارسی سه فرض را متصور است: ۱. تدوین قواعدی برای جدانویسی همهٔ کلمات مرکب و تعیین موارد استثنا، ۲. تدوین قواعدی برای پیوسته‌نویسی همهٔ کلمات مرکب و تعیین موارد استثنا، و ۳. تدوین قواعدی برای جدانویسی الزامی بعضی

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. به‌عنوان مثال، کلمه‌ای مانند « می‌روم» یک قطعهٔ 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).

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 در آخشیک و فتاحی ۱۳۹۱).

د. قطعه‌های پیچیده[1]: منظور از قطعه‌های پیچیده، شکل‌های چندجزئی هستند که در واقع، از چسبیدن مقولات دستوری (مانند: «به»، «را»، «که» و «آن») به ستاک به‌وجود می‌آیند که این تکواژهای وابسته/ مقولات دستوری چسبیده به ستاک، تکواژ شبه‌کلمه[2] خوانده می‌شوند. مانند: «بدین صورت»، «بعقیده»، «اینکار».

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

و. مرز عبارت‌ها: مسائل گوناگونی باعث ابهام در ساختار عبارت‌های اسمی در فارسی می‌شود. تکواژهای آشکار بسیار اندکی در فارسی برای نشان‌دادن مرزهای عبارات اسمی وجود دارد. اغلب هیچ جزئی در خط فارسی برای متصل‌کردن اجزای یک عبارت اسمی وجود ندارد، زیرا تکواژ اضافه[3]، اغلب واکهٔ کوتاهی است که در خط ظاهر نمی‌شود. علاوه بر این، از آنجا که ترتیب پایه و غالب اجزای جمله در فارسی، فاعل-مفعول-فعل[4] است، فقدان تکواژهای آشکار نشان‌دهندهٔ مرزها، تشخیص محل به پایان رسیدن فاعل و شروع مفعول را سخت می‌کند و دشواری‌هایی را در پردازش متن پدید می‌آورد (Megerdoomian 2004).

از میان مسائل مطرح‌شدهٔ فوق، مسائلی که از ساخت‌واژهٔ 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

1. complex tokens
2. word-like morpheme
3. ezafe morpheme
4. subject-object-verb (SOV)

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 تهیهٔ سامانه‌های برچسب‌گذاری در زبان فارسی (نمونهٔ فارسی: سامانهٔ «هضم» sobhe/hazm())، بررسی ساخت‌واژی اسم‌ها و صفت‌ها از دو نظر حائز اهمیت است:

۱. اگر کلمه‌ای قبلاً در پیکرهٔ متنی ظاهر نشده باشد، نمی‌توان از پیکره، اطلاعات دقیقی راجع به آن کلمه به‌دست آورد و حتی از توزیع کلمات در پیکره نیز نمی‌توان استفاده کرد، زیرا توزیع کلمات ناشناخته کاملاً متفاوت از کلمات شناخته‌شده در متن است. بنابراین، اگر در یک پیکرهٔ متنی فارسی، کلمه‌ای در واژگان حضور نداشته باشد (کلمهٔ خارج از واژگان)[1] نمی‌توان برچسب‌هایی را که کلمه به آن‌ها منتسب می‌شود، 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 است که به کلمات متصل می‌شوند. طبق گزارشات، روش‌هایی که بر استفاده از چنین اطلاعاتی مبتنی هستند، حاوی دقت بالاتری در زبان انگلیسی هستند (Brill 1994 و weischedel 1993 نقل در محسنی ۱۳۸۷). به‌عنوان مثال، یکی از این روش‌ها، روش برچسب‌گذار زیراکس[2] است که در آن از قوانینی بر اساس بخش‌های پایانی کلمات، یعنی حروف انتهای کلمات، برای انتساب مجموعه‌ای از برچسب‌ها به کلمات خارج از واژگان استفاده می‌کند (Cutting et al. 1992 نقل در محسنی ۱۳۸۷). بنابراین، در وهلهٔ 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 برچسب‌دهی کرد.

۲. زبان فارسی ظرفیت بالایی برای ساخت هم‌نگاره‌های جدید را نیز که از ساخت‌واژهٔ فارسی نشأت می‌گیرند، دارد. بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابهی دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به‌عنوان 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

1. out of vocabulary word (OOV word)
2. xerox

«اسم» داشته باشند و هم «صفت». بنابراین، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، به‌منظور 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 پیشینهٔ پژوهش و بررسی مطالعات انجام‌شده در زمینهٔ ساخت‌واژهٔ فارسی و سامانه‌های برچسب‌دهی پرداخته می‌شود. سپس، خلاصه‌ای از بررسی ساخت‌واژهٔ اسم‌ها و صفت‌ها در فارسی ارائه خواهد شد و به‌دنبال آن، بخش بعد، مطالعهٔ هم‌نگاره‌ها در فارسی است و در نهایت، روش جمع‌آوری داده‌ها و معرفی پیکره‌ها و پایگاه‌های دادگان و نتیجه‌گیری ارائه خواهد شد.

۲. پیشینهٔ پژوهش

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

۲-۱. پیشینهٔ مطالعاتی که در زمینهٔ برچسب‌گذاری اجزای واژگانی کلام انجام شده است

«هرست» روشی دقیق و ساده برای ابهام‌زدایی از هم‌نگاره‌های اسمی معرفی می‌کند که از پیکرهٔ متنی بزرگی در این راستا بهره می‌گیرد. الگوریتم معرفی‌شده، 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 (Hearst 19991). «مریالدو» «مدل مارکوفی پنهان»[1] را معرفی می‌کند که در آن دو کلمهٔ قبل به‌عنوان سابقه[2] در نظر گرفته می‌شود. در این مدل، دادن هر برچسبی بستگی به برچسب دو کلمهٔ قبل دارد که به‌صورت زیر نمایش داده می‌شود:

1. Hidden Markov Model (HMM) 2. history

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

(Merialdo 1994). «عاصی و عبدالحسینی» سامانهٔ برچسب‌دهی نحوی را به‌عنوان پروژه‌ای در «پژوهشگاه علوم انسانی و مطالعات فرهنگی» معرفی می‌کنند که سامانهٔ مذکور، خود، در خدمت تهیهٔ پیکره‌ای فارسی به نام «پایگاه دادگان زبان فارسی»[1] قرار خواهد گرفت و می‌توان گفت که اولین تلاش برای تهیهٔ سامانه‌ای جهت برچسب‌دهی نحوی زبان فارسی است. آن‌ها به بررسی روشی می‌پردازند که Schuetze (1995) برای برچسب‌دهی نحوی در متون انگلیسی به‌کار برده است و هدف آن‌ها بررسی به‌کارگیری همان روش در زبان فارسی است. آن‌ها معتقدند که با بهره‌گیری از محاسبات دقیق‌تر می‌توان همان روش را در فارسی به‌کار برد (Assi and Abdolhosseini 2000). «هو» و همکاران از طریق اضافه‌کردن اطلاعات نحوی در کنار اطلاعات ساخت‌واژی مربوط به هر کلمه در داخل شبکهٔ مربوطه به بهبود «لینگویستیکا»[2] کمک می‌کنند. آن‌ها علاوه بر درج فهرست ستاک‌ها و وندها، فهرست دیگری که مربوط به برچسب‌های نحوی کلمات است در شبکه وارد کرده‌اند (Hu et al. 2005). «الحاج» نحوهٔ برچسب‌دهی کلمات موجود در «قرآن کریم» را معرفی می‌کند که خود بخشی از سامانه‌های تجزیه و تحلیل رایانه‌ای «قرآن کریم» است (Elhaj 2009). در واقع، سامانه‌ای که وی معرفی می‌کند، تلفیقی است از تجزیه و تحلیلی ساخت‌واژی و «مدل مارکوفی پنهان» که بر اساس ساختار جملات عربی است. «فرودل» برچسب‌دهی با استفاده از شبکه‌های عصبی را معرفی و برای این منظور از دو روش استفاده می‌کند: ۱. استفاده از شبکه‌های پرسپترون چندلایهٔ تکرارشونده که محتمل‌ترین برچسب نحوی برای کلمه را در متن به‌طور احتمالاتی در نظر می‌گیرد. ۲. در روش دوم کلمات به بردارهایی از مشخصه‌ها در فضای چندبعدی تبدیل می‌شوند. وی معتقد است روش معرفی‌شده ۹۴ درصد برچسب‌دهی نحوی درست از متن ارائه می‌دهد (Frodl 2013). «راسخ و فخراحمد» 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 نمایند و به این ترتیب، راهکاری در جهت رفع مشکلات ترجمهٔ ماشینی و بهبود کیفیت آن ارائه می‌کنند (۲۰۱۴). «محسنی» با در نظر گرفتن مسائل و مشکلاتی که در مسیر برچسب‌گذاری

1. Farsi Linguistic Data Base (FLDB) 2. Linguistica

اجـزای کلام در فارسـی وجـود دارد، ابتـدا طرحـی کلـی بـرای برچسـب‌گذاری خـودکار بـا دقـت بـالا در زبـان فارسـی پیشـنهاد می‌کنـد. سـپس، تحلیـل سـاخت‌واژی و اسـتفاده از آن را بـرای پوشـش دادن تعـداد زیـادی از برچسـب‌های پیکـره بـا حفـظ دقـت بـالا در برچسـب‌گذاری کلمـات مـورد بررسـی دقیق‌تـر قـرار داده و تأثیـر وجـود یـک تحلیلگـر سـاخت‌واژی در سـطح تصریـف را بـر برچسـب‌گذاری اجـزای واژگانـی کلام در زبـان فارسـی بررسـی می‌کنـد. وی معتقـد اسـت کـه نتایـج به‌دسـت‌آمده نشـان از کارایـی بسـیار مناسـب ایـن روش پیشـنهادی در برچسـب‌گذاری دارد (۱۳۸۸).

۲-۲. نمونه‌هایـی از مطالعاتـی کـه در زمینـهٔ سـاخت‌واژهٔ اسـم‌ها و بعضـاً، صفت‌هـا در فارسـی صـورت گرفتـه اسـت

«صادقـی» طـی ۱۲ مقالـه، بـه شـیوه‌ها و امکانـات واژه‌سـازی در زبـان فارسـی معاصـر می‌پـردازد. وی ایـن گفتارهـا را بـا بحـث از اشـتقاق در زبـان فارسـی معاصـر آغـاز می‌کنـد و در صـدر آن‌هـا از امکانـات واژه‌سـازی از افعـال مرکـب، کـه در سـال‌های اخیـر به‌عنـوان عامـل عقیـم و غیرزایـا بـودن زبـان فارسـی مـورد حملـهٔ بعضـی از محققـان قـرار گرفتـه اسـت، گفت‌وگـو می‌کنـد. وی همچنیـن، طبقه‌بنـدی از انـواع پسـوندهای اسم‌سـاز و صفت‌سـاز را ارائـه می‌دهـد (۱۳۷۰-۱۳۷۲). «کشـانی» ابتـدا بـه تعریـف پسـوند و نقـش آن در زبـان فارسـی می‌پـردازد، و سـپس، پسـوند و پسـوندواره‌ها و انـواع آن‌هـا را در زبـان فارسـی بـه تفصیـل شـرح می‌دهـد (نوعـاً پسـوندهای اشـتقاقی کـه در سـاختن انـواع اسـم، انـواع صفـت و قیـد بـه‌کار می‌رونـد بـه دقـت بررسـی می‌شـود) و در نهایـت، حرکـت پسـوندها را مـورد مطالعـه قـرار می‌دهـد (۱۳۷۱). «عاصـی» موضوعـات اصلـی در پـردازش زبـان طبیعـی را معرفـی می‌کنـد کـه مشـتمل اسـت بـر حـوزهٔ خـط (کـه خـود شـامل غلط‌یابـی املایـی و بازشناسـی خـودکار متـن[1] اسـت) و حـوزهٔ زبـان (کـه خـود مسـتلزم بهره‌گیـری از حوزه‌هـای آواشناسـی، سـاخت‌واژه، نحـو و معناشناسـی اسـت) (۱۳۸۳). «لازار»[2] بـا رویکـردی زبانشـناختی و نگاهـی موشـکافانه، نکته‌سـنج و بی‌طرفانـه و بـا پرداختـن بـه مسـائلی کـه کمتـر بـه آن‌هـا توجـه شـده اسـت، خصوصیـات آوائـی، صرفـی و نحـوی و نیـز چگونگـی شـکل‌گیری واژه‌هـا را در زبـان فارسـی معاصـر تبییـن می‌کنـد و گاه نیـز بـه کشـف ویژگی‌هایـی در عناصـر سـازندهٔ زبـان و روابـط میـان آن‌هـا دسـت می‌یابـد

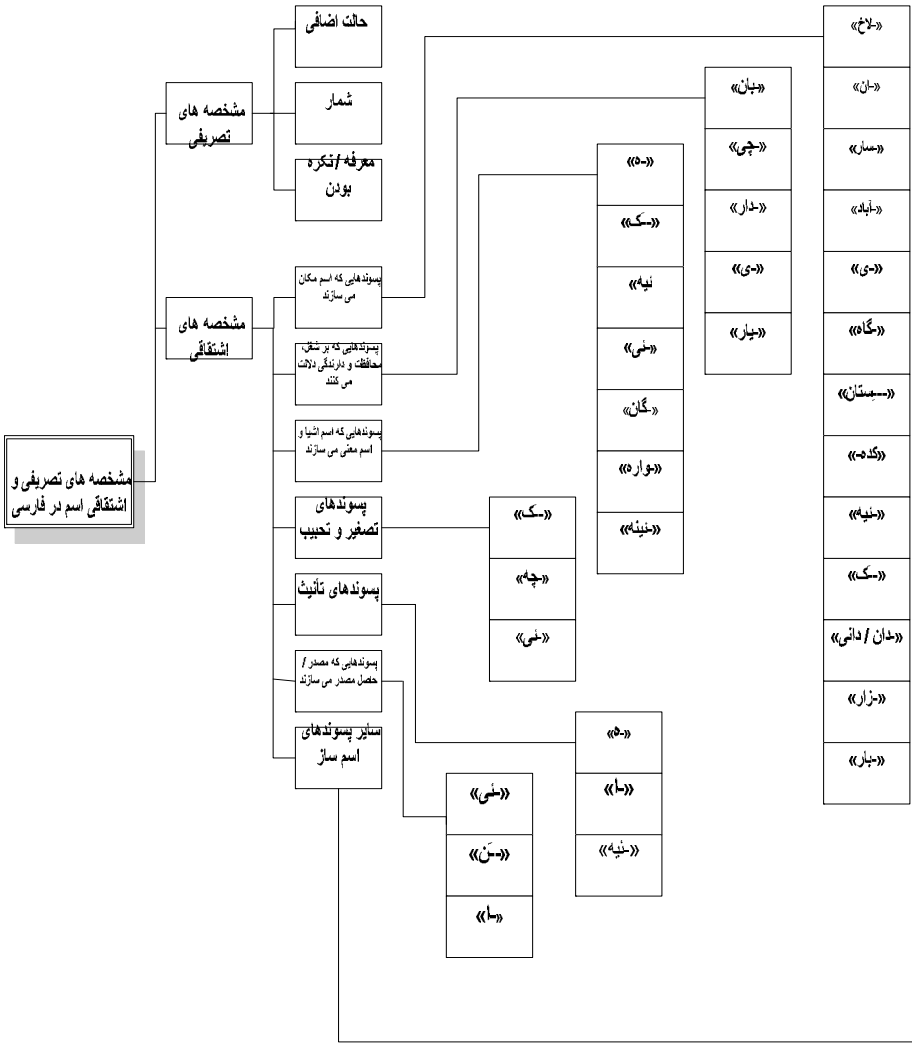
1. Optical Character Recognition (OCR)
2. Lazar

(۱۹۵۷) کـه طـرح آن‌هـا در نوشـته‌های دستورنویسـان فارسـی‌زبان پیـش از او سـابقه نداشـته اسـت. نمونه‌هایـی کـه از زبـان فارسـی معاصـر در ایـن دسـتور زبـان ارائـه شـده، بـر اسـاس آثـار نویسـندگانی چـون «صـادق هدایـت»، «صـادق چوبـک»، «علـی دشـتی»، «جمالـزاده» و دیگـران و نیـز نوشـته‌های روزنامه‌هـا و مجـلات و سـرانجام شـنیده‌های مؤلـف هنـگام اقامتـش در ایـران بـوده اسـت.

بررسـی پژوهش‌هـای انجام‌شـده در زمینـهٔ برچسـب‌گذاری اجـزای کلام نشـان می‌دهـد کـه مسـئلهٔ برچسـب‌دهی در پـردازش متـون حائـز اهمیـت اسـت. برخـی از پژوهش‌هـای اشاره‌شـده در بخـش ۲-۱.، اسـتفاده از ابزارهایـی ماننـد «مـدل مارکوفـی پنهـان»، تلفیقـی از تجزیـه و تحلیلـی سـاخت‌واژی و «مـدل مارکوفـی پنهـان» و شـبکه‌های پرسـپترون چندلایـه تکرارشـونده را پیشـنهاد می‌کننـد و برخـی دیگـر روش‌هایـی را معرفـی می‌کننـد کـه مسـتلزم بهره‌گرفتـن از ابزارهـای تحلیل‌کنندهٔ سـاخت‌واژی و نحـوی اسـت. در ایـن راسـتا و به‌منظـور ارائـهٔ بهینـهٔ چنیـن سیسـتم‌های برچسـب‌دهی بـه اجـزای کلام، ناگزیریـم بـه بررسـی سـاخت‌واژی اجـزای کلام در فارسـی بپردازیـم. بنابرایـن، در بخـش ۲-۲.، نمونه‌هایـی از مطالعـات انجام‌شـده در ایـن زمینـه ارائـه شـده اسـت.

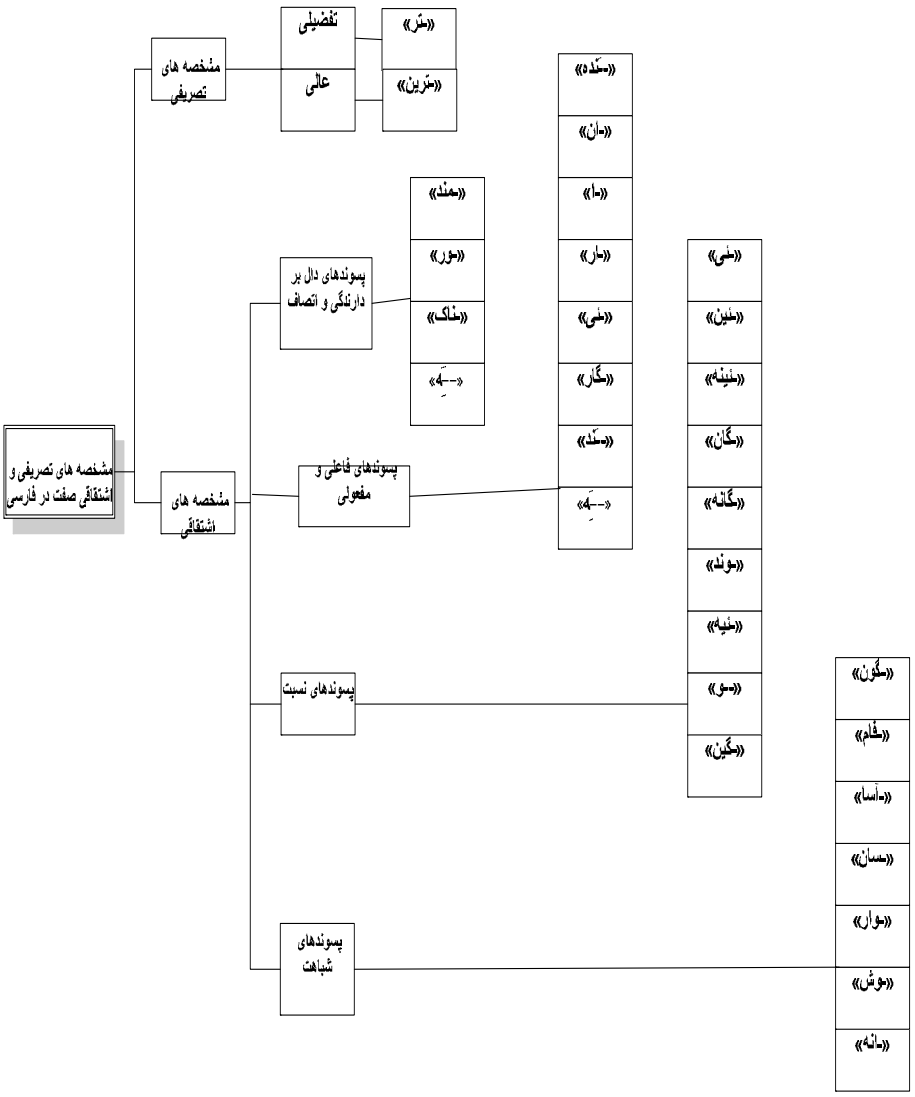
۲-۳. بررسی ساخت‌واژی اسم‌ها و صفت‌ها در فارسی

زبـان فارسـی دارای شـمار نسـبتاً زیـادی از پسـوندها بـرای اشـتقاق در مقولـهٔ اسـم اسـت. اسـم در فارسـی بـا نشـانه‌های تصریفـی خاصـی می‌توانـد همـراه شـود. در زبـان فارسـی معاصـر، بسـیاری از ایـن پسـوندها چنـدان زایـا نیسـتند. شـکل ۱، حاصـل جمع‌بنـدی نشـانه‌های تصریفـی و اشـتقاقی مربـوط بـه اسـم و مسـتخرج از مجموعـهٔ ۱۲ مقالـهٔ «دکتـر علـی اشـرف صادقـی» تحـت عنـوان «*شـیوه‌ها و امکانـات واژه‌سـازی در زبـان فارسـی معاصـر ۱ تـا ۱۲: ۱۳۷۰-۱۳۷۲*» (در بحـث اشـتقاق) اسـت؛ همچنیـن در بخـش تصریـف از کتـاب *دسـتور زبـان فارسـی معاصـر* «لازار» (۱۹۵۷) و مقالـهٔ *مشـخصه‌های تصریفـی در زبـان فارسـی امـروز* «قطـره» (۱۳۸۶) اسـتفاده شـده اسـت. ایـن فهرسـت تنهـا پسـوندهایی را عرضـه مـی‌دارد کـه مهم‌تـر و زایاتـر شـمرده می‌شـوند؛ اگرچـه برخـی از پسـوندهایی کـه در فارسـی معاصـر چنـدان زایـا نیسـتند نیـز بررسـی شـده اسـت.



شکل ۱. مشخصه‌های تصریفی و اشتقاقی اسم در زبان فارسی

برای تشخیص کلی صفت‌ها و به‌منظور بررسی امکان هم‌نگاره بودن با اسم‌ها، ساخت‌واژهٔ 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

۳. بررسی هم‌نگاره‌ها در فارسی

کلماتی هم‌نگاره خوانده می‌شوند که صورت نوشتاری یکسان، اما معنا (و گاهی تلفظ) متفاوتی دارند. در واقع، این کلمات املای یکسان دارند و می‌توان واژهٔ هم‌املا را

نیز در مورد آن‌ها به‌کار 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، به‌دلیل ساختار نوشتاری خاص آن، تعداد هم‌نگاره‌ها نسبت به دیگر زبان‌ها بیشتر است. تعداد زیادی از هم‌نگاره‌ها از ساختار زبان ناشی می‌شوند، به‌عبارت دیگر، هم‌نگاره‌ها در هر زبان به آن زبان خاص وابسته‌اند و هرچه ساخت‌واژهٔ 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 که 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 از جمله هم‌نگاره‌های اسمی در زبان فارسی مورد مطالعه قرار خواهد گرفت.

۳-۱. 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 به‌دست می‌آید، کامل نیست، زیرا واکه‌های کوتاه معمولاً نمود نوشتاری ندارند (Sproat 2000, 163، نقل در علایی و بی‌جن‌خان ۱۳۹۲). تنها تعداد انگشت‌شماری نگاره وجود دارد که برای نشان‌دادن واکه‌های کوتاه در خط به‌کار می‌روند؛ از جمله نگارهٔ «ه»، که برای نشان‌دادن واکهٔ کوتاه /e/ یا /a/ به‌کار می‌رود؛ مانند «به» و «نه»؛ و حرف «و» که می‌تواند نمود نوشتاری واکهٔ کوتاه /o/ باشد؛ مانند «تو». فقدان واکه‌های کوتاه در نظام نوشتاری فارسی باعث ایجاد ابهام می‌شود. به‌عنوان مثال، هم‌نگاره‌هایی مانند «مرد» /mard/ و «مرد» /mord/ در اثر فقدان واکه‌های کوتاه در نظام نوشتاری فارسی ایجاد شده‌اند (علایی و بی‌جن‌خان ۱۳۹۲). همچنین، استفاده‌نکردن 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.

۳-۲. 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 در محسنی ۱۳۸۷). هم‌نگاره‌ها در زبان فارسی می‌توانند واژگانی باشند. منظور از هم‌نگاره‌های واژگانی، هم‌نگاره‌هایی هستند که در فرهنگ‌های لغات به‌صورت مدخل‌های مجزا ذکر شده‌اند و حضورشان وابسته به بافت نحوی و اضافه‌شدن 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:

۳-۲-۱. هم‌نگاره‌هایی که در اثر عدم نمایش علائم زیروزبری در خط فارسی ایجاد شده‌اند؛ مانند: «مرد» /mard/ و «مرد» /mord/، «کند» /kanad/، «کند» /kond/ و «کند» /konad/.

۳-۲-۲. هم‌نگاره‌هایی که در اثر عدم تناظر یک-به-یک میان واج‌ها و نگاره‌ها در فارسی ایجاد شده‌اند؛ مانند: «رود» /rud/ و «رود» /ravad/، «شوم» / um□/ و «شوم»/ avam□/.

۳-۲-۳. هم‌نگاره‌هایی که در اثر یکسانی تظاهر واجی و نوشتاری تکواژها به‌وجود 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 در محسنی ۱۳۸۷):

۳-۲-۴. یکسان‌بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنا یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم‌شخص مفرد، یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یای متصل به گروه اسمی؛ مانند:

الف. یای نکره: استادی که به‌موقع سر کلاس حاضر می‌شود.
ب. یای اسم‌ساز: وی در سال ۱۳۸۸به درجهٔ استادی رسید.
پ. یای شناسهٔ دوم‌شخص مفرد: تو دیگر خود استادی.
ت. یای نسبت: مشکلات جوانی.
ث. یای متصل به گروه اسمی: کارگر ماهری که ما می‌شناسیم، ... .

۳-۲-۵. یکسانی تکواژهای ضمیر متصل سوم‌شخص مفرد با تکواژ اسم‌ساز؛ مانند:
الف. ضمیر متصل سوم‌شخص مفرد: به رویش خندیدم. /□ruja`/
ب. تکواژ اسم‌ساز: رویش گل‌ها را ببین. /ru`je□/

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 می‌شود، با تکواژهای دیگر هم‌نگاره هستند؛ به‌عنوان 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 را که هم‌نگارگی را در سطح تکواژ نشان می‌دهد، میتوان به‌صورت زیر دسته‌بندی کرد:

◇ پسوند «-ین» که هم شمار را در اسم‌ها نشان می‌دهد و هم می‌تواند در صفت‌ها، دال بر نسبت باشد.
◇ پسوند «-ه/ -ه» که می‌تواند نشانهٔ معرفه در گفتار، سازندهٔ اسم اشیا و اسم معنا، تأنیث در اسم‌ها، دال بر دارندگی و اتصاف در صفت‌ها باشد یا صفت فاعلی و مفعولی بسازد.
◇ پسوند «-ِش» که می‌تواند هم نشانهٔ 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 تحبیب را در اسم‌ها نشان می‌دهد و یا اسم اشیا و معنا 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 معنا 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 معنا (و گاهی، تلفظ) متفاوتی دارند.

٤. روش پژوهش

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

در مورد برخی پیکره‌ها که فایل آن‌ها موجود است، باید از روش جست‌وجوی ماشینی جهت تهیهٔ 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 جست‌وجوی ماشینی هم‌نگاره‌ها در پیکرهٔ وابستگی نحوی زبان فارسی و پیکرهٔ متنی زبان فارسی، که فایل آن‌ها موجود است، به‌صورت زیر توضیح داده می‌شود:

جست‌وجو برای هم‌نگاره‌ها با یک دستور در زبان برنامه‌نویسی «پایتون»[1] انجام شده است. به این صورت که کلمات موجود در پیکره، مانند پیکرهٔ متنی زبان فارسی، و برچسب‌هایی را که به آن کلمات تعلق گرفته است، بررسی می‌کند و کلماتی را که دارای بیش از یک برچسب باشند – 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 مجموعهٔ M موجود است یا نه. اگر موجود نباشد، کلمه را به همراه برچسب آن به مجموعهٔ M اضافه می‌کند، اگر موجود باشد، بررسی می‌کند که آیا برچسب کلمه‌ای که الآن مشاهده کرده است، مشابه همان برچسب قبلی است؟ اگر بود، یک واحد به فراوانی برچسب قبلی اضافه می‌کند و اگر برچسب جدید

1. Python

بود، آن را در کنار برچسب قبلی قرار می‌دهد. عملکرد برنامه به‌زبان ساده -به‌عنوان مثال، برای پیکرهٔ «بی‌جن‌خان»[1]، که پیکرهٔ مبسوط‌تر آن، پیکرهٔ متنی زبان فارسی است- به‌صورت زیر است:

- ◇ هر سطر در پیکرهٔ متنی زبان فارسی شامل یک واژه و برچسب اجزای کلام[2] آن است.
- ◇ یک مجموعه به اسم دیکشنری[3] داریم که در واقع، یک ساختار داده در زبان برنامه‌نویسی «پایتون» است. ساختار آن بدین شکل است که هر واژه به‌علاوهٔ برچسب‌های (POS) آن در مجموعه ذخیره می‌شود. نوع دادهٔ دیکشنری در زبان برنامه‌نویسی «پایتون»، به‌صورت فهرستی از کلیدها (key) و ارزش‌ها (value) است. در دیکشنری‌های «پایتون»، ترتیبی وجود ندارد و با دو مفهوم کلید (key) و ارزش (value) سروکار داریم (oruji.org). در مورد پژوهش حاضر، واژه‌ها همان key و برچسب اجزای کلام مربوط به هر واژه value آن در نظر گرفته می‌شوند.
- ◇ برنامه هر سطر از پیکره را می‌خواند. اگر واژهٔ قیدشده در آن سطر قبلاً در دیکشنری وجود نداشت، آن را به‌همراه POS آن به دیکشنری اضافه می‌کند. اگر قبلاً در دیکشنری وجود داشت، بررسی می‌کند که آیا POS قیدشده در این سطر قبلاً در دیکشنری برای این واژه درج شده است یا نه. اگر درج نشده بود ،آن را به لیست POSهای این واژه در دیکشنری اضافه می‌کند.
- ◇ در نهایت، مجموعهٔ دیکشنری بررسی می‌شود و واژه‌هایی که بیش از یک POS برای آن‌ها ثبت شده است، به‌عنوان خروجی برنامه ارائه می‌شوند.

---

۱. پیکرهٔ متنی استاندارد زبان فارسی/ پیکرهٔ «بی‌جن‌خان» که در پژوهشکدهٔ پردازش هوشمند علائم تهیه شده است، مجموعه‌ای از متون نوشتاری و گفتاری زبان فارسی به‌صورت 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 (Bijankhan et al. 2011).

2. Part Of Speech (POS)

3. dictionary

از پایگاه دادگان زبان فارسی[1] در تحقیق حاضر به این صورت استفاده شده است که فهرستی از کلمات پربسامد در پیکره در اختیار نگارنده قرار گرفت و از آنجا که فهرست مذکور فاقد هرگونه برچسبی است، نگارنده، خود به‌صورت دستی کلماتی را که می‌توانند بیش از یک برچسب داشته باشند، برچسب‌گذاری کرده است. برای این منظور، نگارنده هر کلمهٔ موجود در فهرست مذکور را در جملات فرضی، به‌عنوان بافت، در نظر گرفته است تا امکانِ داشتن بیش از یک برچسب برای هر کلمه را با توجه به بافت بررسی کند. سپس، تنها کلماتی را که در بافت بیش از یک برچسب می‌توانسته‌اند داشته باشند، برچسب‌گذاری کرده است و هر تعداد برچسبِ ممکن برای هر یک از کلمات مذکور را کنار آن‌ها مشخص کرده است. وی برای کنترل و تأیید صحت روایی و درستی تشخیص برچسب‌ها، از یک متخصص زبان و ادبیات فارسی کمک گرفته است.

جمع‌آوری داده‌ها در این مرحله دو کاربرد مهم داشت: ۱. نشان‌دهندهٔ تعداد بسیار زیاد هم‌نگاره‌ها در پیکره‌های متنی فارسی که خود نشان می‌دهد در صورت رفع ابهام نکردن از هم‌نگاره‌ها در متون گوناگون، پردازش متن با مسائل و مشکلاتی روبه‌رو خواهد شد. ۲. فهرست تهیه‌شده از هم‌نگاره‌ها در پیکره‌های گوناگون می‌تواند در پژوهش‌های دیگر مورد استفاده قرار گیرد. به این صورت که می‌توان هم‌نگاره‌هایی را که فراوانی بالاتری دارند، در بافت نحوی مورد مطالعه قرار داد تا با توجه به ساخت نحوی که در آن به‌کار رفته‌اند، برچسب‌های درست به آن‌ها اختصاص داده شود.

قسمتی از فهرست تهیه‌شده در جدول ۱، آورده شده است:

---

۱. پایگاه دادگان زبان فارسی، مجموعه‌ای نرم‌افزاری برای ذخیره، پردازش و ارائهٔ داده‌های زبانی فارسی است. این پایگاه دربرگیرندهٔ پیکره‌های گوناگونی از زبان فارسی است که با وجود حجم عظیم و با گستردگی و گوناگونی‌های بسیار، دارای ساختاری بسامان و منطقی است و امکان هرگونه جست‌وجو و دستیابی سریع به آگاهی‌های مورد نیاز را در هر زمان فراهم آورده است. پیکره‌های این پایگاه می‌توانند همواره روزآمد شوند. پایگاه دادگان زبان فارسی فراگیر و متنوع است. در واقع، فراتر از یک یا چند پیکرهٔ خاص است و کاربران بر پایهٔ نیاز و هدف پژوهشی خود می‌توانند پیکرهٔ مناسب را از آن برگزینند.

جدول ۱. قسمتی از فهرست هم‌نگاره‌های تهیه‌شده از پیکره‌های متنی فارسی

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

راهنمای برچسب‌ها:

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

۵. 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، در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جست‌وجو و کمک به مدل‌های آماری اقدامی کاربردی است. با توجه به تعداد بسیار زیاد اسم‌ها در زبان فارسی و هم‌نگاره‌بودن

بسیاری از آن‌ها با صفت‌ها، به‌راحتی نمی‌توان اسم‌ها را در یک پیکرهٔ متنی فارسی برچسب‌گذاری کرد. در تهیهٔ 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 پیکرهٔ آموزشی ظاهر نشده باشد، نمی‌توان از پیکرهٔ آموزشی اطلاعات دقیقی راجع به آن کلمه به‌دست آورد و حتی از توزیع کلمات در پیکره نیز نمی‌توان استفاده کرد، زیرا توزیع کلمات ناشناخته کاملاً متفاوت از کلمات شناخته‌شده در متن است. بنابراین، اگر در یک پیکرهٔ متنی فارسی، کلمه‌ای در واژگان حضور نداشته باشد (کلمهٔ خارج از واژگان)، نمی‌توان برچسب‌هایی را که کلمه به آن‌ها منتسب می‌شود، 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 است که به کلمات متصل می‌شوند. بنابراین، در وهلهٔ اول، بررسی ساخت‌واژی اسم‌ها و صفت‌ها، فارغ از بافت، ضروری به نظر می‌رسد تا از این رهگذر بتوان اسم‌ها و صفت‌ها را در صورت دارابودن شاخص‌های صرفی و اشتقاقی شناسایی و سپس برچسب‌دهی کرد. ۲. زبان فارسی ظرفیت بالایی برای ساخت هم‌نگاره‌های جدید را که از ساخت‌واژهٔ فارسی نشأت می‌گیرند نیز دارد. بسیاری از وندها، اشتقاقی و تصریفی، نمود نوشتاری مشابهی دارند و در اتصال به ستاک، هم‌نگاره‌های گوناگون می‌سازند. به‌عنوان 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 ساخت‌واژی اسم‌ها و صفت‌ها، به‌منظور 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، پس از مقدمه، ابتدا به پیشینهٔ تحقیق پرداخته شد. سپس، ساخت‌واژهٔ 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 برچسب‌دهی کرد. بخش چهارم، مطالعهٔ هم‌نگاره‌ها 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 استفاده‌شده برای این منظور، شامل پیکرهٔ متنی زبان فارسی، پایگاه دادگان زبان فارسی و پیکرهٔ وابستگی نحوی زبان فارسی است و جست‌وجو به دو روش جست‌وجوی ماشینی و دستی انجام گرفت و فهرست مبسوطی از هم‌نگاره‌های به‌دست‌آمده از پیکره‌های مذکور تهیه شد. نمونه‌ای از این فهرست در جدول ۱، آورده شده است. بررسی کلی هم‌نگاره‌ها 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 یکسان‌بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم‌معنا یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم‌شخص مفرد و یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یای متصل به گروه اسمی باشند. مانند:

الف. یای نکره: کشاورزی که محصول خود را به‌موقع درو می‌کند.

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

پ. یای شناسه دوم‌شخص مفرد: تو خود کشاورزی.

و یا در اثر یکسانی تکواژهای ضمیر متصل سوم‌شخص مفرد با تکواژ اسم‌ساز به‌وجود آمده‌اند؛ مانند:

الف.ضمیر متصل سوم‌شخص 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 که خود نشان می‌دهد در صورت رفع ابهام‌نکردن از هم‌نگاره‌ها در متون گوناگون، پردازش متن با مسائل و مشکلاتی روبه‌رو خواهد شد. ۲. فهرست تهیه‌شده از هم‌نگاره‌ها می‌تواند در 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 به‌کار رفته‌اند، برچسب‌های درست به آن‌ها تخصیص داده شود.

٦. 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 به‌دست آمد که تعداد زیادی از هم‌نگاره‌ها در اثر یکسان‌بودن نمود نوشتاری/ املای وندهای اشتقاقی و تصریفی، به‌ویژه 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 آنجا که در فارسی هم‌نگاره‌های دیگری نیز وجود دارد که در اثر عدم نمایش واکه‌های کوتاه در خط فارسی یا نبودِ 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

آخشیک، سمیه، و رحمت‌الله فتاحی. ۱۳۹۱. تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی». *فصلنامه کتابداری و اطلاع‌رسانی* ۱۵ (۳): ۹-۳۰.

صادقی، علی‌اشرف ۱۳۷۰. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱). *نشر دانش* ۶۴: ۱۲-۱۸.

_____. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۲)». *نشر دانش* ۶۵: ۶-۱۲.

_____. ۱۳۷۰. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۳)». *نشر دانش* ۶۷: ۲۸-۳۳.

_____. ۱۳۷۱. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۴). *نشر دانش* ۶۹: ۲۱-۲۵.

_____. ۱۳۷۱. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۵). *نشر دانش* ۷۰: ۳۹-۴۵.

_____. ۱۳۷۱. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۶). *نشر دانش*. ۷۱: ۱۵-۱۹.

_____. ۱۳۷۱. شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۷). *نشر دانش* ۷۲: ۱۹-۲۳.

_____. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۸) ». تهران: *نشر دانش*. شمارهٔ ۷۴: ۹۸-۱۰۵.

_____. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۹)». تهران: *نشر دانش*. شمارهٔ ۷۵: ۹-۱۵.

_____. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۰)». تهران: *نشر دانش*. شمارهٔ ۷۶: ۱۵-۲۳.

_____. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۱)». تهران: *نشر دانش*. شمارهٔ ۷۷: ۲۱-۲۵.

_____. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲)». تهران: *نشر دانش*. شمارهٔ ۷۹ و

۸۰: ۱۲-۱۵.

عاصی، مصطفی. ۱۳۸۳.پردازش دستوری زبان فارسی با رایانه. *ویژه‌نامه فرهنگستان (دستور)* ۱: ۲۹-۵۱.

علایی، الهام، و محمود بی‌جن‌خان. ۱۳۹۲. «عمق خط فارسی». پژوهش‌های زبانی. *مجله سابق دانشکده ادبیات و علوم انسانی دانشگاه تهران* ۴ (۱): ۱-۱۹.

قطره، فریبا. ۱۳۸۶. مشخصه‌های تصریفی در زبان فارسی امروز. *دستور* ۳: ۵۲-۸۱.

کشانی، خسرو. ۱۳۷۱. *اشتقاق پسوندی در زبان فارسی امروز*. تهران: مرکز نشر دانشگاهی.

لازار، ژ. ۱۹۵۷. *دستور زبان فارسی معاصر*. ترجمهٔ مهستی بحرینی و توضیحات و حواشی هرمز میلانیان. ۱۳۸۹. تهران: انتشارات هرمس.

محسنی، مهدی. ۱۳۸۷. *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*. تهران: دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر.

_____، و بهروز مینایی بیدگلی. ۱۳۸۸. سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی. دو فصلنامه *پردازش علائم و داده‌ها* ۲. پیاپی ۱۲: ۱۳-۲۶.

مرادی، مهدی و بهرام وزیرنژاد. ۱۳۹۱. ساخت پیکره‌های نشانه‌گذاری‌شده با رویکرد وب به‌عنوان پیکره. *مجموعه‌مقالات دومین هم‌اندیشی زبان‌شناسی رایانشی*. تهران: انجمن زبان‌شناسی ایران.

مسعودی، بابک، سعید قوچانی، و اعظم استاجی. 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.

References

Assi, M. and M. Haji Abdolhosseini. 2000. Grammatical tagging of a Persian corpu". *International journal of corpus linguistics* 5 69-81 :)1(.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, & M. Ghayoomi. 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 text". *Journal of computer science* 5 (11): 794-800.

Frodl, M. 2013. *Part of speech tagging using Neural Networks*. Czech Republic: 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 corpor. *Proceedings of the 17th annual conference of the University of Waterloo Centre for the new OED and text research.* Oxford.

Hu, Y., I. Matveeva, J. Goldsmith, and C. Sprague. 2005. Using morphology and syntax together in unsupervised learnin". In *Proceeding of the workshop on Psychocomputational Models of Human Language Acquisition*. PMHLA 05. 20-27. Stroudsburg: PA, USA. Association for computational linguistics.

Megerdoomian, K. 2000. Unification-based Persian morphology. *In Proceedings of CICLing, Centro de investigacion en computacion- IPN*, Mexico311-318 :.

_____. 2004. Developing a Persian part-of-speech tagge". *Proceedings of the 1st workshop on Persian*

language and computer. 99-105. University of California. San Diego.

Merialdo, B. 1994. Tagging English text with a probabilistic mode". *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

الهام علایی ابوذر

دانش‌آموختهٔ دکتری تخصصی در رشتهٔ 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، به‌عنوان عضو هیئت علمی آغاز نموده است و هم‌اکنون استادیار پژوهشی این پژوهشگاه است. بیشتر فعالیت‌های وی تاکنون در حوزهٔ زبان‌شناسی رایانه‌ای (سیستم‌های تبدیل متن به گفتار، نظام‌های نوشتاری، یادگیری ماشینی و پردازش زبانی)، زبان‌شناسی نظری و پیکره‌ای بوده است و در حوزه‌های ذکرشده با پژوهشگاه علوم و فناوری اطلاعات ایران همکاری می‌نماید.