بررسی مقایسه‌ای تأثیر برچسب‌زنی مقوله‌های دستوری بر تجزیه در پردازش خودکار زبان فارسی

مسعود قیومی

گروه دستور زبان آلمانی، دانشگاه آزاد برلین، برلین، آلمان

پژوهشکده زبانشناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

چکیده

در این مقاله، به بررسی جایگاه برچسب‌زنی مقوله‌های دستوری[1] در تجزیهٔ نحوی خودکار[2] جملات فارسی پرداخته خواهد شد. به‌همین منظور، تأثیر کیفیت برچسب‌زنی مقوله‌های دستوری و همچنین تأثیرگذاری میزان اطلاعات موجود در مقوله‌های دستوری بر کارآیی[3] تجزیهٔ خودکار جملات مورد مطالعه قرار خواهد گرفت. به‌منظور انجام این دو بررسی، سه سناریو برای تجزیهٔ جملات، ارائه شده و مقایسه می‌شود. در سناریوی نخست، تجزیه‌گر ابتدا دادهٔ ورودی را برچسب‌زنی کرده و سپس جمله را تجزیه می‌کند. در سناریوی دوم، از یک برچسب‌زن خارج از تجزیه‌گر و در سناریوی سوم از برچسب معیار واژه‌ها برای تجزیهٔ جملات استفاده می‌شود. در این بررسی، معیارهای ارزیابی متفاوت مورد استفاده قرار می‌گیرد تا میزان این تأثیرگذاری از ابعاد مختلف نشان داده شود. نتایج حاصل از آزمایش‌ها نشان می‌دهد که کیفیت و میزان اطلاعات در مقولات دستوری واژه، بر کارآیی تجزیه‌گر تأثیر مستقیم دارد. کیفیت بالای برچسب مقوله‌های دستوری، سبب کاهش خطای تجزیه‌گر و افزایش کارآیی آن می‌شود؛ همچنین عدم وجود اطلاعات صرفی‌ـنحوی[4] تأثیر منفی به‌سزایی بر کارآیی تجزیه‌گر دارد که این تأثیرگذاری در مقایسه با کیفیت برچسب مقولات دستوری بسیار بیشتر است.

واژگان کلیدی: پردازش زبان طبیعی، زبان فارسی، برچسب‌زنی مقوله‌های دستوری، تجزیهٔ نحوی خودکار، دانگی اطلاعات زبانشناختی

A Comparative Study on the Impact of Part-of-Speech Tagging on Parsing in Processing the Persian Language

Masoud Ghayoumi

German Grammar Group, Freie Universität Berlin, 14195 Berlin, Germany

Linguistics Department, Institute for Humanities and Cultural Studies, 14377, Tehran, Iran

Abstract:

In this paper, the role of Part-of-Speech (POS) tagging for parsing in automatic processing of the Persian language is studied. To this end, the impact of the quality of POS tagging as well as the impact of the quantity of information available in the POS tags on parsing are studied. To reach the goals, three parsing scenarios are proposed and compared. In the first scenario, the parser assigns the POS tags firstly and then it parses the input sentence. In the second scenario, an external POS tagger is usedto assign the tags, then the sentence is parsed. In the third scenario, the parser uses the gold standard POS tags to parse the input sentence. In this study, various evaluation metrics are used to show the impacts from different points of views. The experimental results show that the quality of the POS tagger and the quantity of the information available in the POS tags have a direct effect on the parsing performance. The high quality of the POS tags causes error reduction in parsing and also it increases parsing performance. Moreover, lack ofmorphological -syntactic information in the POS tags has a high negative impact on parsing performance. This impact is more pronounced than the impact of POS tagger performance.

Keywords: **natural language processing, the Persian language, part-of-speech tagging, parsing, granularity of linguistic information**

* نویسندهٔ عهده‌دار مکاتبات

۱- مقدمه

درک زبان طبیعی، یکی[1] از مهم‌ترین بخش‌های پردازش زبان طبیعی است؛ چون در آن سعی می‌شود با تحلیل عمیق‌تر جملات به مفهوم دست یافته[2] شود. درک زبان طبیعی در سه سطح نحوی، معنایی و کاربردشناسی اتفاق می‌افتد. برای رسیدن به مفهوم، نخستین گام تحلیل جملات در سطح[3] نحو است تا رابطهٔ نحوی کلمات با یکدیگر در یک جمله مشخص[4] شود. این سطح تحلیل شامل دو بخش است: بخش نخست مربوط به مشخص‌شدن نقش دستوری واژه در جمله است. در این بخش، مقولهٔ دستوری و ویژگی‌های معنایی واژه مشخص می‌شود. بخش دوم به نحوهٔ تعامل واژه‌ها با هم و ساخت یک جمله می‌پردازد. در کاربرد عادی یک سامانه، به این دو سطح تحلیل نیاز است تا زمینه برای درک عمیق‌تر جمله فراهم شود.

همان‌طورکه مشخص است، دو بخش تحلیل در سطح نحو رابطهٔ تنگاتنگی با هم دارند و بخش دوم تأثیرپذیر از بخش نخست است. با نگاهی به مطالعات انجام‌شده بر روی زبان فارسی درمی‌یابیم که هریک از این دو سطح به‌طور مستقل بررسی شده است. در این مقاله تلاش می‌کنیم، تعامل این دو بخش با یکدیگر و همچنین تأثیرگذاری مقولات دستوری را از دو جنبهٔ کیفیت و میزان اطلاعات بر تجزیهٔ خودکار جملات بررسی کنیم. در انجام این بررسی، معیارهای ارزیابی متفاوتی را استفاده خواهیم کرد تا ابعاد تأثیرگذاری بهتر نمایان شود.

این مقاله در شش بخش نگارش شده است. بخش دو و سه به بیان کلیات و پیشینهٔ مطالعاتی برچسب‌زنی مقولات دستوری و تجزیهٔ نحوی می‌پردازد. در بخش چهار، معیارهای ارزیابی مورد استفاده در این مقاله معرفی می‌شود. ابزار و دادهٔ مورد نیاز این پژوهش و همچنین نتایج به‌دست‌آمده از آزمایش‌ها در بخش پنجم گزارش و توضیح داده می‌شود. در بخش ششم، خلاصه مقاله و نتیجه‌گیری حاصل از آزمایش‌ها بحث می‌شود.

۲- برچسب‌دهی مقولات دستوری

۲-۱- کلیات

برچسب‌زنی مقولات دستوری نخستین مرحلهٔ پردازش نحوی جملات است که در آن نقش نحوی واژه مشخص می‌شود. برای برچسب‌زنی خودکار می‌توان از برچسب‌زن قاعده‌بنیان[5]، آماری، و یا ترکیبی[6] استفاده کرد.

دو چالش در برچسب‌زنی خودکار واژه وجود دارد که برچسب‌زن می‌بایستی از عهدهٔ آن برآید. یکی از چالش‌ها، انتخاب بهترین برچسب برای واژهٔ مبهم است که ممکن است آن واژه بیش از یک برچسب دستوری داشته باشد. چالش دیگر، پیشنهاد برچسب برای واژه‌های ناشناخته‌ای است که درقبل در دادهٔ آموزش دیده نشده است. این چالش بیشتر برای برچسب‌زن‌های آماری مطرح است. برای چیره‌شدن به هریک از این چالش‌ها روش‌هایی وجود دارد که خارج از بحث در این مقاله است؛ ولی تنها به این نکته می‌توان بسنده کرد که بافت محلی واژه به حدس یا رفع ابهام مقولهٔ دستوری واژه کمک شایانی می‌کند.

۲-۲- پیشینهٔ مطالعاتی برچسب‌زنی خودکار مقوله‌های دستوری در زبان فارسی

تا جایی‌که می‌دانیم، نخستین بررسی برای برچسب‌زنی واژه در زبان فارسی به‌صورت نیمه‌خودکار توسط عاصی و حاجی‌عبدالحسینی [9] در چارچوب کار پژوهشی معرفی‌شده توسط شوتز [37] طراحی و پیاده‌سازی شده است. آنها ۴۳ برچسب معرفی کرده و از دادهٔ پایگاه دادهٔ زبان فارسی [4] استفاده کرده‌اند.

مطالعات متنوعی در استفاده از روش‌های آماری برای برچسب‌زنی واژه‌های فارسی انجام شده است. متأسفانه مطالعهٔ جامعی که مدل‌های ارائه‌شده را با هم مقایسه کند تا انتخاب بهترین مدل برای برچسب‌زنی واژها در فارسی را میسّر سازد، وجود ندارد؛ بنابراین هریک از بررسی‌ها به‌خودی‌خود معتبر بوده که به‌اختصار به این مطالعات می‌پردازیم.

مدل احتمالات بیشینه[7] برای برچسب‌دهی مقولات دستوری استفاده شده است [8]،[29] که برای ساخت این مدل از پیکرهٔ بی‌جن‌خان[8] [1] که دارای ۵۸۶ برچسب است، به‌عنوان دادهٔ آموزش استفاده شده است. برچسب‌زن تی‌اِن‌تی [12] یک برچسب‌زن آماری است که مستقل از

[1] Part-of-Speech Tagging
[2] Syntactic Parsing
[3] Performance
[4] Morpho-syntactic

[5] Rule-based
[6] Hybrid
[7] Maximum Likelihood Estimation
[8] http://ece.ut.ac.ir/dbrg/bijankhan/

زبان بوده و با پیکرهٔ بی‌جن‌خان آموزش دیده است [41]. در یک مدل برچسب‌زنی دیگر، از یک روش ترکیبی برای تحلیل صرف استفاده شده است [40]. تعداد برچسب‌های معرفی‌شده در این پژوهش ۲۵ برچسب است واز پیکرهٔ همشهری [8] در این پژوهش استفاده شده است. محسنی و مینایی‌بیدگلی [27] یک تحلیل‌گر صرفی برای برچسب‌زنی تهیه کرده‌اند؛ به‌صورتی‌که در این پژوهش، تعداد برچسب‌های پیکرهٔ بی‌جن‌خان از ۵۸۶ برچسب به ۱۰۵ برچسب کاهش یافته‌است. سراجی [38] نیز یک برچسب‌زن زبان مجارستانی را با پیکرهٔ بی‌جن‌خان آموزش داده و از آن برای برچسب‌زنی متن فارسی استفاده کرده است. دنیس و زگوت [14] از سامانهٔ Melt برای برچسب‌زنی استفاده کرده‌اند. آنها در پژوهش خود برچسب‌های موجود در پیکرهٔ بی‌جن‌خان را با توجه به نیاز تغییر داده‌اند.

۳- تجزیهٔ خودکار

۳-۱- کلیات

تجزیهٔ خودکار مرحلهٔ پیچیده‌ترِ پردازش نحویِ جملات است؛ چون در این سطحِ تحلیل، روابط واژه‌ها با یکدیگر برای ساخت یک جمله مشخص می‌شود. با توجه به کاربرد و نیاز، درجه‌ای از تجزیهٔ جمله می‌تواند ارائه شود. برای این‌منظور می‌توان از تجزیه‌گر قاعده‌بنیان، آماری، و یا ترکیبی استفاده کرد. دو دیدگاه و شیوهٔ به‌طورکامل متفاوت در ارائهٔ تجزیهٔ نحوی جملات مطرح است: یک دیدگاه مبتنی بر سازه (جاشناختی[1]) و دیدگاه دیگر مبتنی بر وابستگی (ساختار دستوری[2]) است.

در این مقاله، دیدگاه جاشناختی و تجزیهٔ ساخت‌سازه‌ای جملات مدنظر است. در پردازش و تجزیهٔ ساخت‌سازه‌ای جملات دو چالش وجود دارد. یکی از این دو چالش، رفع ابهام در اتصال سازه‌های کوچک‌تر برای ساخت سازه‌های بزرگ‌تر است که ابهام ساختاری جمله در این سطح مطرح است. چالش دیگر، رفع ابهام برای یافتن برچسب صحیحی است که بیان‌گر رابطهٔ واژه‌ها با هم و ساخت یک سازهٔ دستوری است. برای رفع ابهام این دو چالش، روش‌ها و الگوریتم‌های مختلفی، مانند تجزیهٔ بالابه‌پایین یا پایین‌به‌بالا [21]، وجود دارد که سعی می‌کند بهترین تجزیهٔ نحوی را بیابد و آن را به‌عنوان بهترین تجزیه ارائه کند. لازم به ذکر است که بحث و بررسی این روش‌ها از مقالهٔ حاضر خارج است.

۳-۲- پیشینهٔ مطالعاتی تجزیهٔ خودکار در زبان فارسی

تجزیه‌گرهای (مبتنی بر سازه یا وابستگی) را می‌توان به سه دستهٔ اصلی قاعده‌بنیان، آماری و ترکیبی تقسیم کرد. در تجزیه‌گرهای دستهٔ نخست مجموعه‌ای از قواعد زبانی در چارچوب یک نظریهٔ زبانی گردآوری می‌شود که جملات مربوط به آن قواعد را می‌تواند تجزیه کند؛ مانند تجزیه جملات با کمک تشخیص وابستگی‌ها [34]، تجزیه جملات در چارچوب دستور ارتباط[3] [13]، تجزیه جملات در چارچوب دستور تعاملی[4] [17]، تجزیه جملات در چارچوب دستور ساخت‌سازه‌ای هسته‌بنیان[5] [28]، و تجزیه جملات در چارچوب دستور ساخت‌سازه‌ای تعمیم‌یافته[6] [10]. علاوه‌بر این مطالعات، در پژوهشی دیگر شریفی آتشگاه [2] تلاش کرده است با استفاده از روش قاعده‌بنیان به‌طور نیمه‌خودکارجملات را تجزیه و دادگان درختی در چارچوب برنامهٔ کمینگی تهیه کند. نقطه‌ضعف این مطالعه این است که با وجود معرفی یک الگوریتم، تجزیه‌گر کاملی که بتواند تجزیه را در سطح جمله ارائه کند تهیه نشده و کار عملی این مطالعه در سطح تجزیهٔ گروه اسمی باقی مانده است.

تجزیه‌گرهای دستهٔ دوم مستقل از زبان بوده و با استفاده از مدل‌های آماری-احتمالاتی جملات را تجزیه می‌کند. دو روش متفاوت در آموزش این دسته از تجزیه‌گرها وجود دارد: (۱) روش بی‌مربی[7] که به دادهٔ آموزش اولیه نیازی ندارد. فِیلی [5] در رسالهٔ دکترای خود به ابداع روشی برای تهیهٔ یک تجزیه‌گر بی‌مربی آماری برای تشخیص ساخت‌های سازه‌ای پرداخته است. رسولی و فیلی [31] نیز سعی کرده‌اند به‌صورت بی‌مربی، وابستگی بین واژه‌های فارسی را تشخیص دهند تا تجزیهٔ جملات حاصل شود. (۲) روش بامربی که نیازمند دادهٔ آموزش اولیه است. در چند پژوهش برای آموزش تجزیه‌گر آماری بامربی سازه‌ای زبان فارسی از دادگان درختی تهیه‌شدهٔ زبان فارسی [15] استفاده شده‌است [6]، [35]. در پژوهش‌های دیگری، دادگان درختی وابستگی تهیه شده و از آن برای آموزش تجزیه‌گر آماری بامربی وابستگی استفاده شده است [18]،[32]،[39].

[1] Typological
[2] Tecto-grammatical

[3] Link Grammar
[4] Interaction Grammar
[5] Head-driven Phrase Structure Grammar
[6] Generalized Phrase Structured Grammar
[7] Unsupervised

بـه‌هنگام آمـوزش تجزیـه‌گرهای آمـاری می‌تـوان ویژگی‌های[1] به‌کاررفته در مدل دستوری را تغییر داد که ایـن تغییر بر کارآیی تجزیه‌گر تأثیرگذار خواهد بود. ازجملـهٔ ایـن ویژگی‌هـا، دانـه‌ریزی و دانه‌درشـتی ویژگی‌هـای صـرفی و واژگـانی اسـت کـه در چنـدین مطالعـه تـأثیر دانـه‌ریزی یـا دانه‌درشتی واژه بـر کـارآیی تجزیـه‌گر سـازه‌ای در زبان‌هـای فارسـی و بلغـاری بررسـی شـده اسـت [16]،[19]،[20]. در مطالعهٔ دیگر تأثیر ویژگی‌های صـرفی، معنـایی، دانـه‌ریزی و دانه‌درشتی آنها بر تجزیه‌گر وابسـتگی فارسـی بررسـی شـده است [22].همچنین، طول جمله نیز از دیگر عوامل تأثیرگذار بر کارآیی تجزیه‌گر است که مورد بررسی قـرار گرفتـه اسـت [3].

۴- معیارهای ارزیابی

بـرای ارزیـابی کـارآیی برچسـب‌زن و تجزیـه‌گر معیارهـای متفاوتی معرفی شده است. ارزیابی دقـت برچسـب‌زنی سـاده است و به‌طورمعمول از معیار دقت[2] اسـتفاده می‌شـود. ایـن معیار براساس نسبت تعداد کلماتی که صحیح برچسب‌گذاری شـده بـه تعـداد کـل کلمـات برچسب‌گذاری‌شـده محاسـبه می‌شود.

معیارهای ارزیابی تجزیـه‌گر سـازه‌ای تنـوع بیشـتری دارد؛ که در اینجا به چند مورد آن اشاره خواهیم کـرد تـا بـا درنظرگرفتن معیار ارزیابی تجزیه‌گر، تأثیر برچسـب مقولـات دسـتوری بـر تجزیـه‌گر مشـخص شـود. ازجملـه معیارهـای متـداول بـرای ارزیـابی کـارآیی تجزیـه‌گر آمـاری صـحت[3] (رابطهٔ ۱)، پوشش[4] (رابطهٔ ۲)، و امتیاز تأثیر ایـن دو معیـار[5] (رابطهٔ ۳) را می‌توان نام برد.[6]

$$P = \frac{num.\ of\ correct\ constituents}{num.\ of\ cons.s\ in\ parser\ output} \quad (۱)$$

$$R = \frac{num.\ of\ correct\ constituents}{num.\ of\ cons.s\ in\ gold\ standard} \quad (۲)$$

$$f_1 = \frac{2\times P \times R}{P+R} \quad (۳)$$

[1] Feature
[2] Accuracy
[3] Precision
[4] Recall
[5] F-measure
[6] نویسندهٔ مقاله پیش‌تر در [۶] از معادل فارسی «دقت» برای «recall» استفاده کرده‌است. برای رفع ابهام این مفهوم در مقایسه با معادل فارسی «دقت» برای «accuracy»، معادل فارسی «پوشش» برای «recall» پیشنهاد می‌شود.

البته علاوه‌بر ارزیابی اتصـال گره‌هـا در سـاخت‌سـازه، می‌تـوان برچسـب آن سـازه را نیـز دخیـل کـرد تـا ارزیـابی جـامع‌تری صـورت پـذیرد. بـر همـین اسـاس، «صـحت برچسب‌دار[7]» و «پوشش برچسب‌دار[8]» مطرح می‌شود.«تناظر همسان[9]» یک معیار قوی است که امتیاز یک را به تجزیـه‌ای کـه بـه‌طورکامل صـحیح باشـد، می‌دهـد و درغیراین‌صـورت امتیاز صفر می‌دهد.

مشکلی که در بررسی سمپسون [33] درمورد ارزیـابی صحت و پوشش عنوان می‌شود، این است که اگر اشتباهی در یک سازه روی دهد، می‌بایستی فقط بر امتیـاز آن واژه در آن سـازه و نـه بـر امتیـاز ارزیـابی تمـام جملـه تـأثیر گـذارد. برهمین‌اساس، وی معیاری را با عنوان «نیای برگ[10]» معرفی می‌کند. در این معیار، فاصلهٔ واژه تا ریشهٔ درخـت تجزیـه بـا توجه به درخت تجزیه معیار جمله محاسبه شده و معدل این امتیاز به‌عنوان امتیاز نیای برگ تلقی می‌شود. برای محاسـبهٔ این فاصله، از فاصلهٔ لونشتاین[11] [25] استفاده می‌شـود. ایـن امتیاز بیان‌گر میزان تلاش مـورد نیـاز بـرای تغییـر برچسـب سازه‌ای اشتباه به برچسب صحیح است.

نقطهٔ ضعف نیای برگ این است که تعداد قلاب‌هـا در تحلیل درختی نقش زیادی دارد. بر همین اساس، لـین [26] معیاری را معرفی می‌کند که در آن تحلیل سازه‌ای جمله بـه معادل تحلیل وابستگی آن جمله تبدیل می‌شود و وابسـتگی عناصر با هسته به‌جای ساختار سازه‌ای سلسله‌مراتبی ارزیابی می‌شود. در این نحوهٔ ارزیابی، با درنظرداشتن وجود برچسـب وابستگی[12] یا عدم وجود آن[13] بـرای تک‌تـک واژه‌هـا، روابـط وابستگی بـین واژه‌هـا ارزیـابی می‌شـود. در ارزیـابی کـارآیی تجزیه‌گر براساس روابط وابستگی، شـکل سـاده‌ای از تجزیـهٔ جمله مورد توجه قرار می‌گیرد.

۵- نتایج

۵-۱- مقدمه‌سازی برای انجام آزمایش‌ها

۵-۱-۱- ابزار

همان‌طورکه در مقدمـهٔ مقالـه ذکـر شـد، هـدف ایـن مقالـه بررسی تأثیرگـذاری برچسـب مقولـات دسـتوری بـر تجزیـهٔ

[7] Labeled Precision
[8] Labeled Recall
[9] Exact Matching
[10] Leaf Ancestor
[11] Levenshtein Distance
[12] Labeled Attachment
[13] Unlabeled Attachment

خودکار است. در این پژوهش برای انجام آزمایش‌ها از ابزارهای آماری مانند برچسب‌زن تی.اِن.تی[1] [12]، برچسب‌زن استنفورد[2] [42]، تجزیه‌گر استنفورد[3] [23]، تجزیه‌گر بِرْکْلی[4] [30]، و تجزیه‌گر بیتْپَر [36] استفاده می‌کنیم. تجزیه‌گر استنفورد یک تجزیه‌گر آماری واژگانی است؛ ولی تجزیه‌گرهای بِرْکْلی و بیتْپَر آماری غیرواژگانی هستند. هدف استفاده از ابزار متفاوت، بیان این نکته است که شیوه و الگوریتم استفاده‌شده در هریک از ابزارها می‌تواند در نتیجهٔ نهایی تأثیرگذار باشد. لازم به ذکر است که بررسی این تأثیرگذاری در روند محاسباتی داخل برچسب‌زن و تجزیه‌گر و مدل ساخته‌شده از دادهٔ آموزش خارج از بحث مقالهٔ حاضر است و در این مقاله فقط به مقایسهٔ خروجی و کارآیی آن‌ها می‌پردازیم.

۵-۱-۲- دادهٔ آموزش

برای آموزش برچسب‌زن از پیکرهٔ برچسب‌گذاری‌شده بی‌جن‌خان [1] استفاده می‌کنیم. ویژگی این پیکره این است که اطلاعات صرفی-نحوی و معنایی برای واژه‌های پیکره وجود دارد. ساختار برچسب‌ها به‌صورت سلسله‌مراتبی بوده و براساس دستورالعمل لیچ و ویلسون [24] تهیه شده است.

در برچسب واژه‌ها، علاوه‌بر اطلاعات دستوری شامل اسم، صفت، فعل، و غیره، اطلاعات صرفی در مورد واژه‌ها ازجمله اضافه و شخص و شمار نیز وجود دارد. ویژگی معنایی نیز با هدف نشان‌دادن تمایز معنایی هم‌نویسه‌ها استفاده شده است [11]. با درنظرگرفتن این حجم از اطلاعات، ۵۸۶ برچسب به‌دست آمده است که ابهام‌زدایی این تعداد برچسب، کار برچسب‌زن را دشوار کرده و می‌تواند بر دقت آن تأثیر منفی بگذارد. جدول (۱)، نمونه‌هایی از ترتیب اطلاعات صرفی-نحوی و معنایی در برچسب‌های پیکرهٔ بی‌جن‌خان را نمایش می‌دهد.

(جدول- ۱): برچسب واژهٔ «دفتر» در پیکرهٔ بی‌جن‌خان

(Table-1): POS tag of the word «دفتر» /daftar/ in the Bijankhan Corpus

واژه	برچسب اصلی در پیکره
دفتر	N,COM,SING
دفتر	N,COM,SING,EZ
دفتر	N,COM,SING,LOC
دفتر	N,COM,SING,LOC,EZ

در برچسب‌های پیکرهٔ بی‌جن‌خان، دو کاستی و مشکل وجود دارد. مشکل نخست این است که طول برچسب‌ها متفاوت است. براساس جدول (۱)، طول برچسب‌ها بین سه تا پنج برچسب متغیر است. مشکل دوم که متأثر از مشکل نخست است این است که جایگاه هر اطلاع ثابت نیست. همان‌گونه که در جدول (۱) دیده می‌شود، جایگاه واژه‌بست و اطلاعات معنایی در یک جایگاه ظاهر شده است که از نظر منطقی قابل‌پذیرش نیست؛ چون محتوای این دو برچسب متفاوت است. به‌دلیل این دو کاستی و مشکل، قالب برچسب‌های پیکرهٔ بی‌جن‌خان به چارچوب MulText-East[5] تبدیل شده است که نمونه این تغییرات در جدول (۲) آمده است [15].

(جدول- ۲): برچسب واژه «دفتر» در پیکرهٔ بی‌جن‌خان و دادگان درختی زبان فارسی

(Table-2): POS tag of the word «دفتر» /daftar/ in the Bijankhan Corpus and the Persian Treebank

واژه	برچسب اصلی	برچسب تبدیل‌شده
دفتر	N,COM,SING	Ncsp—
دفتر	N,COM,SING,EZ	Ncsp–z
دفتر	N,COM,SING,LOC	Ncspk–
دفتر	N,COM,SING,LOC,EZ	Ncspk-z

در برچسب‌های تبدیل‌شدهٔ جدول (۲)، طول برچسب برای اسم، هفت بوده و در تمام واژه‌های اسم یکسان است. در این شکل جدید برچسب، جایگاه نخست متعلق به مقولهٔ دستوری اصلی واژه است. در جایگاه دوم نوع اسم از نظر عام و خاص مشخص می‌شود. جایگاه سوم مختص شمار است. قطبیت مثبت یا منفی در جایگاه چهارم ظاهر می‌شود. جایگاه پنجم برای ویژگی‌های معنایی است که بیشتر برای ابهام‌زدایی هم‌نویسه‌ها به‌کار رفته است، جایگاه ششم برای اختصار و جایگاه هفتم متعلق به واژه‌بست است. براساس جدول (۲)، میزان اطلاعات زبان‌شناختی برچسب‌های دستوری در پیکرهٔ بی‌جن‌خان و دادگان درختی زبان فارسی تهیه‌شده توسط قیومی [15] یکسان و فقط نحوهٔ نمایش آن با رفع کاستی‌ها تغییر یافته‌است.

برای آموزش تجزیه‌گر سازه‌ای آماری، از دادگان درختی تهیه‌شده توسط قیومی [15] که شامل ۱۰۲۸ جمله

[1] TnT Part-of-Speech Tagger
[2] Stanford Part-of-Speech Tagger
[3] Stanford Parser
[4] Berkley Parser
[5] http://nl.ijs.si/ME/

است، استفاده می‌کنیم. این دادگان نخستین دادگان درختی تهیه‌شده برای زبان فارسی و تنها دادگانی است که براساس دستور ساخت‌سازه‌ای هسته‌بنیان تهیه شده و به‌طور رایگان موجود است.[1] ویژگی این داده این است که علاوه‌بر تحلیل سلسله‌مراتبی ساخت‌سازه‌ای هر جمله، نوع رابطهٔ عناصر و سازه‌ها با هسته از نظر رابطهٔ فاعلی، متممی، یا آویزه مشخص شده است. اگرچه این اطلاعات تحلیل عمیق‌تری از جمله ارائه می‌کند، علاوه‌بر تجزیهٔ سلسله‌مراتبی ساخت جمله و برچسب‌زنی گره‌ها، ابهام‌زدایی در تشخیص نوع رابطهٔ وابستگی در سازه نیز می‌بایستی صورت گیرد که سبب افزایش دشواری تجزیه‌گر شده و ممکن است، منجر به کاهش کارآیی آن شود. شکل (۱)، نمونه درخت تجزیهٔ جملهٔ «بورن به‌اتفاق پدرش به پراگ مهاجرت کرد .» را نمایش می‌دهد. همان‌گونه که در شکل نشان داده شده است، علاوه‌بر برچسب گرهٔ هر سازه، مانندNP،VP، وPP، نوع رابطهٔ وابستگی سازه که بین هسته و واژهٔ دیگر وجود دارد، مشخص شده است. برای مثال، اگر در یک سازه، هستهٔ گروه اسمی واژهٔ دیگر را جزء موضوع[2] هسته انتخاب کند، آن واژه به‌عنوان متمم[3] هسته تلقی شده و برچسب NPC انتخاب می‌شود. چنانچه واژهٔ دیگر سازه جزء موضوع هسته نباشد، آن واژه آویزه[4] است و رابطهٔ وابستگی بین عناصر این سازه آویزه‌ای است و برچسب NPA انتخاب می‌شود.

برای ارزیابی «صحت برچسب‌دار» و «پوشش برچسب‌دار» از ابزار Evalb[5] استفاده می‌کنیم. برای ارزیابی کارآیی تجزیه‌گر براساس روابط وابستگی، نیاز به تبدیل خودکار تجزیه سازه‌ای به وابستگی است که برای این منظور از الگوریتم معرفی‌شده توسط قیومی و کوهن [18] استفاده می‌کنیم.

برای محاسبهٔ کارآیی تجزیه‌گر از «ارزیابی حد وسط ۱۰تایی»[6] استفاده می‌کنیم. در این روش، کل دادهٔ موجود به ده قسمت بدون هم‌پوشی تقسیم شده، و در هر مرحله یکی از این قسمت‌ها به‌عنوان دادهٔ آزمون برای ارزیابی و ۹ قسمت دیگر برای آموزش استفاده می‌شود. در انجام آزمایش‌ها، تمام جملات دادهٔ آموزش و آزمون بدون هرگونه پالایشی مانند حذف جملات بلند یا حذف مقوله‌های تهی مورد استفاده قرار گرفته‌است.

[1] http://hpsg.fu-berlin.de/~ghayoomi/PTB.html
[2] Argument
[3] Complement
[4] Adjunct
[5] http://nlp.cs.nyu.edu/evalb/
[6] 10-fold cross validation

۵-۲- ارزیابی

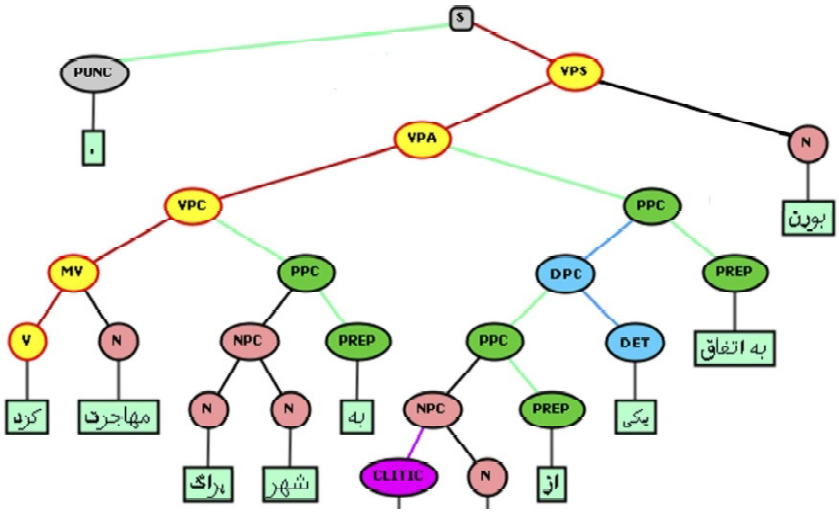
در انجام آزمایش‌ها،سه سناریو پیشنهاد می‌شود تا تصویر واضح‌تری از اثرگذاری برچسب مقوله‌های دستوری بر تجزیه ارائه شود.

در سناریوی نخست، تجزیه‌گر با دادهٔ آموزش حاصل از دادگان درختی آموزش می‌بیند و سپس دادهٔ آزمون به‌صورت خام به آن داده می‌شود. چالش این سناریو این است که تجزیه‌گر می‌بایستی با توجه به دادهٔ آموزش، ابتدا به دادهٔ آزمون ورودی برچسب مقوله‌ای دستوری بزند و سپس شروع به تجزیهٔجمله کند. این سناریو، صورت عادی کاربرد تجزیه‌گر است. نتایج حاصل از این آزمایش در جدول (۳) آمده است. مشکل سناریوی نخست این است که به‌دلیل کمبود دادهٔ آموزش، مدل خوبی برای برچسب‌زنی ساخته نشده و اشتباه در برچسب مقولات دستوری واژه‌ها منجربه اشتباه و کاهش کارآیی تجزیه‌گر می‌شود. برای بررسی عینی‌تر، دقت برچسب‌زن تجزیه‌گرها را محاسبه کرده‌ایم که نتایج آن در جدول (۴) گزارش شده است. همان‌طورکه از نتایج این جدول برمی‌آید، برچسب‌زنی این تجزیه‌گرها دقت بالایی ندارد که این امر سبب گمراه‌ساختن تجزیه‌گر می‌شود. به‌دلیل دقت پایین برچسب مقولات دستوری، سناریوی دوم مطرح می‌شود.

در سناریوی دوم ، تجزیه‌گر با دادهٔ آموزش حاصل از دادگان درختی آموزش می‌بیند و سپس دادهٔ آزمون که به‌صورت مجزا برچسب‌زنی شده است به آن داده می‌شود. در این سناریو، تجزیه‌گر تنها به تجزیهٔ جمله می‌پردازد و دخالتی در برچسب مقولات دستوری ندارد. به‌منظور برچسب‌زنی دادهٔ آزمون در این سناریو، از یک برچسب‌زن خارج از تجزیه‌گر مانند تی‌ـاِن‌ـتی یا استنفورد استفاده شده و آن‌ها را با پیکرهٔ برچسب‌زده‌شدهٔ بی‌جن‌خان آموزش می‌دهیم. ازآنجاکه این پیکره بسیار بزرگ‌تر از دادگان درختی است، انتظار می‌رود فرآیند برچسب‌زنی با دقت بسیار بالاتری صورت پذیرد. گفتنی است که جملات موجود در دادگان درختی از پیکرهٔ آموزش برچسب‌زن حذف شده و به‌عنوان دادهٔ آزمون به برچسب‌زن داده می‌شود تا هیچ‌گونه هم‌پوشی بین دادهٔ آموزش و آزمون برچسب‌زن نباشد.

میزان اطلاعات در برچسب مقولات دستوری می‌تواند کلی (دانه‌ـدرشت[7]) یا جزئی (دانه‌ـریز[8]) باشد. در جدول (۵) اطلاعات مربوط به دادهٔ آزمون سناریوی دوم نمایش داده

[7] Coarse-grained
[8] Fine-grained



(شکل- ۱): نمونه درخت تحلیل از دادگان درختی زبان فارسی
(Figure-1): Sample tree analysis from the Persian Treebank

شده است. همان‌طورکه مشاهده می‌شود، تغییر در میزان اطلاعات برچسب مقولات دستوری سبب تغییر در تعداد برچسب‌ها می‌شود که نتیجهٔ آن آسانی یا پیچیدگی کار برچسب‌زن بوده و بر دقت برچسب‌زن تأثیرگذار است. در جدول (۶)، دقت برچسب‌زن‌های خارجی تی‌اِن ـتی و استنفورد در سناریوی دوم با توجه به دو سطح اطلاع دانه‌ریزی و دانه‌درشتی مقولات دستوری نمایش داده شده است. از مقایسهٔ دقت برچسب‌زن‌ها با استفاده از اطلاعات دانه‌ریز در جدول (۴) و (۶) در می‌یابیم که افزایش حجم دادهٔ آموزش بر افزایش دقت برچسب‌زن تأثیر مستقیم دارد که در ادامهٔ این بخش با انجام آزمایش‌هایی این نکته بررسی می‌شود.

ازآنجاکه دقت برچسب‌زن استنفورد بیش‌تر از تی‌اِن ـتی است، از برچسب‌زن استنفورد برای انجام آزمایش‌های سناریوی دوم استفاده می‌شود. نتایج حاصل از آزمایش‌های سناریوی دوم در جدول (۷) آمده است. از مقایسهٔ جدول (۳) و (۷) در می‌یابیم که بهبود کیفت برچسب مقولات دستوری واژه‌ها بر کارآیی تجزیه‌گر تأثیر مستقیم دارد. این تأثیر در تمامی معیارهای ارزیابی قابل مشاهده است.

گفته شد که افزایش حجم دادهٔ آموزش برچسب‌زن بر افزایش دقت آن تأثیرگذار است. برای این منظور، آزمایشی را انجام داده‌ایم که در آن دقت برچسب‌زن در سناریوی نخست و دوم مقایسه می‌شود. برای بی‌اثرکردن تأثیر ابزار و نحوهٔ استخراج ویژگی‌ها و همچنین حجم دادهٔ آزمون به‌عنوان متغیّر بر دقت، از برچسب‌زن خارجی استنفورد و دادهٔ آزمون یکسان برای هر دو آزمایش استفاده شده است. تنها تفاوت این دو آزمایش در دادهٔ آموزش برچسب‌زن است که در سناریوی نخست فقط از جملات دادگان درختی تهیه‌شدهٔ زبان فارسی، و در سناریوی دوم از جملات پیکرهٔ بی‌جن‌خان استفاده شده است. مقایسهٔ نتایج گویای این است که افزایش حجم دادهٔ آموزش در سناریوی دوم سبب افزایش ۹/۸۲ درصدی نسبت به سناریوی نخست در دقت برچسب‌زن شده است.

سناریوی دوم می‌تواند مدل دو مرحله‌ای از کاربرد عادی تجزیه‌گر باشد؛ ولی همچنان ظرفیت برای ارتقای کیفیت دقت تجزیه‌گر وجود دارد. بر همین اساس سناریوی سوم که در آن نهایت دقت برچسب مقولات دستوری به‌کار رفته است پیشنهاد می‌شود تا حد نهایی کارآیی تجزیه‌گر مشخص شود و زمینهٔ پژوهش را با هدف ارائهٔ روش‌هایی برای بهبود کارآیی تجزیه‌گر در مطالعات آینده فراهم آورد.

در سناریوی سوم، تجزیه‌گر با دادهٔ آموزش حاصل از دادگان درختی آموزش می‌بیند و سپس دادهٔ آزمون برچسب‌زده‌شده به آن داده می‌شود. در این سناریو از برچسب‌زن خارجی استفاده نمی‌شود؛ بلکه برچسب معیار به واژه‌ها ملحق شده، و تجزیه‌گر مجبور به استفاده از این برچسب‌ها است. نتایج حاصل از سناریوی سوم در جدول (۸) آمده است.

از مقایسهٔ جداول (۳، ۷، و ۸) می‌توان به این نتیجه رسید که کیفیت برچسب مقولات دستوری به‌طورمستقیم بر کارآیی تجزیه‌گر مؤثر است؛ زیرا با بهبود کیفیت برچسب‌ها، کارآیی تجزیه‌گرها با الگوریتم‌های متفاوت افزایش می‌یابد. این تأثیر مثبت در تمام معیارهای ارزیابی قابل مشاهده است.

با مقایسهٔ سناریوهای معرفی‌شده، ذکر این نکته حائز اهمیت است که اگرچه سناریوی نخست به‌صورت پیش‌فرض در بیش‌تر کاربردهای عادی تجزیه‌گر مورد استفاده قرار می‌گیرد، دقت بالاتر برچسب‌زن و کارآیی بالاتر تجزیه‌گر در سناریوی دوم نشان‌دهندهٔ ارزش تلاش بیشتر برای پیاده‌سازی سناریوی دوم است تا این سناریو به‌عنوان کاربرد ایده‌آل مورد توجه قرار گیرد. سناریوی سوم نشان‌دهندهٔ نتایجی است که می‌توان از تجزیه‌گر در شرایطی به‌دست‌آورد که برچسب‌زن هیچ خطایی نداشته باشد. این سناریو در کاربرد واقعی قابل پیاده‌سازی نیست؛ ولی می‌تواند برای انجام آزمایش‌های مختلفی که سعی در بهبود کارآیی تجزیه‌گر دارد، مورد استفاده قرارگیرد.

(جدول- ۳): نتایج حاصل از آزمایش‌های سناریو یک
(Table-۳): Experimental results of the first scenario

تجزیه‌گر	صحت	پوشش	امتیاز تأثیر صحت و پوشش	تناظر همسان	نیای برگ	وابستگی	
						بدون‌برچسب	بابرچسب
استنفورد	46.93	46.86	46.90	3.229	82.11	65.74	56.75
بِرکُلی	55.11	54.76	54.95	5.219	81.57	67.63	59.63
بیتْپَر	59.34	58.49	58.86	5.016	84.67	72.08	62.94

(جدول-۴): ارزیابی دقت برچسب‌زنِ تجزیه‌گر(دانه‌دریز)
(Table-4): Evaluation of the parsers' POS tagging (fine-grained)

تجزیه‌گر	دقت
استنفورد	74.67
بِرکُلی	73.72
بیتْپَر	73.79

جدول-۵: اطلاعات مربوط به دادهٔ آزمون
(Table-5): Test data information

میزان اطلاعات	تعداد برچسب	تعداد واژهٔ غیرتکرار	تعداد واژهٔ تکرار	تعداد واژه‌های ناشناخته
دانه‌ریز	247	5700	27026	513
دانه‌درشت	15			

جدول-۶: ارزیابی دقت در برچسب‌زنی کلمات
(Table-6): Evaluation of POS-tagging accuracy

برچسب‌زن	میزان اطلاعات	دقت
تی‌نِ-تی	دانه‌ریز	81.48
	دانه‌درشت	95.70
استنفورد	دانه‌ریز	95.90
	دانه‌درشت	98.37

جدول-۷: نتایج حاصل از آزمایش‌های سناریوی دوم
(Table-7): Experimental results of the second scenario

تجزیه‌گر	صحت	پوشش	امتیاز تأثیر صحت و پوشش	تناظر همسان	نیای برگ	وابستگی	
						بدون برچسب	بابرچسب
استنفورد	52.65	52.30	52.43	4.167	82.30	66.55	57.58
بِرکُلی	57.53	57.66	57.59	6.9	81.90	70.39	63.06
بیتْپَر	64.16	63.31	63.68	8.495	85.49	76.39	69.53

جدول-۸: نتایج حاصل از آزمایش‌های سناریوی سوم
(Table-8): Experimental results of the third scenario

تجزیه‌گر	صحت	پوشش	امتیاز تأثیر صحت و پوشش	تناظر همسان	نیای برگ	وابستگی	
						بدون برچسب	بابرچسب
استنفورد	59.44	59.40	59.42	4.992	85.65	72.44	63.69
بِرْکْلی	62.25	62.28	62.27	8.485	83.51	73.36	66.96
بیتْپَر	71.49	70.43	70.92	10.354	89.78	84.56	78.58

جدول-۹: ارزیابی تأثیر میزان اطلاعات مقولات دستوری بر تجزیه
(Table-9): Evaluation of the impact of POS tag information on parsing

تجزیه‌گر	میزان اطلاعات	صحت	پوشش	امتیاز تأثیر صحت و پوشش	تناظر همسان	نیای برگ	وابستگی	
							بدون برچسب	با برچسب
استنفورد	دانه‌ریز	59.44	59.40	59.42	4.992	85.65	72.44	63.69
	دانه‌درشت	47.39	47.37	47.38	4.046	80.64	59.55	50.82
بِرْکْلی	دانه‌ریز	62.25	62.28	62.27	8.485	83.51	73.36	66.96
	دانه‌درشت	47.18	47.05	47.12	3.013	79.79	58.96	50.50
بیتْپَر	دانه‌ریز	71.49	70.43	70.92	10.354	89.78	84.56	78.58
	دانه‌درشت	51.16	51.37	51.19	3.651	81.16	61.13	54.16

با مقایسهٔ تجزیه‌گرهای متفاوت استفاده‌شده در انجام آزمایش‌ها، مشاهده می‌شود که تجزیه‌گر بیتْپَر کارآیی بالاتری را نسبت به تجزیه‌گرهای استنفورد و بِرْکْلی به‌دست آورده است؛ با این تفاوت که به‌صورت عملی دریافتیم تجزیه‌گر استنفورد مقاوم‌تر[1] از تجزیه‌گر بِرْکْلی و بیتْپَر است؛ چون این تجزیه‌گر تجزیه‌ای را برای تمام جملات ارائه کرده است، درحالی‌که دو تجزیه‌گر دیگر قادر به تجزیهٔ حدود ۱۰ درصد جملات دادهٔ آزمون نبودند.

همان‌طورکه گفته شد، میزان اطلاعات مقوله‌های دستوری می‌تواند در دو سطح دانه‌ریز یا دانه‌درشت برچسب‌زنی شود که میزان اطلاعات این دو سطح می‌تواند بر کارآیی تجزیه‌گر تأثیرگذار باشد. برای بررسی تأثیر میزان اطلاعات مقولهٔ دستوری در دو سطح دانه‌ریزی یا دانه‌درشتی، آزمایش‌های دیگری را انجام داده‌ایم. در این سری آزمایش‌ها، اطلاعات مقوله‌های دستوری از دانه‌ریزی به دانه‌درشتی کاهش می‌یابد. اگرچه سناریوی دوم صورت ایده‌آل کاربرد تحلیل نحوی در پردازش زبان است، برای خنثی‌کردن تأثیر منفی کیفیت برچسب واژه بر تجزیهٔ جملهٔ این آزمایش در چارچوب سناریوی سوم انجام می‌شود. نتایج حاصل از این آزمایش در جدول (۹) آمده است.

همان‌طورکه مشاهده می‌شود در آزمایش‌هایی که میزان اطلاعات مقوله‌های دستوری دانه‌ریز است، کارآیی هر سه تجزیه‌گر استنفورد، بِرْکْلی، و بیتْپَر بیشتر از آزمایش‌هایی است که میزان اطلاعات مقولات دستوری دانه‌درشت است. بر همین اساس، نتایج گزارش‌شده در جدول (۹) گویای این نکته است که افزایش میزان اطلاعات صرفی-نحوی موجود در برچسب مقوله‌های دستوری بر افزایش کارآیی تجزیه‌گر تأثیر مستقیم دارد و این تأثیر در معیارهای ارزیابی متفاوت قابل مشاهده است. با بررسی میزان این تأثیر، می‌توان به این نتیجه دست یافت که هرچند دقت برچسب‌زن صد درصد باشد، میزان اطلاعات صرفی‌نحوی مقولات دستوری تأثیر بسیار بیشتری بر کارآیی تجزیه‌گر نسبت به دقت برچسب‌ها دارد.

۶- نتیجه‌گیری

در این مقاله تأثیر کیفیت و میزان اطلاعات مقوله‌های دستوری بر تجزیهٔ خودکار بررسی شد. به‌همین‌منظور، سه سناریو برای تجزیهٔ خودکار جملات پیشنهاد شد. در سناریوی نخست، تجزیه‌گر ابتدا می‌بایستی دادهٔ ورودی را

[1] Robust

برچسب‌زنی کرده و سپس جمله را تجزیه کند. در سناریوی دوم، از یک برچسب‌زن خارج از تجزیه‌گر استفاده شده و در سناریوی سوم برچسب معیار واژه‌ها برای تجزیهٔ جملات مورد استفاده قرار گرفته‌اند. نتایج عملی نشان داد که کیفیت برچسب مقولات دستوری بر کارآیی تجزیه‌گر تأثیر دارد. همچنین با انجام آزمایش‌ها برای بررسی تأثیرگذاری میزان اطلاعات از نظر دانه‌درشتی یا دانه‌ریزی این نتیجه به‌دست آمد که اطلاعات بیشتر درمورد ویژگی‌های صرفی-نحوی هر واژه بر کارآیی تجزیه‌گر تأثیر مثبت دارد؛ و این تأثیرگذاری نسبت به کیفیت برچسب مقولات دستوری خیلی بیشتر است.

۷- مراجع 7-refrences

[۱] م. بی‌جن‌خان،"نقش پیکرهٔ زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای،" *مجلهٔ زبانشناسی*، ج. ۳۸، صص. ۴۸ـ۶۷، ۱۳۸۳.

M. Bijankhan, "naqše peykarehāye zabāni dar neveštane yek narmafzāre rāyāneyi [The role of corpora in writing a grammar: Introducing a software]," *Iranian Journal of Linguistics*, vol. 38, pp.46-67, 2004.

[۲] م. شریفی آتشگاه، "تولید نیمه‌خودکار درخت‌بانک گروه‌های نحوی در زبان فارسی معاصر،" پایان‌نامهٔ دکترا، دانشکده ادبیات و علوم انسانی، دانشگاه تهران، ۱۳۸۸.

M. SharifiAtashgah,"toulide nimexodkāre deraxtbānke goruhhāye nahvi dar zabāne fārsi [Semi-automatic Generation of Treebanks in Persian Texts]", PhD dissertation, University of Tehran, 2009.

[۳] م.ب. صادق‌زاده، *و همکاران*، "بررسی روش‌های مؤثر بر عملکرد تجزیه‌گر آماری زبان فارسی،" *سومین همایش ملی زبان‌شناسی رایانشی*، تهران، دانشگاه شریف، ۱۳۹۳.

M.B. Sadeghzade *et al.*, "barresiye raveshāye mo'asser bar 'amalkerde taǰziyegare 'āmāriye zabāne fārsi [Study the impact of effecting methods on the performance of a statistical parser for the Persian language]," In *Proceedings of the 3rd National Conference on Computational Linguistics*, Sharif University of Technology, Tehran, 2014.

[۴] م. عاصی، "پایگاه دادهٔ زبان فارسی،" *پژوهشگران*. پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ۱۳۸۴.

M. Assi, "PLDB: The Persian Linguistics DataBase," In *Researchers*, Institute for Humanities and Cultural Studies, Tehran, Iran, 2005.

[۵] ه. فیلی، " استخراج استقرایی گرامر احتمالاتی یک زبان طبیعی به روش بی‌مربی،" پایان‌نامهٔ دکترا، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۸۵.

H. Faili, "estexrāǰe esteqrāyiye gerāmere ehtemālātiye yek zabāne tabi'i be raveše bimorabbi [Unsupervised Grammar Induction for a Natural Language]", PhD dissertation, Sharif University of Technology, Tehran, Iran, 2006.

[۶] م. قیومی، "معرفی دادگان درختی و تجزیه درختی فارسی،" در *مجموعه مقالات هشتمین کنفرانس زبانشناسی ایران*، جلد ۲، صص.۶۶۶ـ۶۷۹، ۱۳۹۲.

M. Ghayoomi, "mo'arrefiye dādegāne deraxti va taǰziyegare xodkāre fārsi [Introducing a treebank and a statistical parser for Persian]," In *Proceedings of the 8th Conference of Iranian Linguistics*, Allāme Tabātabāyi University, vol. 2, pp. 666-679, 2013.

[7] A. AleAhmad *et al.*, "Hamshahri: A standard Persian text collection," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 382–387, 2009.

[8] H. Amiri *et al.*, "Investigation on a feasible corpus for Persian POS tagging," In *Proceedings of the 12th International CSI Computer Conference*, Iran, 2007.

[9] M. Assi and M. HajiAbdolhosseini, "Grammatical tagging of a Persian corpus," *International Journal of Corpus Linguistics*, vol. 5, no. 1, pp. 69–82, 2000.

[10] M. Bahrani *et al.*, "A computational grammar for Persian based on GPSG," *Language Resources and Evaluation*, vol. 45, no. 4, pp. 387–408, 2011.

[11] M. Bijankhan *et al.*, "Lessons from building a Persian written corpus: Peykare," *Language Resources and Evaluation*, vol. 45, no. 2, pp. 143–164, 2011.

[12] T. Brants, "TnT - A statistical part-of-speech tagger," In *Proceedings of the Association for Neuro-Linguistic Programming and NAACL*, pp. 224–231, 2000.

[13] J. Dehdari and D. Lonsdale, "A link grammar parser for Persian," In *Aspects of Iranian Linguistics*, S. Karimi, V. Samiian, and D. Stilo, Eds. Cambridge Scholars Press, vol. 1, 2008.

[14] P. Denis and B. Sagot, "Coupling an annotated corpus and a morpho-syntactic lexicon for state-of-the-art POS tagging with less human effort," In *Pacific Asia Conference on Language, Information and Computation*, Hong Kong, Chine, 2009.

[15] M. Ghayoomi, "Bootstrapping the development of an HPSG-based treebank for Persian," *Linguistic Issues in Language Technology*, vol. 7, no. 1, 2012.

[16] M. Ghayoomi, "Word clustering for Persian statistical parsing," *Advances in Natural Language Processing*, volume 7614 of Lecture Notes in Computer Science: JapTAL '12: Proceedings of the 8th International Conference on Advances in Natural Language Processing, Springer Berlin Heidelberg, pp. 126–137, 2012.

International Symposium on Signal Processing and its Applications, Sharjah, (U.A.E.), 2007.

[42] K. Toutanova and C.D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," In *Proceedings of the Joint SIGDAT Conference on EMNLP and Very Large Corpora*, Hong Kong, pp. 63–70, 2000.

مسعود قیومی عضو هیأت علمی پژوهشگاه علوم انسانی و مطالعات فرهنگی است. وی فارغ‌التحصیل مقطع دکترای رایانه با گرایش زبان‌شناسی رایانشی در سال ۲۰۱۴ از دانشگاه آزاد برلین آلمان است. وی در سال ۲۰۰۹ از دانشگاه سارلند آلمان و در سال ۲۰۰۸ از دانشگاه نانسی۲ فرانسه موفق به اخذ مدرک کارشناسی ارشد در رشته زبان‌شناسی رایانشی شد. همچنین وی در سال ۱۳۸۳ دوره کارشناسی ارشد خود را در رشته زبان‌شناسی همگانی در دانشگاه آزاد واحد تهران مرکز به پایان رساند. ایشان در سال ۱۳۸۰ در رشتهٔ مترجمی زبان انگلیسی از دانشگاه آزاد اسلامی قم فارغ‌التحصیل شد. زمینه‌های تخصصی مورد علاقهٔ ایشان پردازش زبان طبیعی، مدل‌سازی زبانی، یادگیری ماشین، نحو و معناشناسی واژگانی است.

نشانی رایانامهٔ ایشان عبارت است از:

masood.ghayoomi@gmail.com