پروژهٔ دادگان وابستگی زبان فارسی

گروه پژوهشی دادگان

تابستان ۱۳۹۱

بسم الله الرحمن الرحیم

عرض کردیم که کشور باید به عزت علمی برسد. هدف هم باید مرجعیت علمی باشد در دنیا؛ همین طور که بارها عرض کرده‌ایم. یعنی همین طور که شما امروز ناچارید برای علم و دست‌یابی به محصولات علمی به دانشمندانی، به کتاب‌هایی مراجعه کنید که مربوط به کشورهای دیگرند، باید به آنجا برسیم که جویندهٔ دانش، طالب علم، مجبور باشد بیاید سراغ شما، سراغ کتاب شما؛ مجبور باشد زبان شما را یاد بگیرد تا بتواند از دانش شما استفاده کند. هدف باید این باشد. این یک آرزوی خام هم نیست. این چیزی است که عملی است. اینجایی هم که ما امروز از لحاظ علمی و فناوری قرار داریم، این هم یک روزی جزو آرزوهای خام به حساب می‌آمد.

«بخشی از سخنان رهبر معظم انقلاب در جمع استادان و دانشجویان دانشگاه علم و صنعت ایران، ۱۳۸۷»

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

دستور وابستگی

۱.۱ مقدمه

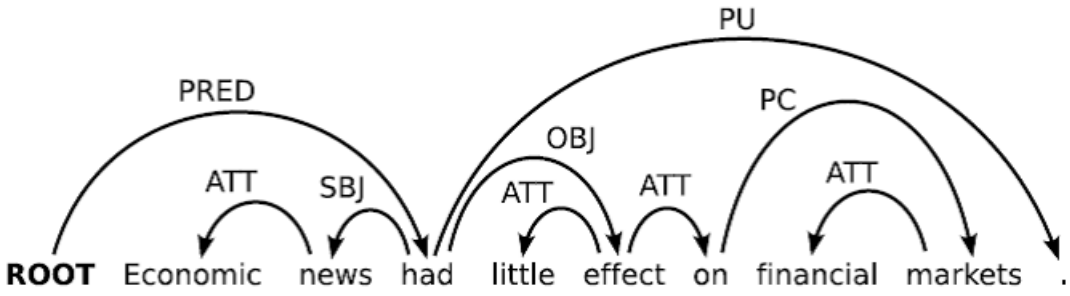
نظریهٔ دستور وابستگی یکی از نظریه‌های ساخت‌گرا و صورت‌گراست که اساساً در آن از طریق بررسی روابط وابستگی بین عناصر هسته و وابسته در زبان، به توصیف ساخت‌های نحوی در زبان‌های گوناگون پرداخته می‌شود [طبیب‌زاده(۱۳۸۵)]. شاید آغاز رویکرد زبانی وابستگی مربوط به اندیشه‌های زبان‌شناسی پانینی[۱] در مورد زبان سانسکریت باشد؛ امّا نظریات تنی‌یر[۲] آغازی بر استفاده از این رویکرد در زبان‌شناسی نوین است. او نخستین بار در کتاب کم‌حجمی با عنوان گفتارهایی در نحو ساختاری [تنی‌یر(۱۹۵۳)] این دیدگاه را مطرح کرد که شرح مبسوط آن پس از مرگش در کتاب مبانی نحو ساختاری [تنی‌یر(۱۹۵۹)] منتشر شد.

پس از تنی‌یر، زبان‌شناسان مختلف روش‌های مختلفی را برای ارائهٔ دستور زبان وابستگی پیشنهاد داده‌اند. در همهٔ این دستورها یک فرض پایه وجود دارد و آن این است که ساختار نحوی شامل واژه‌هایی است که این واژه‌ها با روابط دودویی نامتقارن با هم در ارتباط هستند. به این روابط، ارتباط وابستگی یا وابستگی گفته می‌شود [کوبلر و دیگران(۲۰۰۹)]. دو فرض اساسی در نظریهٔ دستور وابستگی وجود دارد. نخست این که هر جمله یک فعل مرکزی دارد و دوم این که بر اساس نوع و تعداد متمم‌های اجباری و اختیاری، می‌توان ساخت بنیادین جمله‌هایی را که فعل در آن‌ها به کار رفته است، تعیین کرد [طبیب‌زاده(۱۳۸۵)]. در همهٔ این رابطه‌ها یک واژه وابسته و واژهٔ دیگر هسته است. نکته‌ای که در دستور وابستگی بسیار حائز اهمیت است این است که بایستی در هر جمله وضعیت

---

[۱] Panini

[۲] Tesnière



شکل ۱.۱: نمونه‌ای از یک درخت وابستگی در یک جملهٔ انگلیسی

تمام عناصر جمله از این لحاظ که آیا عنصر مورد نظر هسته است یا وابسته مشخص شود. اگر عنصر مورد نظر هسته باشد، باید دید که آن هسته دارای چه وابسته‌هایی است و اگر عنصر مورد نظر وابسته باشد، باید دید که آن عنصر، وابستهٔ کدام هسته در جمله است. در شکل ۱.۱ نمونه‌ای از یک درخت وابستگی نشان داده شده است.

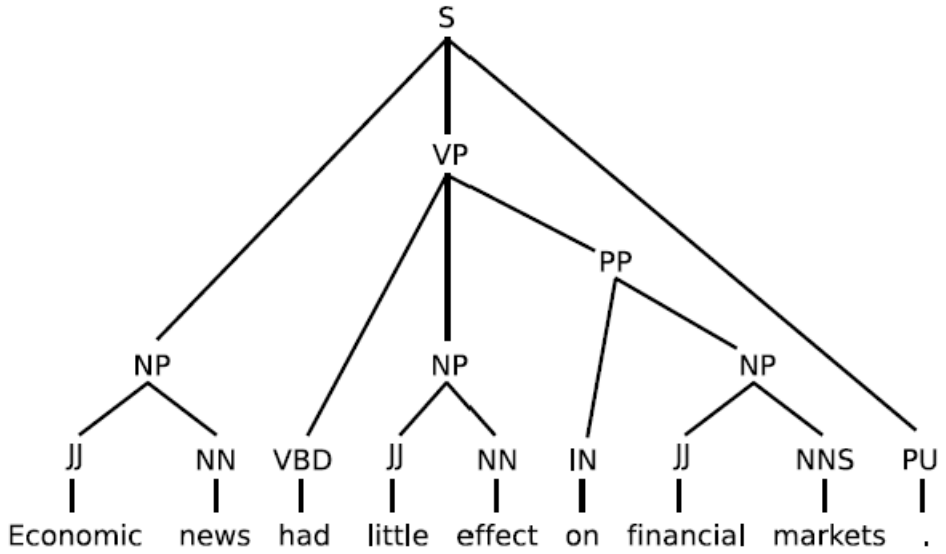
شایان ذکر است که اطلاعات موجود در ساختار وابستگی با اطلاعات موجود در ساختار مبتنی بر گروه‌ها متفاوت است. برای مقایسه می‌توان ساختار موجود در شکل ۲.۱ را با ساختار شکل ۱.۱ مورد بررسی قرار داد. در دستور وابستگی، جمله به دو بخش نهاد و گزاره تقسیم نمی‌شود. در این نظریه این که جمله به دو گروه اسمی و فعلی تقسیم می‌شود، رد می‌شود. به اعتقاد [انگل(۲۰۰۲)]، تجزیهٔ جمله به دو بخش نهاد و گزاره برای تحلیل ساخت اطلاعاتی جمله مفید است ولی در تحلیل نحوی مرکز ثقل ساختاری جمله فعل است [طبیب‌زاده(۱۳۸۵)]. البته تبدیل اطلاعات موجود هر کدام از این دو ساختار به هم، امکان‌پذیر است ولی برای سهولت فرض بر این است که رویکرد مبتنی بر ساختار وابستگی و رویکرد مبتنی بر گروه‌ها، دو رهیافت مختلف و متفاوت هستند [کوبلر و دیگران(۲۰۰۹)].

۲.۱ ضابطه‌های عمومی در دستور وابستگی

با این فرض که ساختار وابستگی نمایندهٔ خوبی برای نشان دادن ساختار نحوی زبان طبیعی است، نیاز به اعمال ضابطه‌هایی بر این ساختار وجود دارد. بنابراین با وجود ساختار زبانی C و واژهٔ وابستهٔ D و واژهٔ هستهٔ H شش ضابطهٔ زیر را خواهیم داشت [کوبلر و دیگران(۲۰۰۹)]:

۱) با استفاده از H مقولهٔ نحوی C قابل تعیین است و H می‌تواند به عنوان نمایندهٔ کل ساختار C جای C را پر کند.

۲) با استفاده از H مقولهٔ معنایی C قابل تعیین است و D به این ساختار ویژگی‌های معنایی می‌افزاید.



شکل ۲.۱: نمونه‌ای از یک درخت نحوی زایشی (مبتنی بر عبارات) در یک جملهٔ انگلیسی

۳) وجود H لازم است ولی وجود D ممکن است اختیاری باشد.

۴) هستهٔ H واژهٔ D را انتخاب و اختیاری یا اجباری بودن آن را تعیین می‌کند.

۵) شکل D وابسته به H است.

۶) موقعیت خطی D با ارجاع به H مشخص می‌شود.

همان‌طور که از ضابطه‌ها پیداست، برخی از این ساختارها مربوط به ساخت‌واژه، برخی مربوط به نحو و برخی مربوط به معناست. البته برخی بر این اعتقادند که باید برای دو نوع ساختار درون‌مرکز[۳] و برون‌مرکز[۴] ضابطه‌های متفاوتی را در نظر داشت. در ساختار درون‌مرکز ممکن است واژهٔ هسته نمایندهٔ مقولهٔ نحوی کل گروه باشد ولی در ساختار برون‌مرکز همهٔ واژه‌ها با هم مقولهٔ نحوی را می‌سازند. به عنوان مثال گروه «میز کهنه» یک ساختار درون‌مرکز است ولی گروه «روی آن میز» یک گروه برون‌مرکز است. به وضوح مشخص است که در ساختار درون‌مرکز هر شش ضابطهٔ ذکرشده صدق می‌کند؛ البته در این ساختار ضابطهٔ ۴ کمی نامناسب به نظر می‌رسد. در ساختار برون‌مرکز ضابطهٔ ۱ درست نخواهد بود [کوبلر و دیگران(۲۰۰۹)].

این دو ساختار قابل مقایسه با دو ساختار افزوده‌های واژهٔ هسته (شبیه به ساختار درون‌مرکز) و متمم‌های واژهٔ هسته (شبیه به

---

[۳] Endocentric

[۴] Exocentric

ساختار برون‌مرکز) است. تفاوت بین افزوده و متمم در مفهوم ظرفیت[۵] نهفته است. در تعریف سنتی دستور وابستگی، ظرفیت مفهومی اساسی دارد. دو زبان‌شناس آلمانی [هلبیگ و شنکل(۱۹۹۱)] با رویکرد دستور زبان زایشی[۶] در مورد ظرفیت فعل در زبان آلمانی پژوهش‌هایی را صورت دادند. تا این که [انگل(۲۰۰۲)] این مفهوم را خصوصاً در مورد نظریهٔ دستور وابستگی مطرح کرد [طبیب‌زاده(۱۳۸۵)]. بر اساس این مفهوم، هر فعل به ساختار جمله، حالات خاصی از وابسته‌ها را تحمیل می‌کند که بر این اساس به دو نوع مقید به ظرفیت و آزاد از ظرفیت تقسیم‌بندی می‌شوند [کوبلر و دیگران(۲۰۰۹)].

در زمینهٔ قواعد موجود در ساختار وابستگی بین زبان‌شناسان مختلف اختلاف نظرهایی وجود دارد. یکی از این اختلاف نظرها در مورد همپایگی‌ها است. از دیدگاه ساختارگراهای سنتی، ساختار همپایگی یک ساختار درون‌مرکز است. دلیل این ادعا، امکان جایگزینی یک یا حتی چند واژه به جای کل گروه نحوی مورد نظر است. مثلاً در جملهٔ «او سیبی را خرید و خورد»، واژهٔ «سیبی» مفعول ساختار همپایگی فعلی «خرید و خورد» است.

۳.۱ مفهوم ظرفیت در دستور وابستگی

مهم‌ترین مبحث در دستور وابستگی، عبارت است از مسألهٔ ظرفیت نحوی که در آن به بحث در مورد وابسته‌های فعل، اسم و صفت پرداخته می‌شود. بر اساس این نظریه، مرکز ثقل ساختاری جمله فعل است [طبیب‌زاده(۱۳۸۵)]. [تنی‌یر(۱۹۸۰)] مفهوم ظرفیت را از شیمی اقتباس کرده بود. ظرفیت در شیمی عبارت است از توانایی یک عنصر در ترکیب با تعداد خاصی از اتم‌های عناصر دیگر. این که ساخت بنیادین جمله حول فعل مرکزی آن صورت می‌گیرد، مبین این واقعیت است که هر فعل پیش از آن که وارد جمله بشود، خود مشخص‌کنندهٔ نوع ساخت بنیادین است [طبیب‌زاده(۱۳۸۵)].

۱.۳.۱ ساخت بنیادین جمله

ساخت‌های بنیادین جمله به ساخت‌هایی اطلاق می‌شود که از بسط و تعریف آن‌ها یا از ترکیب آن‌ها با هم یا از تبدیل آن‌ها به هم یا به ساخت‌های مشتق یا فرعی دیگر، یا از آمیزه‌ای از دو تا یا چند مورد از روش‌های گفته‌شده، بتوان تمام جمله‌های محتمل موجود در زبان را تولید کرد. استخراج چنین ساختارهایی از زمرهٔ مهم‌ترین و ابتدایی‌ترین وظایف تحلیل نحوی است؛ زیرا این ساخت‌های محدود و تکرارشونده تمام ساخت‌های نحوی هر زبان را تشکیل می‌دهند [طبیب‌زاده(۱۳۸۵)]. ظرفیت فعل مفهومی انتزاعی و متعلق به واژه‌هاست ولی ساخت بنیادین مفهومی متعلق به جمله است.

---

[۵] Valency

[۶] Generative

۴.۱ تجزیهٔ وابستگی

تجزیهٔ جملات زبان طبیعی یکی از مهم‌ترین مباحث پژوهشی در پردازش زبان طبیعی است. غالباً این نوع از پردازش بر روی *دستورهای مبتنی بر گروه‌ها* مانند آنچه که در پیکرهٔ درختی پن[۷] [مارکوس و دیگران(۱۹۹۳)] به کار برده شده، بوده است. یکی از روش‌های تجزیهٔ جملات، تجزیهٔ وابستگی است. در تجزیهٔ وابستگی ساختار درخت‌های وابستگی جملات استخراج شده، تجزیه صورت می‌گیرد.

تجزیهٔ وابستگی[۸] رهیافتی برای تجزیهٔ نحوی زبان طبیعی به صورت خودکار است. این رهیافت از زبان‌شناسی سنتی مبتنی بر دستور وابستگی اقتباس شده است. در سال‌های اخیر این روش بیش از پیش مورد توجه قرار گرفته است. چند دلیل عمده برای این اقبال عمومی وجود دارد. نخست این که این گونه از نمایش ساختار نحوی زبان طبیعی، کاربردهای بسیاری در برنامه‌های مربوط به فهم زبان طبیعی از جمله ترجمهٔ خودکار و استخراج اطلاعات دارد. دومین دلیل این است که این نوع از دستور زبان (و تجزیهٔ بر مبنای آن)، در مقایسه با دستور زبان مبتنی بر گروه‌ها، سازگاری بیش‌تری با طبیعت زبان‌های بی‌ترتیب دارد. امّا مهم‌ترین دلیل، نتایج رضایت‌بخش حاصل از اعمال این روش در برخی از زبان‌ها با استفاده از روش‌های یادگیری خودکار بوده است [کوبلر و دیگران(۲۰۰۹)].

۱.۴.۱ روش‌های تجزیهٔ وابستگی

در مجموع می‌توان گفت که در تجزیهٔ وابستگی برای هر جملهٔ ورودی یک گراف وابستگی ساخته می‌شود و دو رهیافت عمومی برای آن وجود دارد : ۱) مبتنی بر داده[۹]؛ و ۲) مبتنی بر دستور زبان[۱۰]. در رهیافت مبتنی بر داده، از روش‌های یادگیری خودکار[۱۱] و در روش‌های مبتنی بر دستور زبان از دستور زبان‌های صوری[۱۲] استفاده می‌شود. البته بدیهی است که می‌توان از هر دو رهیافت به صورت تلفیقی برای تجزیهٔ وابستگی استفاده کرد [کوبلر و دیگران(۲۰۰۹)].

یکی از مهم‌ترین روش‌های یادگیری خودکار، یادگیری باناظر است که معمولاً در این یادگیری دقت بالایی به دست می‌آید. در روش یادگیری باناظر[۱۳] دو مرحلهٔ اصلی در ساخت یک سامانهٔ تجزیهٔ وابستگی وجود دارد. در مرحلهٔ نخست با استفاده از یک پیکرهٔ آموزشی، دستور زبان وابستگی به دست می‌آید. به عمل به دست آوردن دستور زبان، استنتاج دستور زبان[۱۴] گویند. با به دست آمدن دستور زبان وابستگی، الگوی تجزیه به دست خواهد آمد. در مرحلهٔ بعدی بر اساس الگوی به دست آمده در مرحلهٔ قبل، برای هر جملهٔ

---

[۷] Penn Treebank

[۸] Dependency Parsing

[۹] Data Driven

[۱۰] Grammar Based

[۱۱] Machine Learning

[۱۲] Formal

[۱۳] Supervised Learning

[۱۴] Grammar Inference

ورودی، گراف وابستگی تولید خواهد شد. به مرحلهٔ اوّل یادگیری و به مرحلهٔ دوم تجزیه گفته می‌شود [کوبلر و دیگران(۲۰۰۹)].

انواع مختلفی از روش‌های مبتنی بر داده وجود دارد که از مهم‌ترین آن‌ها می‌توان به روش مبتنی بر گذار[۱۵] و روش مبتنی بر گراف اشاره کرد. در همهٔ روش‌های مبتنی بر داده فرض اولیه بر این است که داده‌های ورودی حتماً دارای ساختار نحوی درست هستند ولی در روش‌های مبتنی بر دستور زبان فرض بر این است که اگر ساختاری در قالب هیچ یک از قواعد موجود در پایگاه قوانین نگنجد، آن جمله از نظر دستوری نادرست است. روش‌های مبتنی بر دستور زبان به دو نوع اصلی مستقل از متن و مبتنی بر محدودیت تقسیم می‌شوند. در روش مستقل از متن، ساختار وابستگی تبدیل به عبارات مستقل از متن می‌شود و بر اساس دستور زبان مستقل از متن تجزیه صورت می‌گیرد و در روش مبتنی بر محدودیت، صورت مسأله تبدیل به مسألهٔ ارضای محدودیت[۱۶] می‌شود [کوبلر و دیگران(۲۰۰۹)][۱۷].

---

[۱۵] Transition Based

[۱۶] Constraint Satisfaction

[۱۷] پیش از این در گزارشی مفصل [رسولی(۱۳۸۹)]، در مورد تجزیهٔ وابستگی و انواع روش‌های آن توضیح داده شده است و لذا از تکرار مباحث آن پرهیز شده است.

فصل ۲

پیکرهٔ نحوی وابستگی

۱.۲ قالب‌بندی داده‌ها

این داده‌ها بر اساس قالب معیار همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی بر روی پیکره‌های وابستگی [نیلسون و دیگران(۲۰۰۷)] فراهم آمده است. با این تفاوت که در قالب یادشده وجود فاصله در واژه‌ها به هیچ وجه مجاز نیست در حالی که ما استثنائاً وجود فاصله در اجزای تصریف فعل مرکب را به رسمیت شناختیم[۱].

ویژگی‌های ساخت‌واژی‌ای که برای واژه‌ها در نظر گرفته‌ایم شامل شخص، شمار، وجه و نمود (برای فعل) و چسبیدگی واژه است. لازم به ذکر است چسبیدگی واژه شامل حالتی است که مجبور شده‌ایم واژه را منقطع کنیم و به دو رشته تبدیل نماییم؛ مانند تبدیل «گفتمش» به «گفتم» و «ش».

برای این که امکان گزارش خطا و اصلاح وجود داشته باشد، یک ویژگی مصنوعی با عنوان *senID* نیز اضافه کرده‌ایم که در واقع شمارهٔ جمله‌ای است که برچسب خورده و این شمارهٔ جمله دقیقاً با شمارهٔ جملهٔ موجود در پایگاه دادهٔ دادگان برابری می‌کند. در واقع زمانی که مخاطب با خطایی مواجه می‌شود، می‌تواند شمارهٔ جمله را به همراه توضیحات در صفحهٔ گزارش خطا در وبگاه دادگان درج کرده برای گروه دادگان ارسال نماید.

[۱] اگر این فاصله‌ها را با نویسه‌هایی خاص مانند _ جایگزین نماییم این مشکل مرتفع می‌شود.

۲.۲ منابع متنی مورد استفاده در پیکره

در مرحلهٔ نخست از این پیکره از چندین هزار صفحهٔ خبری موجود در خبرگزاری مهر حدود ۳ هزار جمله استخراج شد و به صورت تدریجی به برچسب‌زنان داده شد. دلیل اعطای تدریجی جملات، استفاده از روش‌های یادگیری هوشمند برای برچسب‌زنی جملات خام بر اساس داده‌های یادگیری قبلی بوده است. سپس از روی جملات داستانی موجود در مجلاتی مانند ادبیات داستانی، برخی از مجلات هنری، برخی از جملات مستخرج از نویسندگان به‌نام ادبیات فارسی[۲]، متون موجود در سخنرانی‌ها و مقالات نویسندگان و اندیشمندان به‌نام[۳] و برخی از منابع آموزش زبان فارسی به غیرفارسی‌زبانان جملاتی به صورت تصادفی وارد پیکره شد. در نهایت با آمارگیری از روی پیکره و استخراج فعل‌هایی که در نسخهٔ ۱ فرهنگ ظرفیت وجود دارند ولی در پیکره نبوده‌اند و با کمک گروه داده‌گزینی به صورت تصادفی از تصریف‌های مختلف هر فعل چند نمونه[۴] با استفاده از موتورهای جستجوی وب به دست آمدند.

۳.۲ قواعد زبانی پیکره

در این بخش مجموعه برچسب‌های وابستگی (روابط وابستگی) موجود در پیکره معرفی می‌شود. این برچسب‌ها و روابط پس از پژوهش‌ها و واکاوی‌های فراوان به شکل زیر درآمده است. پس از توضیحات پیرامون روابط وابستگی، برچسب‌های اجزای سخن و ویژگی‌های ساخت‌واژی به کار رفته در این پیکره مورد بررسی قرار گرفته است.

۱.۳.۲ روابط وابستگی در پیکره

وابسته‌های فعل

- فاعل: SBJ[۵]

مثال ۱.۲. *علی به خانه آمد.*
آمد $\xleftarrow{SBJ}$ *علی*

- مفعول: OBJ

[۲] صادق هدایت، جلال آل احمد، مصطفی مستور، رضا امیرخانی، محمود دولت‌آبادی، سیمین دانشور و صمد بهرنگی از جملهٔ این نویسندگان هستند.

[۳] متون و سخنرانی‌های حضرت آیت‌الله خامنه‌ای، امام خمینی (ره)، شهید مطهری، دکتر شریعتی، بهاءالدین خرمشاهی و وصیت‌نامهٔ شهدا از جملهٔ این متون بوده‌اند.

[۴] بین ۴ تا ۸ نمونه بسته به این که چه مقدار داده از روی وب می‌توانستیم از این فعل‌ها فراهم کنیم

[۵] در مثال‌ها، به صورت قراردادی جهت کمان از سمت هسته به سمت وابسته است.

مثال ۲.۲. *من کتاب خواندم.*
خواندم $\xleftarrow{OBJ}$ *کتاب*

مثال ۳.۲. *کتاب را خواندم.*
خواندم $\xleftarrow{OBJ}$ *را*

- فعل‌یار: NVE

مثال ۴.۲. *با تو صحبت کردم.*
کردم $\xleftarrow{NVE}$ *صحبت*

- فعل‌یار پی‌بستی: ENC

مثال ۵.۲. *از تو خوشم آمد.*
آمد $\xleftarrow{ENC}$ *خوشم*

- مفعول حرف اضافه‌ای: VPP

مثال ۶.۲. *علی به مدرسه آمد.*
آمد $\xleftarrow{VPP}$ *به*

- مفعول دوم: OBJ۲

مثال ۷.۲. *کتاب را به علی هدیه دادم.*
دادم $\xleftarrow{OBJ۲}$ *هدیه*

مثال ۸.۲. *کتابی به علی هدیه دادم.*
دادم $\xleftarrow{OBJ۲}$ *هدیه*

- تمیز: TAM

مثال ۹.۲. *علی را مردی خوب می‌پندارم.*
می‌پندارم $\xleftarrow{TAM}$ *مردی*

- مسند: MOS

مثال ۱۰.۲. *هوا سرد است.*
است $\xleftarrow{MOS}$ *سرد*

- مستمرساز: PROG

مثال ۱۱.۲.
دارم می‌آیم.
می‌آیم $\xleftarrow{PROG}$ *دارم*

- متمم قیدی فعل: ADVC

مثال ۱۲.۲.
تهران ماندم.
ماندم $\xleftarrow{ADVC}$ *تهران*

- بند متممی فعل: VCL

مثال ۱۳.۲.
می‌دانم که می‌آید.
می‌دانم $\xleftarrow{VCL}$ *که*

مثال ۱۴.۲.
می‌دانم می‌آید.
می‌دانم $\xleftarrow{VCL}$ *می‌آید*

- حرف اضافهٔ فعلی: VPRT

مثال ۱۵.۲.
قدرت به دست آورد.
آورد $\xleftarrow{VPRT}$ *به*

- جزء همکرد: LVP

کارخانه به تهران انتقال پیدا کرد.

مثال ۱۶.۲.
کرد $\xleftarrow{LVP}$ *پیدا*
کرد $\xleftarrow{NVE}$ *انتقال*

- بند وصفی: PARCL

مثال ۱۷.۲.
به خانه رفته، خوابیدم.
خوابیدم $\xleftarrow{PARCL}$ *رفته*

- قید: ADV

مثال ۱۸.۲.
برای خرید رفتم.
رفتم $\xleftarrow{ADV}$ *برای*

مثال ۱۹.۲. *عمداً به او فحش دادم.*
دادم $\xleftarrow{ADV}$ *عمداً*

- بند افزودهٔ فعل: AJUCL

مثال ۲۰.۲. *اگر بیایی خوشحال می‌شوم.*
می‌شوم $\xleftarrow{AJUCL}$ *اگر*

مثال ۲۱.۲. *بیایی خوشحال می‌شوم.*
می‌شوم $\xleftarrow{AJUCL}$ *بیایی*

- افزودهٔ پرسشی فعل: PART

مثال ۲۲.۲. *آیا حرفم را باور داری؟*
داری $\xleftarrow{PART}$ *آیا*

- همپایهٔ فعل: VCONJ

مثال ۲۳.۲. *به خانه رفت و خوابید.*
خوابید $\xleftarrow{VCONJ}$ *و*

وابسته‌های اسم

- صفت پیشین اسم: NPREMOD

مثال ۲۴.۲. *بهترین دوست کتاب است.*
دوست $\xleftarrow{NPREMOD}$ *بهترین*

- صفت پسین اسم: NPOSTMOD

مثال ۲۵.۲. *دوست خوب نعمتی است برای خودش.*
دوست $\xleftarrow{NPOSTMOD}$ *خوب*

- حرف اضافهٔ اسم: NPP

مثال ۲۶.۲. *جدال در تاسوکی را دیدم.*
جدال $\xleftarrow{NPP}$ *در*

مثال ۲.۲۷. *به خدا اتکا کنید.*
اتکا $\xleftarrow{NPP}$ *به*

- بند اسم: NCL

مثال ۲.۲۸. *مردی که دیدی پدرم بود.*
مردی $\xleftarrow{NCL}$ *که*

- مضافٌ‌الیه: MOZ

مثال ۲.۲۹. *کتاب حسن را پیدا کردم.*
کتاب $\xleftarrow{MOZ}$ *حسن*

- بدل: APP

مثال ۲.۳۰. *سعدی شاعر شیرازی از به‌نام‌ترین شاعران تاریخ ادبیات فارسی است.*
سعدی $\xleftarrow{APP}$ *شاعر*

- همپایهٔ اسم: NCONJ

مثال ۲.۳۱. *قیصر و سلمان با هم دوست و رفیق بودند.*
قیصر $\xleftarrow{NCONJ}$ *و*

- قید اسم: NADV

مثال ۲.۳۲. *تهران سکونت دارم.*
سکونت $\xleftarrow{NADV}$ *تهران*

- اسم‌یار: NE

مثال ۲.۳۳. *اخراج کردن کارمندان به صلاح شرکت نیست.*
کردن $\xleftarrow{NE}$ *اخراج*

- ممیز: MESU

مثال ۲.۳۴. *دو جلد کتاب از آنجا خریدم.*
کتاب $\xleftarrow{MESU}$ *جلد*

- جزء اسمی: NPRT

مثال ۲.۳۵. *از دست دادن عزیزان غمی بزرگ بر دل می‌نهد.*
دادن $\xleftarrow{NPRT}$ *از*

وابسته‌های صفت

- حرف اضافهٔ تفضیلی: COMPPP

مثال ۲.۳۶. *دانایی از همه چیز بهتر است.*
بهتر $\xleftarrow{COMPPP}$ *از*

- قید صفت: ADJADV

مثال ۲.۳۷. *تاکسی سوار شدم.*
سوار $\xleftarrow{ADJADV}$ *تاکسی*

- متمم بندی صفت: ACL

مثال ۲.۳۸. *آگاه هستم که می‌آیی.*
آگاه $\xleftarrow{ACL}$ *که*

- متمم حرف اضافه‌ای صفت: AJPP

مثال ۲.۳۹. *با شهرتان آشنا هستم.*
آشنا $\xleftarrow{AJPP}$ *با*

- متمم نشانهٔ اضافه‌ای صفت: NEZ

مثال ۲.۴۰. *نگران او هستم.*
نگران $\xleftarrow{NEZ}$ *او*

- همپایهٔ صفت: AJCONJ

مثال ۲.۴۱. *تو خوب و زیبا هستی.*
خوب $\xleftarrow{AJCONJ}$ *و*

- وابستهٔ پیشین صفت: APREMOD

مثال ۲.۴۲. *تو بسیار شاد هستی.*
شاد $\xleftarrow{APREMOD}$ *بسیار*

- وابستهٔ پسین صفت: APOSTMOD

مثال ۲.۴۳. *پیراهن آبیِ آسمانی به تن دارد.*
آبی $\xleftarrow{APOSTMOD}$ *آسمانی*

دیگر روابط وابستگی

- وابستهٔ پیشین: PREDEP

مثال ۴۴.۲. *علی را دیدم.*
را $\xleftarrow{PREDEP}$ *علی*

مثال ۴۵.۲. *شاد کردن دل دیگران کار بسیار نیکویی است.*
کردن $\xleftarrow{PREDEP}$ *شاد*

مثال ۴۶.۲. *حتی معلم هم به درد من پی برد.*
معلم $\xleftarrow{PREDEP}$ *حتی*

- وابستهٔ پسین: POSDEP

مثال ۴۷.۲. *کتاب را به علی دادم.*
به $\xleftarrow{POSDEP}$ *علی*

مثال ۴۸.۲. *کتاب و دفتر وسایل مورد نیاز برای تحصیل هستند.*
کتاب $\xleftarrow{POSDEP}$ *و*

- همپایهٔ حرف اضافه: PCONJ

مثال ۴۹.۲. *در تهران و با ما بود.*
در $\xleftarrow{PCONJ}$ *و*

- همپایهٔ قید: AVCONJ

مثال ۵۰.۲. *این که عمداً یا سهواً می‌نویسید به من ربطی ندارد.*
عمداً $\xleftarrow{AVCONJ}$ *یا*

- گزاره: PRD

مثال ۵۱.۲. *آمدم تا ببینم چه بلایی سرتان آمده است.*
تا $\xleftarrow{PRD}$ *ببینم*

- ریشهٔ جمله: ROOT

مثال ۲.۵۲. *آمدم تا ببینم چه بلایی سرتان آمده است.*
ریشهٔ جمله $\xleftarrow{ROOT}$ آمدم

- PUNC : علامت نگارشی:

مثال ۲.۵۳. *آمدم تا ببینم چه بلایی سرتان آمده است.*
آمدم $\xleftarrow{PUNC}$.

۲.۳.۲ ویژگی‌های ساخت‌واژی و برچسب‌های اجزای سخن

در این پیکره از ۱۷ نوع برچسب اجزای سخن استفاده شده است. بسته به این که چه قدر نیاز به وارد شدن به جزئیات پیکره بوده‌ایم، هر برچسب اجزای سخن دارای ویژگی‌ها و برچسب‌های ریزتری خواهند بود که در جدول‌های ۲.۲-۴.۲ نشان داده شده است.

۴.۲ آماره‌های پیکرهٔ وابستگی

پس از حذف جملات خدشه‌دار و نامناسب که در حدود ۱۰۰۰ جمله بوده است و نیز تلفیق جملاتی که از عمد به صورت تکراری در پیکره درج شده بودند، در نهایت پیکره‌ای با ۲۹۹۸۲ جمله به دست آمد. در نهایت این پیکره با نسبت ۱۰-۱۰-۸۰ و به صورت تصادفی به بخش‌های یادگیری، ارزیابی و آزمون تبدیل شد. آماره‌های مربوط به فراوانی واژه‌ها در پیکرهٔ وابستگی در جدول ۵.۲ نشان داده شده است. همچنین آماره‌های مربوط به درصد حضور هر برچسب اجزای سخن و هر رابطهٔ وابستگی در جدول‌های ۶.۲ و ۷.۲ نشان داده شده است.

۱.۴.۲ اصلاحات و میزان هماهنگی برچسب‌زنی

از روی ۸۴۳ جملهٔ تکراری نیز میزان توافق برچسب‌زنان با یکدیگر اندازه‌گیری شده است تا بدین‌وسیله معیاری از یکسانی دیدگاه‌های زبانی برچسب‌زنان در پروژه داشته باشیم. این نتایج در جدول ۸.۲ نشان داده شده است. همچنین برای این که در نهایت کار دارای کیفیت بالایی شود، برخی از قواعد زبانی به صورت دستی مورد بررسی قرار گرفت و با استفاده از این قواعد برنامه‌ای نوشته شد که با آن خطاهای موجود در پیکره گوش‌زد شود. میزان تغییرات پس از این اصلاحات نیز در جدول ۹.۲ نشان داده شده است.

۲.۴.۲ فعل‌ها در پیکرهٔ وابستگی

همان طور که اشاره شد هدف اصلی از حجمی که برای این پیکره در نظر گرفته‌ایم، پوشش عمدهٔ فعل‌های فارسی است. به همین خاطر با یاری جستن از گروه داده‌گزینی و با استفاده از موتورهای جستجو (مانند گوگل و بینگ)، سعی بر آن داشتیم که از همهٔ فعل‌های درون فرهنگ ظرفیت در پیکره گنجانده باشیم. گرچه به دلیل نبود نمونه جمله برای همهٔ فعل‌ها این سعی کامل نبود ولی درصد بالایی از افعال زبان فارسی را پوشش داد. بدین وسیله یک بستر اولیهٔ مناسب برای پیکره‌های معنایی نیز خواهیم داشت. در جدول ۱۰.۲ آماره‌های مربوط به فعل‌های پیکره نشان داده شده است. لازم به ذکر است که این آماره‌ها تنها از روی شمارش فعل‌های روساختی بوده، فعل‌هایی که بر اثر حذف یا همپایگی در پیکره وجود دارند، مورد شمارش قرار نگرفته‌اند.

XePersian

اختصار	توضیح	اختصار	توضیح
ACL	متمم بندی صفت	ADV	قید
ADVC	متمم قیدی فعل	AJCONJ	همپایهٔ صفت
AJPP	متمم حرف اضافه‌ای صفت	AJUCL	بند افزودهٔ فعل
APOSTMOD	وابستهٔ پسین صفت	APP	بدل
APREMOD	وابستهٔ پیشین صفت	AVCONJ	همپایهٔ قید
COMPPP	حرف اضافهٔ تفضیلی	ENC	فعل‌یار پی‌بستی
LVP	جزء همکرد	MESU	ممیز
MOS	مسند	MOZ	مضافٌ‌الیه
NADV	قید اسم	NCL	بند اسم
NCONJ	همپایهٔ اسم	NE	اسم‌یار
NEZ	متمم نشانهٔ اضافه‌ای صفت	NPOSTMOD	صفت پسین اسم
NPP	حرف اضافهٔ اسم	NPREMOD	صفت پیشین اسم
NPRT	جزء اسمی	NVE	فعل‌یار
OBJ	مفعول	OBJ۲	مفعول دوم
PARCL	بند وصفی	PART	افزودهٔ پرسشی فعل
PCONJ	همپایهٔ حرف اضافه	POSDEP	وابستهٔ پسین
PRD	گزاره	PREDEP	وابستهٔ پیشین
PROG	مستمرساز	PUNC	علامت نگارشی
ROOT	ریشهٔ جمله	SBJ	فاعل
TAM	تمیز	VCL	بند متممی فعل
VCONJ	همپایهٔ فعل	VPP	مفعول حرف اضافه‌ای
VPRT	حرف اضافهٔ فعلی		

جدول ۱.۲: روابط وابستگی در پیکرهٔ وابستگی زبان فارسی

ویژگی‌های ساخت‌واژی و برچسب اجزای سخن در پیکرهٔ وابستگی				
برچسب اجزای سخن	برچسب ریز	شخص (Person)	شخص (Number)	وجه نمود زمان (TMA)
ADJ	AJP AJCM AJSUP			
ADR	PRADR POSADR			
ADV	SADV			
CONJ				
IDEN				
N	ANM IANM		SING PLUR	
PART				
POSNUM				
POSTP				
PR	SEPER JOPER DEMON INTG CREFX UCREFX RECPR	۱ ۲ ۳	SING PLUR	
PREM	EXAJ QUAJ DEMAJ AMBAJ			
PRENUM				
PREP				
PSUS				
PUNC				
V	ACT PAS MOD	۱ ۲ ۳	SING PLUR	ر.ک. به جدول ۴.۲
SUBR				

جدول ۲.۲: ویژگی‌های ساخت‌واژی موجود در پیکرهٔ وابستگی زبان فارسی

اختصار	توضیح	اختصار	توضیح
ACT	معلوم	ADJ	صفت
ADR	نقش‌نمای ندا	ADV	قید
AJCM	صفت تفضیلی	AJP	صفت مطلق
AJSUP	صفت عالی	AMBAJ	صفت مبهم
ANM	جاندار	CONJ	نقش‌نمای همپایگی
CREFX	بازتابی مشترک	DEMAJ	صفت اشاره
DEMON	اشاره	EXAJ	صفت تعجبی
IANM	بی‌جان	IDEN	شاخص
INTG	پرسشی	JOPER	شخصی پیوسته
MOD	وجهی	N	اسم
PART	جزء دستوری	PAS	مجهول
PLUR	جمع	POSADR	نقش‌نمای ندا پسین
POSNUM	صفت شمارشی پسین	POSTP	حرف اضافهٔ پسین
PR	ضمیر	PRADR	نقش‌نمای ندا پیشین
PREM	پیش‌توصیف‌گر	PRENUM	صفت شمارشی پیشین
PREP	حرف اضافهٔ پیشین	PSUS	شبه‌جمله
PUNC	علامت نگارشی	QUAJ	صفت پرسشی
RECPR	ضمیر متقابل	SADV	قید مختص
SEPER	ضمیر شخصی جدا	SING	مفرد
SUBR	نقش‌نمای وابستگی	UCREFX	ضمیر بازتابی غیرمشترک
V	فعل		

جدول ۳.۲: اختصارات موجود در برچسب‌های اجزای سخن و ویژگی‌های ساخت‌واژی درج‌شده در جدول ۲.۲

وجه نمود زمان	اختصار	مثال (خوردن)
حال امری	HA	بخور
آیندهٔ اخباری	AY	خواهم خورد
گذشتهٔ نقلی استمراری اخباری	GNES	می‌خورده‌ام
گذشتهٔ بعید استمراری اخباری	GBES	می‌خورده بودم
گذشتهٔ استمراری اخباری	GES	می‌خوردم
گذشتهٔ نقلی اخباری	GN	خورده‌ام
گذشتهٔ بعید اخباری	GB	خورده بودم
حال اخباری	H	می‌خورم
گذشتهٔ سادهٔ اخباری	GS	خوردم
گذشتهٔ بعید استمراری التزامی	GBESE	می‌خورده بوده باشم
گذشتهٔ استمراری التزامی	GESEL	می‌خورده باشم
گذشتهٔ بعید التزامی	GBEL	خورده بوده باشم
حال التزامی	HEL	بخورم
گذشتهٔ التزامی	GEL	خورده باشم

جدول ۴.۲: وجه-نمود-زمان در فعل‌های زبان فارسی

تعداد کل واژه‌ها	۴۹۸،۰۸۱
میانگین طول هر جمله	۱۶/۶۱
تعداد واژه‌های منحصر به فرد	۳۷۶۱۸
تعداد بن‌واژه‌ها منحصر به فرد	۲۲۰۶۴

جدول ۵.۲: آماره‌های فراوانی واژه‌ها در پیکرهٔ وابستگی زبان فارسی

برچسب اجزای سخن	تعداد حضور	درصد حضور
PREP	۵۷۹۶۰	۱۱/۶۴
N	۱۹۵۴۳۱	۳۹/۲۴
PUNC	۴۵۵۹۷	۹/۱۵
ADJ	۳۶۸۳۳	۷/۳۹
PREM	۱۰۷۹۵	۲/۱۷
CONJ	۲۲۸۴۵	۴/۵۹
V	۶۲۸۷۶	۱۲/۶۲
SUBR	۱۴۱۸۲	۲/۸۵
PRENUM	۶۱۹۵	۱/۲۴
POSTP	۱۵۱۸۹	۳/۰۵
PR	۱۸۹۷۳	۳/۸۱
ADR	۱۸۲	۰/۰۴
ADV	۸۵۵۵	۱/۷۲
IDEN	۸۵۱	۰/۱۷
PSUS	۵۲۶	۰/۱۱
POSNUM	۵۷۱	۰/۱۱
PART	۵۲۰	۰/۱۰

جدول ۶.۲: فراوانی برچسب‌های اجزای سخن در پیکرهٔ وابستگی

اختصار رابطه	تعداد حضور	درصد حضور	اختصار رابطه	تعداد حضور	درصد حضور
ADV	۳۶۲۲۴	۷/۲۷	VPP	۸۳۸۴	۱/۶۸
POSDEP	۷۵۱۰۷	۱۵/۰۸	VCL	۸۳۵۲	۱/۶۸
MOZ	۵۰۶۳۹	۱۰/۱۷	PARCL	۵۳۲	۰/۱۱
NPP	۱۸۲۸۷	۳/۶۷	AVCONJ	۶۰	۰/۰۱
PUNC	۴۵۹۷۵	۹/۲۳	ACL	۴۷۶	۰/۱۰
NPOSTMOD	۲۳۲۸۸	۴/۶۸	AJCONJ	۲۱۴۸	۰/۴۳
NPREMOD	۱۷۰۵۱	۳/۴۲	AJPP	۲۴۴۵	۰/۴۹
SBJ	۲۸۶۹۹	۵/۷۶	MESU	۷۴۳	۰/۱۵
NCONJ	۱۱۲۳۰	۲/۲۵	COMPPP	۵۵۴	۰/۱۱
NVE	۳۱۷۵۶	۶/۳۸	ADVC	۶۸۱	۰/۱۴
ROOT	۲۹۹۸۱	۶/۰۲	NE	۴۲۶	۰/۰۹
AJUCL	۳۹۶۹	۰/۸۰	APREMOD	۶۰۶	۰/۱۲
MOS	۱۰۸۱۹	۲/۱۷	PCONJ	۵۸۶	۰/۱۲
NEZ	۱۲۱۵	۰/۲۴	PART	۳۹۱	۰/۰۸
PRD	۱۴۲۰۷	۲/۸۵	NADV	۱۷	۰/۰۰
NCL	۹۴۸۳	۱/۹۰	LVP	۴۴۱	۰/۰۹
PREDEP	۲۸۳۱۴	۵/۶۸	NPRT	۴۲	۰/۰۱
VCONJ	۱۱۰۰۱	۲/۲۱	PROG	۱۶۳	۰/۰۳
APP	۱۲۰۷	۰/۲۴	ENC	۱۰۲	۰/۰۲
OBJ	۲۰۶۰۵	۴/۱۴	OBJ۲	۳۰	۰/۰۱
TAM	۷۵۳	۰/۱۵	APOSTMOD	۴۳	۰/۰۱
VPRT	۱۰۴۷	۰/۲۱	ADJADV	۲	۰/۰۰

جدول ۷.۲: فراوانی روابط وابستگی در پیکرهٔ وابستگی

توافق در رابطهٔ وابستگی	٪۹۷/۰۶
توافق در رابطهٔ وابستگی برچسب‌دار	٪۹۵/۳۲
توافق در برچسب اجزای سخن	٪۹۸/۹۳

جدول ۸.۲: آماره‌های مربوط به میزان توافق

تغییر در رابطهٔ وابستگی	٪۴/۹۱
تغییر در رابطهٔ وابستگی برچسب‌دار	٪۶/۲۹
تغییر در برچسب اجزای سخن	٪۴/۲۳

جدول ۹.۲: آماره‌های مربوط به میزان اصلاحات پس از بازبینی پیکره

تعداد فعل در پیکره	۶۰۵۷۹
تعداد فعل منحصر به فرد	۴۷۸۲
میانگین حضور هر فعل	۱۲/۶۷
انحراف معیار حضور هر فعل	۹۱/۴۳

جدول ۱۰.۲: آماره‌های مربوط به فعل‌های موجود در پیکرهٔ وابستگی زبان فارسی

کتاب‌نامه

[انگل(۲۰۰۲)] Engel, U. (2002), *Kurze Grammatik der deutschen Sprache*, Iudicium Verlge.

[هلبیگ و شنکل(۱۹۹۱)] Helbig, G. and Schenkel, W. (1991), *Wörterbuch zur Valenz und Distribution deutscher Verben*, M. Niemeyer.

[کوبلر و دیگران(۲۰۰۹)] Kübler, S., McDonald, R., and Nivre, J. (2009), *Dependency parsing*, vol. 1 of *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers.

[مارکوس و دیگران(۱۹۹۳)] Marcus, M., Marcinkiewicz, M., and Santorini, B. (1993), "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, 19, 313–330.

[نیلسون و دیگران(۲۰۰۷)] Nilsson, J., Riedel, S., and Yuret, D. (2007), "The CoNLL 2007 shared task on dependency parsing," in *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pp. 915–932.

[تنی‌یر(۱۹۵۳)] Tesnière, L. (1953), *Esquisse d'une Syntaxe structurale*, Klincksieck.

[تنی‌یر(۱۹۵۹)] — (1959), *Éléments de syntaxe structurale*, Klincksieck.

[تنی‌یر(۱۹۸۰)] — (1980), *Grundgzüge der Strukturalen Syntax*, Klettcotta.

[رسولی(۱۳۸۹)] رسولی، م.ص. (۱۳۸۹)، "تجزیهٔ نحوی با استفاده از دستور وابستگی،" گزارش طرح تحقیقی، مرکز تحقیقات کامپیوتری علوم اسلامی (معاونت تهران).

[طبیب‌زاده(۱۳۸۵)] طبیب‌زاده، ا. (۱۳۸۵)، *ظرفیت فعل و ساخت‌های بنیادین جمله در زبان فارسی امروز*. نشر مرکز.