تحلیل زبان آذری مبتنی بر تقطیع به کمک دستور زبان پیوندی

مریم عرب‌زاده[1]
کارشناسی ارشد زبان‌شناسی همگانی
دانشگاه آزاد اسلامی؛ واحد اهر

سیدمهدی عراقی[2]
دکتری آموزش زبان انگلیسی؛ استادیار مدعو
دانشگاه آزاد اسلامی؛ واحد اهر

دریافت: ۱۳۹۱/۰۶/۱۴ پذیرش: ۱۳۹۲/۰۷/۱۵

چکیده: انواع مختلفی از نظریه ها دربارهٔ مشکل تقطیع نحوی و ایجاد دستور زبان‌های مربوط به زبان‌های طبیعی وجود دارند. این مقاله یک دستور زبان نحوی بر مبنای صورت‌بندی دستور زبان رابطه‌ای برای زبان ترکی (آذری) که جزو زبان‌های پیوندی می‌باشد، ارائه می کند. در صورت‌بندی دستور زبان رابطه‌ای، کلمات یک جمله بر طبق نقش‌های نحوی که دارند به یکدیگر متصل می گردند. زبان ترکی (آذری) دارای ساخت واژه‌ای صرفی و اشتقاقی پیچیده می‌باشد و تکواژهای صرفی و اشتقاقی، نقش‌های نحوی مهمی در جملات بازی می کنند. به منظور طراحی نرم‌افزار دستور زبان رابطه‌ای برای زبان ترکی (آذری)، بخش‌های لغوی در بازنمایی ساخت واژه‌ای کلمات ترکی حذف شده‌اند و یال‌ها بر مبنای نشانه‌های ادات سخن و تکواژهای تصریفی در کلمات ایجاد می شوند. کلمات مشتق شده توسط مرزهای اشتقاقی از دیگر کلمات جدا می‌شوند. یک صورت‌بندی منحصربه‌فرد دستور زبان رابطه‌ای تطابق‌یافته با زبان ترکی، دارای انعطاف‌پذیری لازم برای ایجاد ساخت‌های اتصال می‌باشد و در نهایت با استفاده از زبان برنامه‌نویسی دلفی، نرم‌افزار دستور زبان رابطه‌ای برای زبان آذری طراحی و اجرا گردید و سپس با انتخاب ۲۵۰ جملهٔ تصادفی، این نرم‌افزار مورد بررسی و آزمون قرار گرفت. برای ۸۴/۳۱٪ از جملات، نتیجهٔ تقطیع کننده شامل تقطیع کننده‌های صحیح بود.

کلیدواژه‌ها: تقطیع کننده؛ دستور زبان رابطه‌ای؛ پردازش زبان طبیعی؛ تحلیل ساختواژه‌ای؛ زبان برنامه‌نویسی دلفی

دانش و مدیریت اطلاعات

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، ISC و SCOPUS
http://jipm.irandoc.ac.ir
دورهٔ ۲۹ | شمارهٔ ۳ | ص‌ص۸۴۵-۸۷۰
بهار ۱۳۹۳

نوع مقاله: پژوهشی

۱. پدیدآور رابط
ma25286@yahoo.com
2. m_araghi@pnu.ac.ir

۱. مقدمه

انواعی از نظریه‌ها دربارهٔ تقطیع نحوی و ایجاد دستور زبان‌های مربوط به زبان‌های طبیعی وجود دارند. یکی از این صورت‌گرایی[1]ها عبارت است از دستور زبان مقوله‌ای[2] که توسط اصول ترکیب به‌وجود می‌آید (Sag and Wasow 1999). بر طبق این صورت‌گرایی، اجزای نحوی به عنوان توابع یا در یک رابطهٔ دلیل و مدلولی با هم ترکیب می‌شوند (Chomsky 1981). علاوه بر دستور زبان‌های مقوله‌ای، دو گونهٔ دیگر از دستور زبان‌ها وجود دارند که شامل دستور زبان‌های ساخت‌گروهی[3] و دستور زبان‌های وابستگی[4] می‌باشند (Schneider 1998). دستور زبان‌های ساخت‌گروهی، سازه‌ها را در نمودار درختی سازماندهی می‌کنند (Gazdar et al. 1985). اما دستور زبان‌های وابستگی روابط ساده‌ای بین جفت کلمات به‌وجود می‌آورند (Melçuk 1998). چون دستور زبان‌های وابستگی به‌ترتیب خاص کلمات اهمیت نمی‌دهند؛ در نتیجه این‌گونه از دستور زبان‌ها برای زبان‌هایی که ترتیب آزاد کلمات را دارند مناسب هستند، مانند زبان‌های ترکی و چک (دبیر مقدم ۱۳۸۷). دستور زبان رابطه‌ای مشابه دستور زبان وابستگی می‌باشد (Sleator and Temperley 1993).

اما در دستور زبان رابطه‌ای[5] علاوه بر نبود یک رابطهٔ وابسته به رأس، دارای یک گراف جهت‌دار بین کلمات نیز می‌باشد (Antworth 1999). برخی تحقیقات در زمینهٔ تحلیل‌های محاسباتی نحو ترکی صورت گرفته است. یکی از این تحقیقات، دستور زبان نقشی/ لغوی[6] می‌باشد (Güngördü 1993). علاوه بر این، دستور زبان "ATN" نیز برای زبان ترکی وجود دارد (Demir 1993). دستور زبان دیگری برای زبان ترکی بر مبنای صورت‌بندی "HPSG"[7] وجود دارد (Şehitoğlu 1996). نحو ترکی نیز با چشم‌انداز دستور زبان وابستگی مورد بررسی قرار گرفته است. «اوفلازر» یک تقطیع‌کنندهٔ وابستگی با استفاده از یک شیوهٔ محدود بسط‌یافته ارائه می‌کند (Oflazer 1999). این تقطیع‌کننده برای

1. formalism
2. categorical grammar
3. phrase structure grammars
4. dependency grammars
5. link grammar
6. lexical functional grammar
7. Head driven Phrase Structure Grammar

ایجاد دستور زبان درخت افزایشی[1] ترکی به‌کار می‌رود (Akbik 2009).

از درخت افزایشی وابستگی زبان ترکی، برای آموزش و آزمودن تقطیع‌کننده وابستگی زبان ترکی استفاده شده است (Eryiğit and Oflazer 2006). همچنین کارهای دیگری در زمینهٔ دستور زبان‌های مقوله‌ای برای زبان ترکی صورت گرفته‌اند (Hoffman 1995).

تحلیل نحوی در لایهٔ زیرین بسیاری از کاربردهای زبان طبیعی قرار گرفته است و بنابراین گام بسیار مهمی برای هر زبان می‌باشد (فیلی ۱۳۸۲). کارهای متعددی در زمینهٔ تحلیل‌های محاسباتی زبان ترکی صورت گرفته، اما این مقاله دربارهٔ اولین نرم‌افزار طراحی‌شده برای دستور زبان رابطه‌ای زبان ترکی آذری است که جزو زبان‌های پیوندی می‌باشد. در این کار، ساختار واژگانی‌شدهٔ صورت‌گرایی دستور زبان رابطه‌ای برای بیان نقش‌های نحوی صورت‌های مشتق‌شدهٔ طبقه میانی کلمات در زبانی که دارای ساخت واژه‌ای بسیار خلاقانهٔ اشتقاقی و تصریفی است، مورد استفاده قرار می‌گیرد. این هدف زمانی به‌دست می‌آید که با هر یک از این صورت‌های اشتقاقی طبقهٔ میانی، به‌صورت یک کلمهٔ مجزا رفتار شود. صورت‌بندی دستور زبان رابطه‌ای تطابق‌یافته می‌تواند برای پیشرفت دستور زبان‌های رابطه‌ای دیگر زبان‌هایی که ساخت‌واژهٔ خلاقانه دارند نیز به‌کار رود.

امروزه یکی از مهم‌ترین موضوعات در زمینهٔ مدیریت اطلاعات، سازماندهی و کنترل ساختار، ترکیب و پردازش حجم عظیم اطلاعات است و در چنین شرایطی بازیابی و استخراج اطلاعات اهمیت ویژه‌ای می‌یابد، زیرا به کاربر کمک می‌کند تا اطلاعات مورد نظر را با بالاترین سرعت از میان حجم انبوهی از اطلاعات ساختارنیافته به‌دست آورد (عبدالله‌زاده ۱۳۸۳). گسترش فناوری‌های اطلاعاتی و ارتباطی و افزایش حجم متون و مدارک تخصصی در رسانه‌ها و قالب‌های مختلف از یک‌سو و نیاز کاربران و متخصصان به بازیابی اطلاعات مرتبط در کمترین زمان از سوی دیگر، کارشناسان را به فکر ایجاد تغییر و تحول در زمینهٔ نحوهٔ دسترسی و پردازش اطلاعات انداخته است. از این‌رو ایجاد نرم‌افزارهای پردازش زبان طبیعی در سال‌های اخیر، امکانات و قابلیت‌های فراوانی برای رفع مشکلات افزایش دسترسی به اطلاعات در سیستم‌های اطلاعاتی ایجاد کرده است.

---

1. Tree Adjoining Grammar

نرم‌افزار دستور زبان رابطه‌ای این امکان را برای کاربر فراهم می‌آورد که با توجه به حجم عظیم واژه‌ها و کلمات، در کوتاه‌ترین زمان ممکن، جملات را پردازش و تقطیع نماید و با یک عملکرد بهینه و با کمترین ابهام، اطلاعات مورد نظر در مورد تمامی کلمات را به کاربر ارائه می‌دهد (Magerman 1993).

۲. پیشینهٔ پژوهش

در ادامه به بررسی نحوهٔ عملکرد پژوهش‌هایی که از طریق دستور زبان رابطه‌ای بر روی دیگر زبان‌ها صورت گرفته است می‌پردازیم.

۲-۱. حوزهٔ زبان فارسی

«سجادی» و «عبدالله‌زاده» برای اولین بار به بررسی نحوهٔ عملکرد دستور زبان رابطه‌ای برای زبان فارسی پرداختند. در این تحقیق آنها با شروع از روابط اصلی و اساسی در زبان فارسی، توسعهٔ کار را مبتنی بر این روابط قرار دادند و در هر مرحله تعداد این روابط را افزایش دادند تا حدی که مجموعه قواعدی برای متون رسمی نوشتاری و ساخت‌های ابتدایی فراهم آوردند. آنان برای اطمینان از انسجام مجموعه، قواعد را در هر مرحله به کمک تقطیع‌گر دستور زبان رابطه‌ای و روی یک بستر مناسب آزمایش کردند (سجادی و عبدالله‌زاده ۱۳۸۷).

در تحقیقی دیگر، «سجادی» و «همایون‌پور» به بررسی ساختاروندی زبان فارسی پرداختند. آنها با معرفی سیستمی که قادر به تجزیهٔ کلمه به همه واج‌های تشکیل‌دهندهٔ آن در زبان فارسی و انعکاس ویژگی‌های کلمهٔ مورد نظر می‌باشد، تکواژشناختی زبان فارسی را به‌طور مفصل و از دیدگاه محاسباتی بررسی نمودند. در سیستم آنها بر خلاف سیستم‌های موجود، همهٔ واج‌های صرفی و اشتقاقی مورد پوشش قرار داده شدند، که به کمک روش‌های معمولی امکان‌پذیر نمی‌باشد. با توجه به این که دستور زبان‌های رابطه‌ای فاقد ویژگی هستند و ویژگی، خاصیت لازم هر تحلیلگر تکواژشناختی است(فرخ ۱۳۸۱، ۷۸)، آنها در تحقیق خود روشی برای بازنمایی و استخراج ویژگی‌ها در این صورت‌گرایی ارائه دادند. بر خلاف تحقیق‌های مشابه، همهٔ دانش سیستم ارائه شد تا دیگر افراد بتوانند در تحقیقات مشابه از آن استفاده کنند. توسعهٔ سیستم آنها برای پذیرش موارد جدید به‌سادگی

صورت می‌گیرد و بر خلاف پیچیدگی‌های فراوان سیستم‌وندی به‌خصوص وندهای اشتقاقی، آن‌ها توانستند به کمک قواعد زبانی، وندها را مدیریت کنند. پیچیده‌ترین بخش تکواژشناختی زبان فارسی، مربوط به وندهای اشتقاقی می‌باشد که استفاده از یک صورت‌گرایی مستقل از متن را ضروری می‌کند (Assi and Abdolhosseini 2000). بزرگ‌ترین مشکلی که پوشش همهٔ این قواعد ایجاد می‌کند، ابهام‌های بسیار زیاد است که برخی از آن به کمک تحلیل‌های نحوی و معنایی رفع شده بود و رخداد تعدادی از آن‌ها نیز با اعمال ارزش بر روی پیوندها قابل رفع به نظر می‌رسید. در این مقاله ساختارهای فعلی، غیرفعلی، صرفی و اشتقاقی به‌صورت کامل بررسی شده و واژه‌نامهٔ کامل برای پرهیز از هر گونه ابهام ذکر شده است (سجادی و همایون‌پور ۱۳۸۷).

«لوندزدیل» و «دهداری» در پژوهش خود به مطالعهٔ تقطیع‌کنندهٔ نحوی زبان فارسی پرداختند (Dehdari and Londsdale 2005).

تقطیع‌کننده مبتنی بر دستور رابطه‌ای و دستور وابستگی است (Sleator and Temperley 1993). در ابتدا تکواژهای تصریفی فردی را توسط قسمت تکواژشناختی[1] تجزیه نمودند و سپس این تکواژها را از لحاظ نحوی در یک حالت کارآمد به یکدیگر متصل کردند. هر مؤلفه با جزئیات و با تشریح جایگاه کنونی سیستم و کاربردهای ممکن ارائه شده است. سیستم مقیاسی که آنها توصیف کرده‌اند موتورهای تکواژشناختی کارآمد را با تقطیع‌کنندهٔ نحوی تنومند یکپارچه می‌کند. این امر مهم است زیرا بسیاری از مشکلات در پردازش زبان فارسی، همانند ابهام در قوانین املایی و ساخت‌واژه، می‌توانند قبل از رسیدن تقطیع‌کننده در قسمت تکواژشناختی حل شوند. کاربرد سیستم آنها بیشتر در آموزش زبان، بازیابی و استخراج اطلاعات، و لغت‌نامه‌های برخط می‌باشد.

«سجادی» و «بروجردی» در مقالهٔ خود فرمالیسم جدیدی برای تحلیل نحوی زبان طبیعی معرفی نمودند که حاصل افزودن پارادایم‌های همسان‌سازی به دستور زبان‌های رابطه‌ای می‌باشد (سجادی و بروجردی ۱۳۸۵). فرمالیسم ارائه‌شده مزایای بسیاری دارد که از جمله می‌توان به توان بیشتر و پیچیدگی کمتر لغت‌نامه اشاره کرد. همچنین توان توصیفی دانش در این فرمالیسم، سازمان‌یافته‌تر و قابل استفاده‌تر می‌باشد. این فرمالیسم همچنین چارچوبی را برای تحلیل تکواژشناختی (هم استفاده و هم طراحی) یا هر سازوکار

1. Perstem or PC-kimmo

پیش‌تحلیل دیگری مهیا می‌کند. بیشتر مثال‌های مورد بررسی از زبان فارسی، و استخراج‌شده از همین سیستم می‌باشد و این سیستم نسبت به سیستم‌های قبلی کاراتر و ساده‌تر می‌باشد.

«کشاورزی» تقطیع‌گری برای تقطیع جملات سادهٔ خبری بر اساس دستور ساخت گروهی هسته‌بنیان و الگوریتمی بالا به پایین ارائه داده است. این تقطیع‌گر قادر به شناسایی گروه اسمی، شامل وابستهٔ پیشین اسم، گروه اسمی هم‌پایه، گروه پیش‌اضافه، گروه پس‌اضافه، و گروه فعلی است. تقطیع‌گر علاوه بر این، ساده یا ترکیبی‌بودن گروه فعلی را تشخیص می‌دهد؛ از میان ترکیب‌ها، فعل مرکب و پیشوندی را به اجزای آنها تقطیع می‌کند. وی قواعد ساخت ۴۵۰ جمله و واژگان را برای تقطیع به تقطیع‌گر داده است. تقطیع‌گر پس از دریافت جملهٔ ورودی، درختی ارائه می‌دهد که ساخت نحوی جمله را در شش مرحله مشخص می‌کند (کشاورزی ۱۳۸۷).

«رضایی» در پایان‌نامهٔ دکتری، نتیجهٔ سه تحقیق خود را منعکس کرده است. در ابتدا وی برای تقطیع جملات سادهٔ زبان فارسی، سیستمی مبتنی بر شبکهٔ انتقالی برافزوده[1] طراحی کرد (Rezaei 1999). این تقطیع‌گر توالی‌های ممکن درون‌بند ساده را تبیین می‌کند، اما قادر به تقطیع‌بندهای درونه‌ای نیست. بنابر تحقیق بعدی وی، تقطیع‌گر قلب نحوی را نیز در بر می‌گیرد. ایشان در تحقیق آخر، پدیده‌هایی از قبیل برجسته‌سازی و جابه‌جاییِ بندهای متمم به آخر جمله را مطرح می‌کند. پدیده‌های زبانی، در دو تقطیع‌گر آخر وی، در قالب نظریهٔ حاکمیت و مرجع‌گزینی توصیف می‌شود.

۲-۲. حوزهٔ زبان ترکی

انواع مختلفی از نظریه‌ها دربارهٔ مشکل تقطیع نحوی و ایجاد دستور زبان‌های مربوط به زبان‌های طبیعی وجود دارند. «چیچکلی» و «ایستک» در پژوهش خود یک دستور زبان نحوی بر مبنای صورت‌بندی دستور زبان رابطه‌ای برای زبان ترکی که جزو زبان‌های پیوندی می‌باشد ارائه کردند (Cicekli and Istek 2006). در صورت‌بندی دستور زبان رابطه‌ای، کلمات یک جمله بر طبق نقش‌های نحوی که دارند به یکدیگر متصل می‌گردند (Mollá et al. 2000). زبان ترکی دارای تکواژشناختی صرفی و اشتقاقی پیچیده می‌باشد و

1. Augmented Transition Network

تکواژهای صرفی و اشتقاقی، نقش‌های نحوی مهمی در جملات بازی می‌کنند. به منظور طراحی نرم‌افزار دستور زبان رابطه‌ای برای زبان ترکی، آن‌ها بخش‌های لغوی در بازنمایی تکواژشناختی کلمات ترکی را حذف کردند و یال‌ها را بر مبنای نشانه‌های ادات سخن و تکواژهای تصریفی در کلمات ایجاد کردند و کلمات مشتق‌شده را توسط مرزهای اشتقاقی از دیگر کلمات جدا نمودند و یک صورت‌بندی منحصربه‌فرد دستورزبان رابطه‌ای تطابق‌یافته با زبان ترکی که دارای انعطاف‌پذیری لازم برای ایجاد ساخت‌های اتصال است ایجاد نمودند. اما چون اسم‌ها در زیرشاخهٔ زمان، مکان و موضوع تقسیم‌بندی نشده‌اند، این فرایند منجر به پدید آمدن گروه‌های اسمی صفتی نامحدود نادرست شده است. علاوه بر آن برخی از جملات، متشکل از کلماتی با ترتیب تکواژهای اشتقاقی خیلی پیچیده می‌باشد؛ یعنی بسیاری از صورت‌های اشتقاقی طبقهٔ میانی که باعث ایجاد شماری از یال‌های ممکن بین این صورت‌های اشتقاقی طبقهٔ میانی می‌شوند افزایش یافته‌اند.

۲ –۳. حوزهٔ زبان‌های خارجی

«کولبر» در پایان‌نامهٔ خود الگوریتمی برای یادگیری دستور زبان رابطه‌ای زبان آلمانی ارائه داده و مشکل پراکندگی اطلاعات را از طریق به‌کارگیری تجزیه و تقطیع‌های جزئی و قطعات دستوری در دسترس، حل کرده است (Kübler 1998). به گفتهٔ وی، از آنجا که اسلیتور و تمپرلی[1] به‌طور اکید بر یال‌های مکانی تأکید می‌ورزند (یعنی یال‌ها باید کلمات را به کلمات همسایه که نزدیکشان می‌باشد وصل کنند). گرچه ممکن است این حالت در زبان‌شناسی مورد تأیید نباشد. بنابراین با توجه به این که توافق بین کلمات در زبان آلمانی بسیار وسیع‌تر از زبان انگلیسی می‌باشد، پس ضروری است که به عوض وصل شدن به نزدیک‌ترین کلمهٔ همسایه، کلمات بر طبق ملزومات توافقی خود به یکدیگر متصل گردند. با این حال این دیدگاه منجر به ایجاد یال‌های بسیار طولانی شده است. همچنین چون ترتیب کلمات در زبان آلمانی آزاد است (یعنی مکان کلمات، آزاد و متغیر می‌باشد)، این پدیده باعث به‌وجود آمدن برچسب‌های متفاوتی که گونهٔ یکسانی از یال‌ها اما با ترتیب متفاوت را توصیف می‌کنند، می‌شود. از این‌رو نویسنده ایدهٔ کنترل را معرفی می‌کند؛ یعنی هر یال یا کنترل می‌شود یا کنترل می‌کند.

1. Sleator and Temperley

۳. مبانی نظری

۳-۱. دستور زبان رابطه‌ای

دستور زبان رابطه‌ای یک سیستم دستور زبانی رسمی تعریف‌شده توسط اسلیتور و تمپرلی در سال ۱۹۹۱ می‌باشد که همراه با هم، الگوریتم‌های برنامه‌ریزی کارآمد پویایی برای پردازش دستور زبان‌ها بر مبنای صورت‌گرایی و ایجاد یک دستور زبان رابطه‌ای به منظور پوشش وسیع زبان انگلیسی ابداع کردند (Sleator and Temperley 1991). این صورت‌گرایی، بر خلاف دستور زبان‌های بافت آزاد، لغوی است و نه از سازه‌ها استفاده می‌کند و نه از دسته‌ها (Pollard and Sag 1994). در حقیقت دستور رابطه‌ای می‌تواند تحت طبقهٔ دستورهای وابستگی طبقه‌بندی شود. در این صورت‌گرایی، زبان توسط دستور زبانی تعریف می‌شود که شامل کلمات زبان و ملزومات ارتباطی‌شان باشد (Jurafsky and Martin 2000). جملهٔ داده‌شده توسط سیستم، زمانی پذیرش می‌شود که ملزومات ارتباطی[1] همهٔ کلمات در جمله ارضاء شود[2]؛ هیچ یک از یال‌ها[3]ی بین کلمات، همدیگر را قطع نکنند[4]؛ و حداکثر یک یال بین جفت کلمات وجود داشته باشد[5]. مجموعهٔ یال‌های بین کلمات یک جمله که توسط سیستم پذیرش می‌شود، مجموعهٔ اتصال یا حلقه‌های زنجیر[6] نامیده می‌شود. دستور زبان در یک فایل لغت‌نامه تعریف شده و هر یک از ملزومات ارتباطی کلمات با واژه‌های اتصال‌گر[7]، در فایل لغت‌نامه بیان شده است. زمانی که توالی کلمات پذیرفته می‌شوند، تمامی یال‌ها بالای کلمات کشیده می‌شوند.

مثلاً ملزومات ارتباطی کلمات داده‌شدهٔ زیر در مقابل آنها آمده است:

apardi (برد): O- & S-;

şabnam (شبنم): S+;

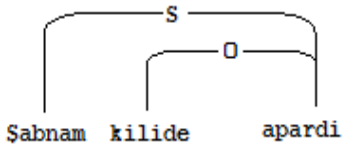
kilide (کلید را): O+;

در اینجا فعل "apardi" دارای دو ملزوم ارتباطی از سمت چپ می‌باشد: یکی "S"

1. linking requirements
2. connectivity
3. links
4. planarity
5. exclusion
6. linkage
7. connectors

(فاعل) و دیگری"O" (مفعول). از طرف دیگر، اسم "şabnam" به‌خاطر اتصال‌گر "+S" نیاز به اتصال از سمت راست به یک کلمه دارد و اسم "kilide" به‌خاطر اتصال‌گر "+O" باید از سمت راست به کلمه‌ای متصل گردد. چون کلمات "apardi" و "şabnam" دارای اتصال‌گر یکسان"S" هستند (یعنی دارای ملزومات ارتباطی یکسان با علامت متفاوت می‌باشند) پس می‌توانند از طریق رابط"S" به‌هم وصل شوند. این موقعیت مشابه می‌تواند بین کلمات"kilide" و "apardi" به‌خاطر اتصال‌گر "O" نیز رخ دهد. بنابراین اگر این کلمات به شیوهٔ مندرج در شکل ۱ به یکدیگر متصل شوند، تمام ملزومات ارتباطی این کلمات تأمین می‌شود.

Şabnam kilide apardi (۱)



شکل ۱. نتیجهٔ تقطیع جملهٔ ۱

۳-۲. نحو ترکی آذری

در زبان آذری ترتیب اصلی کلمه "SOV" (subject- object – verb) می‌باشد؛ اما ترتیب سازه بر حسب متن مکالمه ممکن است به‌راحتی تغییر یابد و دستور زبان زبان آذری محدودیتی در به‌وجودآمدن انواع ترتیب سازه‌ها در درون جمله ندارد. بنابراین هر شش حالت ترتیب کلمات ("SOV"، "OSV"، "VSO"، "SVO"، "OVS"، "VOS") می‌توانند به‌وجود آیند.

زبان آذری هسته پایانی است. از این‌رو در یک جملهٔ عادی زبان آذری، معرف کلمه همیشه در سمت چپ قرار می‌گیرد و کلمه‌ای که آن را مشخص می‌کند در سمت راست قرار دارد. به این دلیل ملزومات ارتباطی سمت چپ یک کلمه، مشابه با معرفش می‌باشد و ملزومات ارتباطی سمت راست کلمه نیز مشابه با کلمه‌ای است که آن را توصیف می‌کند. مثال زیر را بررسی کنیم.

Gözal gəz (دختر زیبا)

در این عبارت، صفت "gözal" (زیبا) معرف اسم "gəz" (دختر) می‌باشد.

همانند دیگر زبان‌های آلتایی، زبان ترکی (آذری) نیز پیوندی می‌باشد. در زبان آذری پسوندهای تصریفی نقش‌های دستور زبانی دارند. علاوه بر آن، کلمات ممکن است چندین پسوند اشتقاقی که ادات سخن‌شان را تغییر می‌دهد دارا باشند و هر صورت مشتق‌شدهٔ میانی می‌تواند پسوندهای تصریفی خودش را بگیرد و هر یک از آنها در نقش‌های نحوی کلمه شرکت می‌کنند. بنابراین در زبان آذری میزان قابل توجهی از برهم‌کنش بین نحو و ترتیب تکواژها وجود دارد. مثلاً حالت مطابقه، منسوب‌کردن اسم‌ها و زمان، شرط، وجه، مجهول‌سازی، نفی، سببی و انعکاس فعل‌ها توسط پسوندها مشخص می‌شود.

Sağlam +laş$_1$ +tir$_2$ +max$_3$ (sağlamlaştirmax :نیرومندکردن)

Sağlam + Noun +A3sg + Pnon +Nom ^DB +Verb +Become$_1$
^DB +Verb + Caus$_2$ + Pos^DB +Noun +Infl$_3$ + A3sg+Pnon+Nom

در این مثال رابطهٔ بین یک تکواژ و مشخصه‌اش از طریق نشان‌دار کردن هر دوی آنها با زیرنگاشت عددی یکسان نشان داده شده است و صورت‌های اشتقاقی طبقهٔ میانی یک واژه از طریق اتصالگر "^DB" به‌هم مرتبط شده‌اند. فهرست کامل مشخصه‌های تکواژشناختی زبان ترکی آذری در پیوست همین مقاله قابل مشاهده است.

٤. روش پژوهش

٤-١. روش انتخاب جملات و جامعهٔ آماری

در ابتدا با مراجعه به سایت ویکی‌پدیا لیست روزنامه‌های کشور آذربایجان که به زبان آذری چاپ می‌شوند بررسی گردید. سپس با مراجعه به وب‌سایت این روزنامه‌ها و با توجه به اهداف این تحقیق، چندین روزنامه انتخاب شد[1] و به‌صورت تصادفی از بین این روزنامه‌ها جملات متنوع و در حیطهٔ امور داخلی و مالی و ورزشی کشور آذربایجان و همچنین مسائل مربوط به کشورهای خارجی که در این روزنامه‌ها مطرح گردیده بود

---

۱. اسامی روزنامه‌های انتخاب‌شده برای تحقیق عبارت‌اند از:

Azerbaycan (www.azerbaijan-news.az), Respublika (www.respublica-news.az), Baki Xeber (bakixeber.com), Zaman (www.zaman.az), Iki Sahil (www.ikisahil.com), Kaspi (www.kaspi.az)

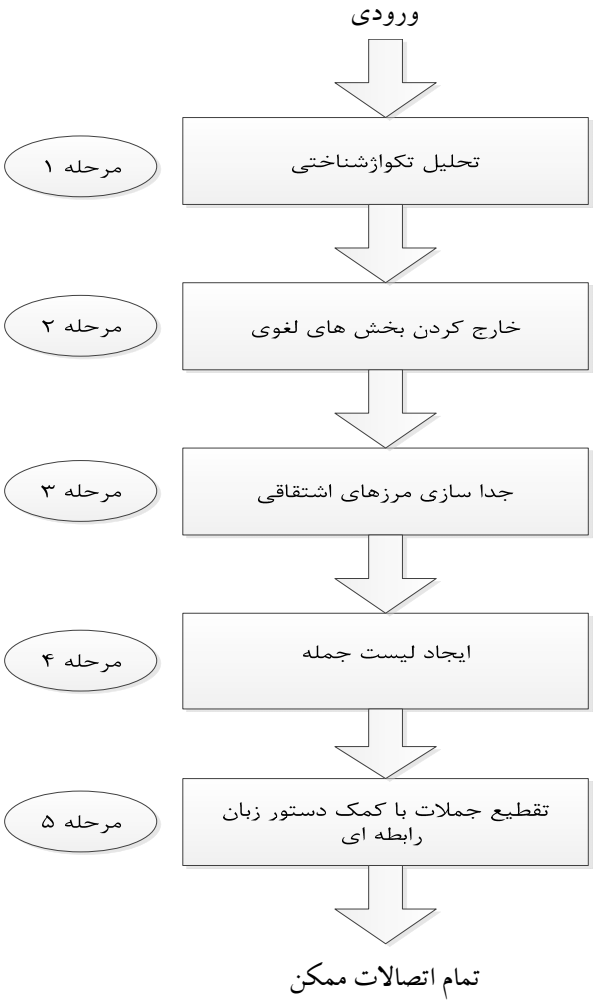
انتخاب شد. این جملات به‌صورت دستی انتخاب، و توسط نرم‌افزار تقطیع و تحلیل شده‌اند.

٤-٢. معماری سیستم

هدف این کار، طراحی نرم‌افزار دستور زبان نحوی زبان آذری در فرم دستور زبان رابطه‌ای می‌باشد. خروجی غیرواژگانی یک تجزیه‌گر تکواژشناختی، بعد از برخی پیش‌پردازش‌ها، همچون یک ورودی برای دستور زبان به‌کار برده می‌شود. اگر تجزیه‌گر تکواژشناختی نتواند یک کلمه را تقطیع کند، آن کلمه یا یک کلمهٔ نامعتبر در زبان آذری است، یا یک کلمهٔ ناشناخته می‌باشد. تجزیه‌گر دستور زبان رابطه‌ای، برخی عاملیت‌ها را فراهم می‌کند تا کلمات ناشناخته را به‌کار بگیرد. در نتیجه، این کلمات همچون ورودی برای تجزیه‌گر استفاده می‌شوند و قوانین ضروری برای این کلمات ناشناخته به دستور زبان اضافه می‌شوند. از این‌رو سیستم کنونی ما، کلمات ناشناخته را به‌کار می‌گیرد. در حال حاضر دستور زبان نمی‌تواند نشانه‌های نقطه‌گذاری را به‌کار گیرد، اما آنها به آسانی می‌توانند مجتمع شوند. تجزیه‌کننده تنها از اطلاعات نحوی و تکواژشناختی استفاده می‌کند و در نتیجه از اطلاعات معنایی استفاده نمی‌کند.

در معماری سیستم، تقطیع‌کننده از تجزیه‌گر ساخت‌واژه‌ای زبان ترکی آذری و دستور زبان رابطه‌ای ایستا بهره می‌گیرد. در شکل ۲، جملهٔ داده‌شده به فرم‌های طبقهٔ میانی هر مرحله منتقل می‌شود و در نهایت تمام اتصالات ممکن در جمله توسط تقطیع‌کننده تولید می‌شوند. در ما بقی این بخش هر مرحله به‌طور جداگانه توضیح داده شده است.

مرحلهٔ ۱. تحلیل ساختواژه‌ای: در تحلیل تکواژشناختی کلمات درون جمله در مرحلهٔ اول، سیستم تجزیه‌گر بعد از گرفتن جملهٔ ورودی، تکواژشناختی خارجی را برای هر کلمه از جمله فرا می‌خواند تا تحلیل تکواژشناختی آنها را به‌دست آورد. اگر تجزیه‌گر تکواژشناختی نتواند یک کلمه را تجزیه و تحلیل کند، از خود کلمه در سیستم استفاده می‌شود.



شکل ۲. معماری سیستم تقطیع‌کنندهٔ زبان ترکی (آذری)

مثلاً اگر جملهٔ ورودی به مرحلهٔ ۱ وارد شود، خروجی آن به‌صورت زیر خواهد بود:

ورودی مرحلهٔ اول:

San kitabi oxudun (تو کتاب را خواندی)

خروجی مرحلهٔ اول:

- تو San

I . san +Pron + A2sg + Pnon + Nom

- کتاب Kitab

I . kitab + Noun + A3sg + Pnon + Acc
II . kitab + Noun + A3sg + P3sg + Nom

- خواندن Oxu

I . oxu + Verb + Pos + Past + A2sg

مرحلهٔ ۲. خارج‌کردن بخش‌های لغوی: در مرحلهٔ دوم، خروجی مرحلهٔ اول برای بخش تقطیع، پیش‌پردازش می‌شود. در این مرحله به غیر از حروف ربط، جنبه‌های لغوی کلمات برداشته می‌شوند. در حقیقت دستور زبان رابطه‌ای ما برای زبان آذری به عوض خود کلمات، مبتنی بر طبقات کلمات و مشخصهٔ ساختارشان (یعنی POS) می‌باشد.

خروجی طبقهٔ میانی مرحلهٔ دوم، فهرستی از مشخصهٔ ساختارهای تکواژشناختی غیرلغوی کلمات می‌باشد. اگر کلمه‌ای با مساعدت حداقل یک پسوند اشتقاقی از کلمه‌ای دیگر اشتقاق یابد، گفته می‌شود که مشخصهٔ ساختارش شامل مرز اشتقاقی است.

خروجی مرحلهٔ دوم:

- San

I. Pron + A2sg + Pnon + Nom

- Kitab

I. Noun + A3sg + Pnon + Acc
II. Noun + A3sg + P3sg + Nom

- Oxu

II. Verb + Pos + Past + A2sg

مرحلهٔ ۳. جداسازی مرزهای اشتقاقی: اگر کلمه‌ای با کمک حداقل یک پسوند اشتقاقی از کلمهٔ دیگر جدا شود، پس مشخصهٔ ساختارش حداقل باید شامل یک مرز اشتقاقی باشد. مشخصهٔ ساختار کلمات دارای مرزهای اشتقاقی با یک روش خاص در سیستم ما به‌کار برده شده است. در مرحلهٔ سه، کلمات در مرزهای اشتقاقی جدا می‌شوند و نشانه‌های ادات سخن هر صورت اشتقاق‌یافته، ارزش‌گذاری می‌گردد تا جایگاهش در آن کلمه مشخص شود.

ورودی:

Noun + A3sg + Pnon + Acc

خروجی:

Noun + A3sg + Pnon + Acc

ورودی:

Noun + A3sg + P1p1 + Loc ^DB + Adj + Re1^DB + Noun + Zero + A3sg + Pnon + Gen

خروجی:

NounRoot + A3sg + P1p1 + Loc
Adj DB
NounDBEnd + A3sg + Pnon + Gen

مرحلهٔ ٤. ایجاد فهرست جمله: هنگامی که نشانگر ادات سخن در سیستم به‌کار گرفته نشود، شمار مشخصات ساختارهای یافت‌شده برای کلمات بسیار زیاد می‌شود. به این دلیل بعد از مرحلهٔ ۴، جمله‌ای جداگانه برای هر یک از تقطیع‌های ساختواژه‌ای کلمه اضافه می‌شود.

ورودی مرحلهٔ ۴:

I . Pron + A2sg + Pnon + Nom
I . Noun + A3sg + Pnon + Acc
II . Noun + A3sg + P3sg + Nom
I . Verb + Pos + Past + A2sg

خروجی مرحلهٔ ۴:

I . Pron + A2sg + Pnon + Nom
Noun + A3sg + Pnon + Acc
Verb + Pos + Past + A2sg
II . Pron + A2sg + Pnon + Nom
Noun + A3sg + P3sg + Nom
Verb + Pos + Past + A2sg

مرحلهٔ ۵. تقطیع جملات: در انتها برای هر یک از این جملات، دستور زبان رابطه‌ای فراخوانده می‌شود و در مرحلهٔ پنجم هر یک از جملات با توجه به دستور زبان رابطه‌ای زبان ترکی آذری تقطیع می‌شود.

ورودی مرحلهٔ ۵:

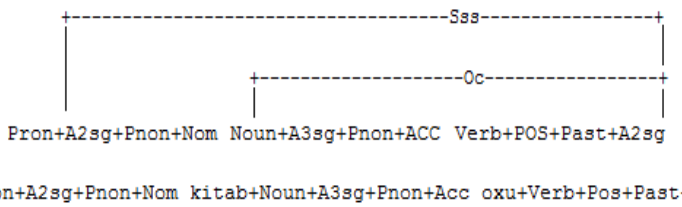
I . Pron+A2sg+Pnon+Nom Noun+A3sg+Pnon+Acc Verb+Pos+Past+A2sg

II . Pron+A2sg+Pnon+Nom Noun+A3sg+P3sg+Nom Verb+Pos+Past+A2sg

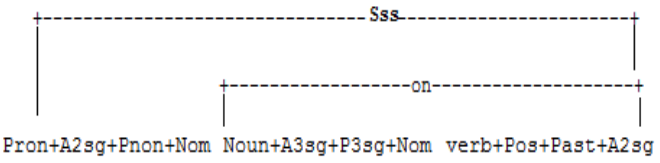
خروجی مرحلهٔ ۵:

San kitabi oxudun

۱-۱



۱-۲



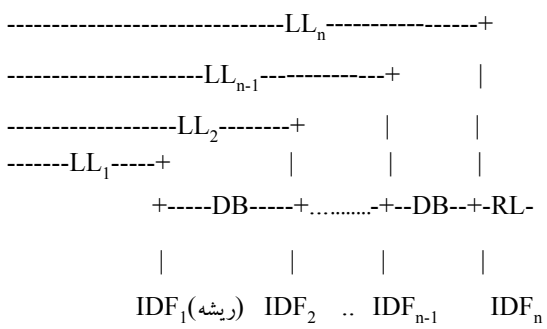
٤-۳. ملزومات ارتباطی مربوط به زبان‌های پیوندی

به منظور حفظ نقش‌های نحوی که صورت‌های اشتقاقی طبقهٔ میانی یک کلمه ایفا می‌کنند، ما با آنها همچون کلمات مجزا در دستور زبان رفتار می‌کنیم. از طرف دیگر، به منظور نشان‌دادن این که آنها اشتقاق‌های طبقهٔ میانی مربوط به یک کلمه هستند، همهٔ آنها به اتصال‌گر "DB"[1] متصل می‌شوند.

در شکل ۳ ملزومات ارتباطی یک کلمه، با n صورت اشتقاق طبقهٔ میانی (IDF_1.....IDF_n) توضیح داده شده است. در شکل ۳، "LL" بیانگر رابطه‌هایی است که به کلمات سمت چپ و "RL" بیانگر رابطه‌هایی است که به کلمات سمت راست اشاره دارند. "IDF"های کلمه توسط رابطه‌های "DB" وصل شده‌اند. همچنان که دیده می‌شود، همهٔ "n IDF"ها می‌توانند به کلمات سمت چپ یعنی "LL" متصل شوند، اما فقط آخرین "IDF" یعنی "IDF_n" می‌تواند به کلمات سمت راست یعنی "RL" متصل شود. علاوه بر آن، "IDF_1" که ستاک ریشه است فقط باید به سمت راستش با اتصال "DB" متصل شود؛ در

1. Derivational Boundary

حالی که آخرین "IDF" یعنی "IDF_n" باید به سمت چپش با همان اتصال‌گر متصل شود. از طرف دیگر، تمام "IDF"های بین این دو، باید هم به سمت چپ و هم به سمت راست خود با رابط‌های "DB" متصل شوند تا مشخص کنند که وابسته بـه کلمـهٔ یکسـانی مـی‌باشـند. از این‌رو همـین کلمـه، در حقیقـت همـین "IDF"، دارای الزامـات ربطـی متفـاوت وابسـته بـه مکانش در کلمه می‌باشد. برای استفاده‌کردن از این موقعیت، آیتم‌هـای متفـاوت در درون دستور زبان قرار می‌گیرند که هر یک از این‌ها ۳ جایگاه همان کلمه را نشان می‌دهند.



شکل ۳. ملزومات ارتباطی فرم طبقهٔ میانی یک کلمه

در ادامه، ملزومات ارتباطی مربوط به حوزهٔ تمامی کلمات ارائه می‌شود.

```
Adverb : (<affix-bound>  & ({EE-} & (EE+ or Ea+))) or <Sffxlss- Adverb-to-Verb-Drv>;
Adverb+AdjMdfy : ({EE-} & (EE+ or Ea+ or EA+)) or <Sffxlss- Adverb-to-Verb-Drv>;
```

شکل ٤. ملزومات ارتباطی قیدها

```
Adj :(<affix-bound> & (({EA-} & A+) or ([<n-noun-right>]))) or <Sffxlss-Adj-to-Verb-Drv>;
Num+Ord :(<affix-bound> & (({NN-} & {EA-} & A+) or ([<n-noun-right>]))) or <Sffxlss-Ord-to-Verb-Drv>;
Adj+Ques :(<affix-bound> & (({EA-} & Aq+) or ([<n-noun-right-q>]))) or <Sffxlss-Adj-to-Verb-Drv>;
```

شکل ۵. ملزومات ارتباطی صفت‌ها

```
<llr_noun>:{{@AN-}& ({{Dn-}&{@A-}} or {{@A-}&{Dn-}})}
```

شکل ٦. ملزومات ارتباطی سمت چپ مشترک برای تمامی اسم‌ها

```
%takes nouns in nominal case :          Postp: (Jn- & (Ap+ or (Ep+ or EEp+ or EAp+)));
%takes nouns in genitive case :         Postp: (Jg- & (Ap+ or (Ep+ or EEp+ or EAp+)));
%takes nouns in dative case :           Postp: (Jd- & (Ap+ or (Ep+ or EEp+ or EAp+)));
%takes nouns in ablative case :         Postp: (Ja- & (Ap+ or (Ep+ or EEp+ or EAp+)));
%takes nouns in instrumental case :     Postp: (Ji- & (Ap+ or (Ep+ or EEp+ or EAp+)));
```

شکل ۷. ملزومات ارتباطی پس‌اضافه‌ها

```
Adj :(<affix-bound> &  (({EA-} & A+) or ([<n-noun-right>])))  or <Sffxlss-Adj-to-Verb-Drv>;
Num+Ord :(<affix-bound> &  (({NN-} & {EA-} & A+) or([<n-noun-right>])))  or <Sffxlss-Ord-to-Verb-Drv>;
Num+Card: (<affix-bound> & ({NN-} & (Dn+ or NN+))) or <Sffxlss- Card-to-Verb-Drv>;
```

شکل ۸. ملزومات ارتباطی اعداد

```
%nominative pronouns can be subject of the verb
Pron+A1sg+Pnon+Nom: (<affix-bound> & (Sfs+)) or <Suffixless-Pron-to-Verb-Drv>;
Pron+A2sg+Pnon+Nom: (<affix-bound> & (Sss+)) or <Suffixless-Pron-to-Verb-Drv>;
Pron+A3sg+Pnon+Nom: (<affix-bound> & (Sts+)) or <Suffixless-Pron-to-Verb-Drv>;
Pron+A1pl+Pnon+Nom: (<affix-bound> & (Sfp+)) or <Suffixless-Pron-to-Verb-Drv>;
Pron+A2pl+Pnon+Nom: (<affix-bound> & (Ssp+)) or <Suffixless-Pron-to-Verb-Drv>;
Pron+A3pl+Pnon+Nom: (<affix-bound> & (St+)) or <Suffixless-Pron-to-Verb-Drv>;
```

شکل ۹. ملزومات ارتباطی ضمایر فاعلی

```
<genitive-noun-right>:(Dg+ or Jg+);
<dative-noun-right>:(IOd+ or Jd+);
<ablative-noun-right>:(IOa+ or Ja+);
<accusative-noun-right>:(Oc+);
<locative-noun-right>:(IOl+);
<instrumental-noun-right>:(Ei+);

<nominative-noun-right-A3sg>:(Sts+ or On+ or Jn+);
<nominative-noun-right-A3pl>:(St+ or On+ or Jn+);

<nominative-noun-right-A3sg-PnonP3sg>: (Sts+ or On+ or Jn+ or [AN+]);
<nominative-noun-right-A3pl-PnonP3sg>: (St+ or On+ or Jn+ or [AN+]);
```

شکل ۱۰. ملزومات ارتباطی سمت راست اسم‌ها

% Linking Requirements of Genitive Pronouns: Rule Part 1

Pron+A1sg+Pnon+Gen: (<affix-bound> & (Dfs+ or Jg+)) or <Suffixless-Pron-to-Verb-Drv>;

Pron+A2sg+Pnon+Gen: (<affix-bound> & (Dss+ or Jg+)) or <Suffixless-Pron-to-Verb-Drv>;

Pron+A3sg+Pnon+Gen: (<affix-bound> & (Dts+ or Jg+)) or <Suffixless-Pron-to-Verb-Drv>;

Pron+A1pl+Pnon+Gen: (<affix-bound> & (Dfp+ or Jg+)) or <Suffixless-Pron-to-Verb-Drv>;

Pron+A2pl+Pnon+Gen: (<affix-bound> & (Dsp+ or Jg+)) or<Suffixless-Pron-to-Verb-Drv>;

Pron+A3pl+Pnon+Gen: (<affix-bound> & (Dtp+ or Jg+)) or<Suffixless-Pron-to-Verb-Drv>;

% Linking Requirements of Accusative Pronouns: Rule Part 2

Pron+A1sg+Pnon+Acc Pron+A2sg+Pnon+Acc Pron+A3sg+Pnon+Acc

Pron+A1pl+Pnon+Acc Pron+A2pl+Pnon+Acc Pron+A3pl+Pnon+Acc: (<affix-

ound> & {Oc+}) or <Suffixless-Pron-to-Verb-Drv>;

شکل ۱۱. ملزومات ارتباطی ضمایر اضافی و مفعولی

```
Pron+A1sg+Pnon+Loc Pron+A2sg+Pnon+Loc
Pron+A3sg+Pnon+Loc Pron+A1pl+Pnon+Loc
Pron+A2pl+Pnon+Loc Pron+A3pl+Pnon+Loc: (<affix- bound> & {IOl+}) or
<Suffixless-Pron-to-Verb-Drv>;

Pron+A1sg+Pnon+Abl Pron+A2sg+Pnon+Abl Pron+A3sg+Pnon+Abl
Pron+A1pl+Pnon+Abl Pron+A2pl+Pnon+Abl Pron+A3pl+Pnon+Abl: (<affix- -
bound>&{IOa+ or Ja+}) or <Suffixless-Pron-to-Verb-Drv>;

Pron+A1sg+Pnon+Dat Pron+A2sg+Pnon+Dat
Pron+A3sg+Pnon+Dat Pron+A1pl+Pnon+Dat Pron+A2pl+Pnon+Dat
Pron+A3pl+Pnon+Dat: (<affix-bound>&{IOd+ or Jd+}) or <Suffixless-Pron-to-
Verb-Drv>;

Pron+A1sg+Pnon+Ins Pron+A2sg+Pnon+Ins Pron+A3sg+Pnon+Ins
Pron+A1pl+Pnon+Ins Pron+A2pl+Pnon+Ins Pron+A3pl+Pnon+Ins: (<affix-bound>
& {Ep+}) or <Suffixless-Pron-to-Verb-Drv>;
```

شکل ۱۲. ملزومات ارتباطی ضمایر مکانی، مفعول به‌ای، مفعول باواسطه، وسیله‌ای

۵.تحلیل یافته‌ها

کارآیی سیستم تحت پوشش با فایلی که شامل جملاتی دربارهٔ خبرهای ورزشی، مالی، خارجی، داخلی (که به‌صورت تصادفی از روزنامه‌های کشور آذربایجان انتخاب شده بود) مورد سنجش قرارگرفت. این جملات به‌صورت دستی انتخاب، و توسط نرم‌افزار، تقطیع و تحلیل شده‌اند. قبل از شروع فرایند سنجش، نشانه‌های نقطه‌گذاری از درون جملات حذف شده‌اند. همچنین تحلیل‌های ساختواژه‌ای نادرست از نتایج حذف گشته‌اند.

در این پژوهش، به‌علت حجم بسیار بالای برنامه‌نویسی و با توجه به این که با افزایش تعداد کاراکترهای کلمه ممکن است حالت‌های بسیاری پدید آیند و درصد خطاها افزایش یابند، برنامه‌نویسی برای کلماتی که بیش از هفت حرف دارند صورت نگرفته است. بنابراین اگر کلمه‌ای بیش از هفت حرف داشته باشد توسط نرم‌افزار بر مبنای اطلاعاتی که برای کلمات هفت حرفی و کمتر نوشته شده است تجزیه و تحلیل می‌شود.

در نتیجه، بسط این پژوهش در حیطهٔ عمل یک تیم برنامه‌نویسی می‌باشد؛ زیرا از آنجا که زبان ترکی آذری، یک زبان پیوندی است، هر کلمه می‌تواند پذیرای تعداد زیادی پسوند باشد که این بند در مورد دیگر زبان‌ها از جمله زبان فارسی صدق نمی‌کند و این زبان‌ها چون پیوندی نمی‌باشند به‌راحتی قابل تقطیع می‌باشند- به‌گونه‌ای که حجم برنامه‌نویسی برای این پژوهش در حدود صد صفحه می‌باشد که با به‌کارگیری زبان برنامه‌نویسی دلفی و با استفاده از قواعد کلی دستور زبان ترکی آذری و بهره‌گیری از قواعد دستور زبان رابطه‌ای و تجزیه‌کنندهٔ ساختواژه‌ای برای زبان آذری، دستور رابطه‌ای ارائه شده است. جدول ۱ نتایج این آزمون را نشان می‌دهد.

جدول ۱. نتایج آماری فرایند سنجش و آزمون

شمار جملات	۲۵۰
شمار متوسط کلمات در هر جمله	۵/۱۹
درصد جملاتی که برایند تقطیع‌کننده‌ها، شامل تقطیع‌کنندهٔ صحیح می‌باشند	۸۴/۳۱
شمار متوسط تقطیع‌ها	۷/۴۹
رده‌بندی متوسط تقطیع‌کننده‌های صحیح	۱/۷۸

در این آزمایش ۲۵۰ جمله مورد بررسی قرار گرفته‌اند و همان‌گونه که پیشتر بیان شد، این جملات به‌صورت تصادفی و دستی از روزنامه‌های کشور آذربایجان انتخاب، و به نرم‌افزار وارد گردیدند. شمار متوسط تقطیع‌کننده‌ها برای جملات ۷/۴۹ می‌باشد- هر چند برای دو تا از جملات، شمار تقطیع‌کننده‌ها بسیار بالا یعنی ۲۲ و ۵۰ می‌باشد. هر دوی این جملات شامل اسم‌های متوالی می‌باشند. چون اسم‌ها در زیرشاخهٔ زمان، مکان، و موضوع تقسیم‌بندی نشده‌اند؛ در نتیجه این فرایند منجر به پدید آمدن گروه‌های اسمی صفتی نامحدود نادرست می‌شود و این، مشکل این دو جمله می‌باشد. علاوه بر آن، یکی از این جملات متشکل از کلماتی با ترتیب تکواژهای اشتقاقی بسیار پیچیده می‌باشد؛ یعنی بسیاری از صورت‌های اشتقاقی طبقهٔ میانی که باعث ایجاد شماری از یال‌های ممکن بین این صورت‌های اشتقاقی طبقهٔ میانی می‌شوند، افزایش می‌یابند. علاوه بر آن برای ۸۴/۳۱

درصد از جملات، نتیجهٔ تقطیع‌کننده شامل تقطیع‌کننده‌های صحیح می‌باشند. این فرایند بیانگر این است که گرچه برخی از موضوعات خارج از محدودهٔ ما هستند، اما ما با پدیده‌های بسیار مهمی سروکار داشتیم و موضوعات مبهم به‌ندرت در زبان اتفاق می‌افتند.

سرانجام رده‌بندی متوسط تقطیع‌کننده‌های صحیح ۱/۷۸ می‌باشد. اما برای ۶۲/۳۹ درصد از جملات، اولین تقطیع‌کننده، تقطیع‌کنندهٔ صحیح می‌باشد و برای ۸۰/۹۴ درصد از جملات، یکی از سه تقطیع‌کنندهٔ اول صحیح می‌باشد. این امر بدین علت می‌باشد که برخی از پسوندهای یک گروه در زبان آذری (به‌عنوان مثال، برخی از پسوندهای زمانی) با برخی از پسوندهای دیگر گروه‌ها (به‌عنوان مثال، پسوندهای ملکی) کاملاً شبیه به‌هم بودند و این امر باعث تداخل در فرایند صحیح تقطیع کلمات جمله، و در نتیجه بعضاً باعث به‌وجود آمدن تقطیع‌های اشتباه می‌شد.

٦. نتیجه‌گیری و کارهای آینده

در این پژوهش ما نرم‌افزار دستورزبان ترکی (آذری) را بر مبنای صورت‌بندی دستور زبان رابطه‌ای طراحی کردیم. در این نرم‌افزار از یک تجزیه‌گر ساختواژه‌ای کاملاً توصیف‌شده بهره گرفتیم که این امر برای زبان‌های پیوندی نظیر ترکی آذری بسیار مهم است. نرم‌افزار دستور زبان رابطه‌ای ترکی آذری که ما آن را طراحی نمودیم بر مبنای دستور زبان لغوی نیست. ما از مشخصهٔ ساختارهای ساختواژه‌ای برای طبقهٔ کلمه بهره گرفتیم. همچنین نقش‌های نحوی صورت‌های اشتقاقی طبقهٔ میانی کلمات در سیستم، از طریق جداسازی کلمات مشتق‌شده از مرزهای اشتقاقی‌شان و برخورد با هر یک از این صورت‌های طبقهٔ میانی به‌عنوان یک کلمهٔ مجزا، نگهداری شدند.

در سال‌های اخیر رابطهٔ حیاتی بین زبان و تفکر در حیطهٔ تشخیص صوت و ایجاد شبکه‌های اینترنتی پیشرفت نموده و باعث شده که پردازش زبان طبیعی رواج یابد و به حیطهٔ تحقیقاتی مهمی تبدیل شود (عرب سرخی، ذوالقدری، و فیلی ۱۳۸۵). برخی از محیط‌های کاربرد پردازش زبان طبیعی عبارت‌اند از درک و تولید زبان طبیعی، بازیابی اطلاعات، استخراج اطلاعات، و ترجمهٔ ماشینی. تمام این محیط‌های کاربردی نیازمند برخی از صورت‌های تجزیهٔ نحوی به‌عنوان مرحلهٔ اساسی می‌باشند. گرچه شمار بسیاری از

این برنامه‌های کاربردی برای زبان‌هایی نظیر انگلیسی وجود دارند، اما زبان آذری از دیدگاه محاسباتی کمتر مورد مطالعه قرار گرفته است. به این دلیل تصمیم گرفتیم که نحو زبان آذری را در پرتو همزمان نظریه‌های زبان‌شناسی مطالعه کنیم و با یک توصیف نحوی که همچون ابزاری در آینده برای ساختن بسیاری از برنامه‌های کاربردی سطح بالا می‌باشد، به پایان برسانیم و در نهایت، دیگر افرادی که علاقه‌مند به پژوهش در این زمینه می‌باشند، می‌توانند بخش‌هایی که در سیستم ما وجود ندارند را به سیستم اضافه کنند تا باعث کامل‌ترشدن طرح نهایی شوند.

همچنین اگرچه این دستور زبان رابطه‌ای، انطباق‌یافته با زبان ترکی آذری می‌باشد، اما می‌تواند برای پیشرفت دستور زبان‌های رابطه‌ای در زبان‌های دیگری که دارای ساختواژهٔ پیچیده می‌باشند به‌کار گرفته شود.

۷. فهرست منابع

دبیرمقدم، محمد. ۱۳۸۷. زبان‌شناسی نظری: پیدایش و تکوین دستور زایشی (ویراست دوم). تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت)، چاپ سوم.

سجادی، آرمین و احمد عبدالله‌زاده بار فروش. ۱۳۸۷. تحلیل نحوی زبان فارسی به کمک گرامر پیوندی. *مجله پردازش علائم و داده‌ها*، ۱ (۹): ۲۵-۴۰.

سجادی، آرمین و محمدمهدی همایون‌پور. ۱۳۸۷. مدل‌سازی دانش تکواژشناختی زبان فارسی به کمک گرامرهای پیوندی. *مجله پردازش علائم و داده‌ها*، ۱(۹): ۴۱-۵۶.

سجادی، آرمین، و محمدرضا مطش بروجردی. ۱۳۸۵. تحلیل نحوی مبتنی بر همسان‌سازی به کمک گرامر پیوندی. دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران، ۱-۳ اسفند، دانشگاه شهید بهشتی تهران.

عبدالله‌زاده، ا. ۱۳۸۳. یادداشت‌های درس پردازش زبان طبیعی. دانشگاه صنعتی امیرکبیر.

عرب‌سرخی، محسن، منصور ذوالقدری جهرمی و هشام فیلی. ۱۳۸۵. استخراج گرامر زبان فارسی با استفاده از الگوریتم‌های ژنتیک. *دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران*، ۱-۳ اسفند، دانشگاه شهید بهشتی.

فرخ، ماندانا. ۱۳۸۱. بررسی ساختمان افعال ساده و مرکب فارسی و تدوین روش‌های سرواژه‌سازی به کمک رایانه. *پایان‌نامه کارشناسی ارشد*، دانشگاه آزاد اسلامی واحد تهران مرکزی.

فیلی، هشام و غ، قاسم ثانی. ۱۳۸۲. استفاده از گرامر درخت افزایشی برای ترجمه انگلیسی به فارسی. *مجموعه مقالات نهمین کنفرانس سالانه انجمن کامپیوتر ایران*، جلد ۱، ص ۶۳۹-۶۴۷.

کشاورزی، نیما. ۱۳۷۸. تقطیع نحوی جملات ساده فارسی بر اساس دستور گروه ساختی هسته‌بنیاد. پایان‌نامه کارشناسی ارشد، پژوهشگاه علوم انسانی و مطالعات فرهنگی تهران.

Akbik, A. 2009. Creating a Semantic Wiki Using a Link Grammar-Based Algorithm for Relation Extraction. Master Thesis of Computer Engineering, Berlin University.

Antworth, E.L. 1999. PC-KIMMO: A Two–Level Processor for Morphological Analysis. Summer Institue of Linguistics.

Assi, M., and M. H. Abdolhosseini. 2000. *Grammatical Tagging of a Persian Corpus*, International Journal of Corpus Linguistics, Vol. 5, No. 1, pp. 69-81.

Chomsky, N. 1981. Lectures on Government and Binding: The Pisa Lectures. 7th Edition, Berlin and New York: Mouton de Gruyter, USA, 1993.

Cicekli, I., and O. Istek. 2006. *A Link Grammar for an Agglutinative Language*. Bilkent University, Ankara, Turkey.

Dehdari, J., and D. Lonsdale. 2005. *A Link Grammar Parser for Persian*. Oral Presentation at First International Conference on Aspect of Iranian Linguistics, Leipzig, Germany.

Demir, C. 1993. *An ATN Grammar for Turkish*. M.S. Thesis, Bilkent University.

Eryiğit, G., and K. Oflazer. 2006. *Statistical Dependency Parsing of Turkish*. In Proceedings of EACL 2006 11th Conference of the European Chapter of the Association for Computational Linguistics, April, Trento, Italy.

Gazdar, G., E. Klein, G. Pullum, and I. Sag .1985. *Generalized Phrase Structure Grammar*. Cambridge: Harvard University Press.

Güngördü, Z. 1993. *A Lexical Functional Grammar for Turkish*. M.S. Thesis, Bilkent University.

Hoffman, B. 1995. The Computational Analysis of the Syntax and Interpretation of 'Free' Word Order in Turkish. PhD Thesis, University of Pennsylvania.

Jurafsky, D., and J. H. Martin. 2000. *Speech and Languag Processing*. Prentice Hall, New Jersey, USA.

Kübler, S. 1998. *Learning a Lexicalized Grammar for German*. Master Thesis of Computational Linguistics, Duisburg, Germany.

Lewis, G. L. 1988. *Turkish Grammar*, Oxford University Press.

Magerman, D. 1993. *Parsing as Statistical Pattern Recongnition*. PhD Thesis, Stanford University.

Melçuk, I. A. 1998. *Dependency Syntax: Theory and Practice*. State University of NewYork Press.

Mollá, D., G. Schneider, R. Schwitter, and M. Hess. 2000. *Answer Extraction Using a Dependency Grammar in Extrans*. T.A.L., Vol. 41, No. 1, pp. 127-156.

Oflazer, K. 1999. *Dependency Parsing with an Extended Finite State Approach*. In Proceedings of 37th Annual Meeting of the Association for Computational Linguistics, June, Maryland, USA.

Pollard, C., and I. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Pub.

Rezaei, S. 1999. Linguistic and Computational Analysis of Word Order and Scram Bling in Persian. Ph.D. Dissertation, Edinburgh, University of Edinburgh.

Sag, I., and T. Wasow. 1999. *Syntactic Theory: A Formal Introduction*. Stanford, California, p.48.

Schneider, G. 1998. *A Linguistec Comparison of Constituency, Dependency and Link Grammar*. Master Thesis, Institut fur Informatic, University of Zurich, Switzerland.

Sleator, D., and D. Temperley. 1993. *Parsing English with a Link Grammar*. Third International Workshop on Parsing Technologies.

Sleator, D., and D .Temperley. 1991. *An Introduction to the Link Grammar Parser*. http://www.link.cs.cmu.edu/link/dict/introduction.html.(accessed 30 July 2014)

Şehitoğlu, O. T. 1996. *A Sign-Based Phrase Structure Grammar for Turkish*. M.S. Thesis, Middle East Technical University.

پیوست

مشخصه‌های تکواژشناختی زبان ترکی (آذری)

^DB	Derivation boundary	Nom	Nominative case for nominal
A1sg	First person singular agreement	Noun	Noun
A2sg	Second person singular agreement	Num	Number
A3sg	Third person singular agreement	Ord	Ordinal numbers
A1pl	First person plural agreement	P1sg	First person singular possessive agreement
A2pl	Second person plural agreement	P2sg	Second person singular possessive agreement
A3pl	Third person plural agreement	P3sg	Third person singular possessive agreement
Abl	Ablative case for nominal	P1pl	First person plural possessive agreement
Acc	Accusative case for nominal	P3pl	Third person plural possessive agreement
Adj	Adjective	Past	Past tense for verbs
Card	Cardinal numbers	PCNom	Postpositions that take nominative nominal
Cond	Conditional for verbs	PCAbl	Postpositions that take ablative nominal
Conj	Conjunctive	PCDat	Postpositions that take dative nominal
Dat	Dative case for nominal	PCIns	Postpositions that take instrumental nominal

Fut	Future tense for verbs	PCGen	Postpositions that take genitive nominal
Gen	Genitive case for nominal	Pnon	No possessive agreement
Imp	Imperative for verbs	Pos	Positive Polarity
Ins	Instrumental case for nominal	Postp	Postposition
Loc	Locative case for nominal	Pres	Present tense for verbs
Narr	Narrative tense for verbs	Prog1	Progressive time for verbs
Neg	Negative Polarity	Prop	Proper Name
P2pl	Second person plural possessive agreement	Pron	Pronoun

Analyzing Turkish Language Based on Parsing Using Link Grammar

Maryam Arabzadeh[1]
MA. in General Linguistics
Islamic Azad University of Ahar; Iran

Seyed Mehdi Araghi[2]
Professor of Language Teaching
Assistant Professor of General Linguistics
Islamic Azad University of Ahar; Iran

Iranian Journal of Information Processing & Management

Iranian Research Institute for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in SCOPUS, ISC & LISA
Vol.29 | No.3 | pp: 845-870
Spring 2014

Abstract: There are different classes of theories for the natural lanuguage syntactic parsing problem and for creating the related grammars. This paper presents a syntactic grammar developed in the link grammar formalism for Turkish which is an agglutinative language. In the link grammar formalism, the words of a sentence are linked with each other depending on their syntactic roles. Turkish has complex derivational and inflectional morphology, and derivational and inflection morphemes play important syntactic roles in the sentences. In order to develop a link grammar software for Turkish, the lexical parts in the morphological representations of Turkish words are removed, and the links are created depending on the part of speech tags and inflectional morphemes in words. Furthermore, derived words are separated at the derivational boundaries. The adapted unique link grammar formalism for Turkish provides flexibility for the linkage construction, and similar methods can be used for other languages with complex morphology. Finally, using the Delphi programming language, the link grammar related to the Azeri language was developed and implemented and then by selecting 250 random sentences, this grammar is evaluated and then tested. For 84.31% of the sentences, the result set of the parser contains the correct parse.

Keywords: Parser; Link Grammar; Natural Language Process; Morphological Analysis; Delphi Programming Language

1. Corresponding Author
ma25286@yahoo.com
2. m_araghi@pnu.ac.ir