پردازش دستوری زبان فارسی با رایانه

مصطفی عاصی (پژوهشگاه علوم انسانی)

موضوع اصلی در زبان‌شناسی رایانه‌ای پردازش زبان طبیعی است و هنگامی که کارشناسان رایانه از پردازش زبان طبیعی سخن می‌گویند اغلب مسائلی را عنوان می‌کنند که نشان می‌دهد تمایز روشنی میان خط و زبان قائل نیستند. به سخن دیگر، در بسیاری موارد، این دو حوزه را در هم می‌آمیزند. بد نیست در آغاز به برخی از کاربردهای رایانه در حوزهٔ خط اشاره‌ای نماییم و سپس به نقش آن در حوزهٔ زبان به‌ویژه ساخت واژه بپردازیم.

۱ حوزهٔ خط

۱-۱ غلط‌یابی املائی

یکی از ابزارهای سودمندی که واژه‌پردازها به‌تدریج از آن سود جستند، خطایاب یا غلط‌یاب املائی[1] است. در نسخه‌های جدید نرم‌افزار Word، برای متنهای انگلیسی، می‌توان از کاراییهای بالای غلط‌یاب آن بهره گرفت. از چند سال گذشته، برخی از شرکتهای سازندهٔ نرم‌افزار واژه‌پرداز فارسی کوشیدند نمونه‌هایی از چنین غلط‌یابهایی را برای فارسی به کار گیرند، از جمله **پیشکار**، **زرنگار**، **گستره‌نگار** و **نقش**. در دو نرم‌افزار نخست به نظر می‌رسد بیشترین اتکای برنامه بر جستجوی قاعده‌مند و الگوریتمی

1) spelling checker

ساختهای واژه قرار داشته باشد، در صورتی که در غلط‌یابِ گستره‌نگار که نسخه‌ای از آن در واژه‌پردازِ نقش نیز به کار رفت، تکیه بر جستجوی واژه در یک فهرست دویست هزار واژه‌ای است. روش دوم، که امکان استفاده از آن در نرم‌افزارهای دیگر نیز هست، با سرعت و دقت بیشتری کار می‌کند؛ اما هر دو روش دارای مشکلاتی هستند.

مهم‌ترین اشکال در اغلب غلط‌یابها عدم توجه آنها به بافت (واژه‌های همسایه) است، که باعث می‌گردد واژه‌ای با ظاهر درست در جمله‌ای نابه‌جا به کار رود و غلط شناخته نشود. مثلاً واژهٔ اسب در جملهٔ زیر درست به شمار می‌آید: امروز هوا گرم اسب.

یکی از امکانات غلط‌یابها افزودن واژه‌های جدید به فهرست است، که این ویژگی در جستجوهای فهرستی کارآئیِ بیشتری دارد. از سوی دیگر، امکان پیشنهاد واژهٔ درست در برابر واژهٔ غلط، شمشیری دولبه است. گر چه پیشنهادهای به‌جا می‌تواند در سرعت غلط‌گیری بسیار مؤثر باشد، اما پیشنهادهای نامربوط ــ که تعداد آنها در واژه‌پردازهای فارسی بسیار است ــ بیشتر باعث کندی کار می‌گردد. نکتهٔ دیگر وابستگی این‌گونه غلط‌یابها به یک دستور خط خاص است و تا هنگامی که دستور خطی استاندارد (فراتر از دستور خط فرهنگستان) و بدون موارد استثنائی تدوین نگردد، آشفتگی و سردرگمی این نرم‌افزارها نیز پایان نخواهد گرفت.

۱-۲ بازشناسی خودکار متن (OCR)

درون‌داد متن کاری وقت‌گیر و پر خطاست و حجم متنهایی که پیش‌تر چاپ شده‌اند بسیار زیاد است. از اینجاست که اندیشهٔ درون‌دادِ خودکار متنها شکل می‌گیرد. درون‌داد نگاره‌ایِ یک متن (با روش عکس‌برداری یا پویش) تصویری غیر قابل استفاده برای پردازش فراهم می‌نماید. روشها و سیستمهای متعددی برای بازشناسی خودکار متنهای زبانهای اروپایی (با حروف لاتین) به وجود آمده و بسیاری از آنها نیز با درجهٔ دقت بالایی کار می‌کنند. شرکت صَخر وابسته به مایکروسافت، که بیشتر برای خط و زبان عربی فعالیت می‌کند، نخستین بار برنامه‌ای برای بازشناسی متنهای عربی تهیه کرد. نسخهٔ ابتدائی حتی برای متنهای عربی مشکلاتی داشت، ولی می‌توانست در مورد خط فارسی پایه‌ای برای آغاز به شمار آید. یکی از شرکتهای ایرانی، بر همین اساس، نرم‌افزاری به نام شناسا تولید کرد که تا مدتی تنها برنامهٔ OCR فارسی به شمار می‌آمد. متأسفانه، با

وجود نارساییهای متعدد و درجهٔ دقت پایین، تلاشی از سوی تهیه‌کنندگان آن برای بهبود و افزایش دقت نرم‌افزار صورت نگرفت. به‌تازگی شرکت صخر نسخهٔ ۶ برنامهٔ متن‌خوان خودکار خود را عرضه نموده که برای زبان و خط فارسی نیز امکاناتی را ارائه می‌دهد.

۲ حوزهٔ زبان

پردازش زبان فارسی در سطوح چهارگانهٔ آوایی، ساخت‌واژی، نحو، و معنایی و در حوزه‌های کاربردی و میان‌رشته‌ای به صورت پراکنده و در نهادهای دانشگاهی و پژوهشی انجام پذیرفته و متأسفانه ارتباط منظمی میان آنها وجود نداشته است. از این رو، فعالیتهای مشابه و موازی بسیار مشاهده می‌شود. شاید بتوان امیدوار بود، با ایجاد مراکز پژوهشی مشخص و انجام پژوهشهای هدف‌دار و برنامه‌ریزی شده، تا اندازه‌ای از پراکنده‌کاری و دوباره‌کاری جلوگیری شود. به دلیل یادشده، تنها به برخی از پژوهشهای نمونه در هر زمینه اشاره می‌گردد.

۲-۱ آواشناسی

سیر منطقی بررسی این حوزه باید به شناسائیِ واجهای زبان فارسی و مشخصه‌های آنها با روشهای آزمایشگاهی و روشن کردن بسیاری از موارد ابهام یا مورد اختلاف دربارهٔ آنها بپردازد؛ از جمله تعیین دقیق واکه‌ها[۲] و همخوانها[۳]ی اصلی فارسی معیار و گونه‌های آنها، وجود واکه‌های مرکب[۴] و تعداد و کیفیت آنها، ماهیت همزه به عنوان یک واج در جایگاههای مختلف واژه و گونه‌های آن، و بسیاری نکات دیگر در این زمینه.

مرحلهٔ دیگر بررسی واحدهای زبرزنجیری[۵] یا نواهای گفتار است؛ عواملی مانندِ زیر و بمی[۶]، تکیه[۷]، نواخت[۸]، آهنگ[۹]، و درنگ[۱۰] که در گفتارِ پیوسته بر روی زنجیرهٔ آواها و در سطوح آوا، هجا، واژه و جمله تأثیر می‌گذارند. این واحدها نه‌تنها از نظر تعیین، تغییر و تمایزِ معنی بلکه از لحاظ ایجاد لحن طبیعی گفتار و بیان حالات گوناگون عاطفی دارای اهمیت‌اند.

---

2) vowels 3) consonants 4) diphthongs 5) suprasegmentals
6) pitch 7) stress 8) tone 9) intonation
10) juncture

پـردازش گـفتار فـارسی بـدون تـوجه دقیق بـه فـرایـندهای آوایـی، کـه مـجموعهٔ دگرگونیهایی است که در اثر هم‌نشینی آواها و تأثیر ویژگیهای آنها در هـنگام گـفتار بـر یکدیگر به وجود می‌آید، چه در مـرحـلهٔ درک و بـازشناسی و چـه در مـرحـلهٔ تـولید، غیرواقعی و حتی غیرعملی خواهد بود. آنچه گفتار بازسازی‌شده را از حالت ماشینی، خشک و مقطّع به گفتاری طبیعی و انسانی تبدیل می‌کند دخالت دادن عـواملی مـانند واحدهای زبرزنجیری و نیز فرایندهای آوایی در تولید گفتارِ پیوسته است.

۲-۱-۱ برخی بررسیهای انجام‌شده در حوزهٔ آواشناسی

اسلامی (۱۳۷۹) ویژگیهای آهنگی زبان فارسی را بررسی کرده و نشان داده که زیر و بمی دارای یک نظام واج‌شناختی است. به نظر وی، عناصر آهنگی یعنی تکیهٔ زیـر و بـمی، نواختِ گروه و نواختِ مرزنما هویت مستقل از یکـدیگر دارنـد و بـه‌طور مستقل نـیز می‌توانند تغییر کنند و معنای آهنگیِ متفاوتی به پاره‌گفتار ببخشند. این عناصرِ آهنگی در ترکیب با هم الگوهای آهنگی می‌سازند که هـر کـدام نـمایندهٔ بـافت خـاصی هسـتند. اسلامی نشان می‌دهد که از ترکیب منطقی تکیه‌ها و نواختهای کناری در زبـان فـارسی شانزده الگوی آهنگ به دست می‌آید. ایشان، با طرح مباحثی مربوط به مرز گروههای آهنگی، نشان می‌دهد که گفتار پیوسته بـه‌صورت قـاعده‌مند بـه واحـدهای کـوچک‌تر تقسیم می‌شود. از اطلاعات نحوی می‌توان در شناسائیِ مرز گروههای آهنگی استفاده کرد. آنگاه از بحث مرز گروههای آهنگی در بازسازی گفتار استفاده مـی‌شود و بـدین وسیله گفتارِ بازسازی‌شده به واقعیت نزدیک‌تر می‌شود.

وی نشان داده که، در تأکید کلی، الگوی برجستگی واحدهای نحوی بر اساس اصل هسته‌گریزی است و، در آن، تکیهٔ زیر و بمی روی دورترین وابستهٔ هسته قرار می‌گیرد. در جـملات پـیچیده، هـر کـدام از گـروههای نـحوی مؤکد، بـر اسـاس هـمان اصـل هسـته‌گـریزی، تکـیه مـی‌گیرند. در آخـر، ایشـان، بـر اسـاس اطـلاعات واج‌شـناختی، واحدهای واژگانی را شناسایی می‌کند. از آن‌جـایی کـه هـجای تکـیه بـر ایـن واحـدها مشخص شده است، در بازسازی گفتار می‌توان پیش‌بینی کرد که کدام هجای یک واحد واژگانی می‌تواند بالقوه جایگاه تکیهٔ زیـر و بـمی بـاشد. سـپس چگـونگی اسـتفاده از اطلاعات واژگانی در بازشناسیِ مرز واژه در گفتار پیوسته مورد بررسی قرار مـی‌گیرد.

وی متذکر می‌شود که، به‌جز در موارد معدودی، واحدهای واژگانی تکیه‌پایانی‌اند و، اگر در پاره‌گفتار برجسته شوند، هجای تکیه‌برِ واژگانیِ آنها محل تکیهٔ زیر و بمی خواهد بود. بنابراین، تکیهٔ زیر و بمی در هر جایی از گفتار که ظاهر شود نشانهٔ مرز واژه است.

نتیجهٔ کار غلام‌پور (۱۳۷۹) تشخیص رشتهٔ آوائیِ ورودی است. سپس، با رجوع به واژگان، صورتهای متفاوتی از مجموعه کلماتی که می‌توانند با آن رشتهٔ آوایی متناظر باشند شناسایی می‌شود. در نهایت، از بین رشته کلمات موجود، با کمک تقطیع‌گرِ پایین به بالا جملهٔ صحیح تشخیص داده می‌شود. تقطیع‌گرِ وی از حدود هشتصد قاعده استفاده می‌کند که حدود صد ساخت فارسی مشمول آن است.

قاسمی (۱۳۷۷) مبنای آواشناختی برای انتخاب و استخراج واحدهای آوایی به‌منظور سنتزگفتار فارسی معرفی کرده است. وی یازده فرایند آوایی را بررسی کرده که، از آن میان، پنج فرایند را در امر بازسازیِ گفتار مهم و شش مورد از آنها را قابل اغماض می‌شمارد.

در فرایند تولید ناقصِ دو همخوانِ همانند در مرز دو هجا، وی توصیه می‌کند که در مورد واجهای انسدادی-سایشی در کلمهٔ بازسازی‌شده بین دو هجا مکثی به اندازهٔ تولید همخوان اول در نظر گرفته و سپس آن همخوان حذف شود. در مورد واجهای دارای مشخصهٔ پیوسته یا خیشومی، بین دو هجای منظور مکثی وجود ندارد و از واحدهای معمولی می‌توان استفاده کرد. به نظر وی، مکث بین دو هجا در گروه اول در حالت بازسازی در پاره‌گفتار یا جمله ۷۰ میلی ثانیه و در واژه ۱۳۰ میلی ثانیه است. به‌کار بردن [h] واک‌دار به جای [h] سایشی ایجاد اشکال می‌کند، ولی عکس آن بدون اشکال است. از این رو، قاسمی توصیه می‌کند که تمام واحدهای دارای [h] را از محیطی استخراج کنیم که دارای [h] سایشی باشند. وی نشان داده است که در هجای cvc، اگر همخوان آخر [n] و واکه از نوع بلند (â, u, i) باشد، کشش واکه در این محیط از کشش واکه در محیطهای دیگر حدود ۶۰ میلی‌ثانیه کمتر است و هرگاه در واژه‌ای [i] قبل از [y] بیاید کشش آن نسبت به محیطهای دیگر کمتر (تقریباً ۷۰ میلی‌ثانیه) می‌شود. واحدهای دارای همزهٔ بسیار خفیف را نمی‌توان به جای بقیهٔ واحدهای نظیر آنها به کار برد. بنابراین، در استخراج واحدها نباید از همزهٔ بسیار خفیف استفاده کرد. یعنی این‌گونه واحدها را نباید از گروه یا جمله استخراج کرد بلکه باید از واژه استخراج شوند. واج‌گونهٔ لرزشی [r] و

واج‌گونهٔ واکه‌گونهٔ [ɪ] را می‌توان به جای هم و به جای واج‌گونه‌های دیگر /ɪ/ به‌کار برد، ولی واج‌گونهٔ زنشی [ɾ] را نمی‌توان به جای واج‌گونه‌های دیگر به‌کار برد. بنابراین، در استخراج واحدها نباید واج‌گونهٔ [ɾ] زنشی داشته باشیم. به این ترتیب، فرایندهای مذکور از فرایندهای مهم در سنتز طبیعی گفتار به شمار می‌آیند.

۲-۲ ساخت‌واژه و نحو

ساخت‌واژه[11] به بررسی کوچک‌ترین واحد معنی‌دار زبان یعنی تکواژ[12]، انواع آن، آرایش و چگونگی شرکت آن در ساخت واحدهای بزرگ‌تر، یعنی واژه‌ها، می‌پردازد. تکواژ نیز واحدی انتزاعی است که ممکن است صورتهای کاربردی و عینی گوناگونی داشته باشد.

مجموعه واژه‌های هر زبان موجودی یا دارائی آن زبان به شمار می‌رود و هرچه فهرست واژگانِ[13] یک زبان بلندبالاتر باشد آن زبان غنی‌تر به شمار می‌آید. زبان فارسی، با تنوع و انعطاف بسیار، تقریباً از همهٔ الگوها و فرایندهای واژه‌سازی چه آنها که بیشتر در زبانهای تصریفی به کار می‌روند (مانند اشتقاق) و چه آنها که ویژهٔ زبانهای ترکیبی هستند (یعنی ترکیب) و حتی فرایند وندافزایی و پیوند که بیشتر مربوط به زبانهای پیوندی است بهره می‌گیرد و از این رو، برخلاف تصور برخی از افراد، توانائیِ واژه‌سازی بالقوهٔ آن بسیار بالاست.

به گمان گروهی «اولین گام در تحلیل نحوی، شناسائیِ مقولاتی است که واژه‌های یک زبان بدان تعلق دارند» (اگرادی و دیگران، ۱۳۸۰، ص۲۰۸). اما شاید در حوزهٔ ساخت‌واژه نیز نخستین گام فراهم آوردن فهرستی از اقلام واژگانی زبان است که امروزه با ایجاد پایگاههای داده‌های زبانی عملی می‌گردد و در بخش دیگری به آن اشاره خواهد شد. اما تنها فهرست واژگان نیست که مورد نیاز برنامه‌های گوناگون پردازش زبان طبیعی است بلکه فهرستهای ویژهٔ دیگری نیز در فعالیتهای خاص به کار گرفته می‌شوند. پیش از هرگونه تحلیل خودکار یا پردازش رایانه‌ای متنهای زبانی، توصیف دقیق زبان‌شناختی آن ضروری است. خوشبختانه بررسیهای علمی زیادی انجام شده است که می‌تواند زمینهٔ این‌گونه تحلیلها را فراهم سازد.

---

11) morphology 12) morpheme 13) lexicon

۲-۲-۱ برخی بررسیهای انجام‌شده در حوزهٔ ساخت‌واژه و نحو

بقایی (۱۳۸۰) و امامی (۱۳۸۰)، با بهره‌گیری از پایگاه داده‌های زبان فارسی، به جداسازی همهٔ تکواژهای فارسی پرداختند و فهرست کامل آنها را ارائه دادند.

ماه‌جانی (۱۳۷۸) مدلی برای نمایش اطلاعات نحوی و معنائیِ مدخل واژگانی فعل ارائه داده است. مدل پیشنهادی وی از نحو به سوی معنا حرکت می‌کند. وی، در سطح اول که سطح ساختاری است، اطلاعات دستوری اعم از مقولهٔ نحوی نهاد، چارچوب زیرمقوله‌ای (متممهای اجباری فعل) و مقولهٔ نحوی متممهای اختیاری (ادات) را نشان داده است. در سطح بعدی که ساختار موضوعی است، اطلاعات ساختار موضوعی، نقشهای معنایی، محدودیتهای گزینشی، هستهٔ واژگانی و بالاخره پربسامدترین ساختها با هستهٔ فعل نشان داده می‌شود.

سمائی (۱۳۷۷) در پایان‌نامهٔ دکتری، با توجه به داده‌هایش، دوازده حوزهٔ دستوری را باز شناخته است. این حوزه‌ها عبارت اند از صفت، ضمیر، اسم، فعل، قید، حرف اضافه، علائم سجاوندی، جمله‌سازی، گشتار، صرف، املا و واژگان. وی، سپس، ویژگیهای هر حوزه را استخراج و قواعد حاکم بر آن را ارائه کرده است. این کار، به ادعای سمائی، بر اساس فرضیهٔ استقلال نحو چامسکی است.

یکی از چالشهای بزرگ در پردازش خودکار متنهای زبانی شناسائیِ واژه‌ها و نشانه‌گذاری آنهاست. نشانه‌گذاری دستوری را معمولاً برچسب‌دهی می‌نامند و تعیین مجموعهٔ برچسبهای دستوری هر زبان، به‌جز چارچوبها و قواعد عمومی، شرایط ویژهٔ خود را نیز دارد.

فرّخ (۱۳۸۱)، با بررسی مفصّل فعل در زبان فارسی، نوعی دسته‌بندی ارائه داده است که با توجه به آن بتوان برنامه‌ای برای رایانه نوشت تا شناسائیِ افعال در متن به‌طور خودکار انجام و سپس اجزاء و نوع آنها تعیین شود.

دانش‌کار آراسته (۱۳۸۱) برنامه‌ای چهارصد خطی، به زبان Visual Basic، برای تشخیص فعل در زبان فارسی نوشته است. این نرم‌افزار قادر است ویژگیهای زمان، شخص، عدد، معلوم، مجهول، سببی و ریشهٔ فعل را اعلام نماید.

در مرحلهٔ اول، کاربر متنی راکه ممکن است شامل یک واژه، یک عبارت، یک جمله یا چندین جمله باشد وارد می‌کند. واحد متنْ جمله در نظر گرفته شده است؛ بنابراین،

باید پایان متن را با یکی از علائم سجاوندی به برنامه اعلام نمود. مرحلهٔ دوم تشخیص واژه است که مرز آن فاصله است. تکواژهایی که مربوط به فعل‌اند اما جدا نوشته می‌شوند برای برنامه تعریف شده‌اند؛ بنابراین، برنامه به‌طور خودکار فاصلهٔ بین این تکواژها و فعل را حذف می‌کند و این کلمات را به صورت یک واژهٔ یکپارچه به حساب می‌آورد.

مرحلهٔ بعد بررسی فعل‌بودن یا فعل‌نبودنِ واژه است. این قسمت بدنهٔ اصلی برنامه است و بیشترین بخشهای برنامه را در بر می‌گیرد. برای این برنامه یک پایگاه داده شامل ستاکهای گذشته و حال تهیه شده است. ابتدا همهٔ واژه در فهرست جستجو می‌شود. افعالی که هیچ‌گونه پیشوند یا پسوندی ندارند به‌راحتی در فهرست پیدا می‌شوند. سپس، مشخصات فعل مورد نظر، بر اساس اجزای اعلام‌شده در فهرست دیگری، اعلام می‌شود. پایگاه داده‌های این برنامه شامل ۴۵۰ واژه است. این برنامه، با طراحی مرحله‌به‌مرحله، اقدام به شناسائیِ اجزای واژه می‌کند و با جداسازی و تجزیهٔ این افعال نوع فعل را مشخص می‌کند.

عاصی و حاج عبدالحسینی (Assi and H. Abdolhosseini 2000)، برای تعیین مقوله‌های دستوری واژه‌های متنهای پیوستهٔ فارسی، از روشی ریاضی و آماری بهره می‌گیرند. روش مورد استفاده، که برچسب دهی توزیعی[14] نامیده شده، نخستین‌بار به‌وسیلهٔ شوتس (Schuetze 1995) برای زبان انگلیسی به کار گرفته شد. در این روش، فرض بر این است که رفتار نحوی واژه‌ها در الگوهای هم‌وقوعی[15] آنها بازتاب می‌یابد. برنامه، با ایجاد بردارهای آماری از همسایه‌های دو سوی هر واژه و بررسی شباهتهای رفتار نحوی‌شان، احتمالهای ممکن مقولهٔ دستوری آن را محاسبه می‌کند و برچسب مناسب را از میان یک مجموعهٔ ۴۵ تایی برمی‌گزیند. تعیین مجموعه برچسبهای هر زبان و برای هر منظور تابع شرایط و معیارهای متعددی است که این کار را به چالشی بزرگ تبدیل می‌کند. مجموعه برچسب این طرح نیز با نشانه‌های دقیق و با نظم سلسله‌مراتبیِ حساب‌شده‌ای مشخص گردیده و برچسبها هیچ‌گونه تداخل یا هم‌پوشانی با یکدیگر ندارند.

جدول زیر مجموعه برچسبهای مورد بحث را نشان می‌دهد.

14) Distributional Part-of-Speech Tagging 15) co-occurrence

No.	Tag	Complete Tag Name	Description	Example
1	**ADJ**	Adjective	Any word or compound distinctly functioning as an adjective	*bozorg* (big)
2	**ADJC**	Adjective-Comparative	Comparative adjectives bearing the ending *-tar* (-er)	*bozorg-tar* (bigger)
3	**ADJN**	Adjective-Noun	Forms ambiguous between adjectives and nouns	*por* (full) and *par* (feather) have identical spelling
4	**ADJS**	Adjective-Superlative	Superlative adjectives bearing the ending *-tarin* (-est)	*bozorg-tarin* (biggest)
5	**ADVI**	Adverb-Interrogative	Equivalent to wh-words in English questioning adverbs	*chetor* (how)
6	**ADV**	Adverb	Any distinctly recognizable adverb other than those specified in this tagset	*šetäb-än* (hurriedly)
7	**ADV/C**	Adverb-Complement	Prepositional phrases appearing as single forms in orthography	*be-to* (to-you) or *baräy-aš* (for-him)
8	**ADVJ**	Adverb-Adjective	Forms ambiguous between adverbs and adjectives	*xūb* (good/well)
9	**ADVN**	Adverb-Noun	Forms ambiguous between adverbs and nouns	*sar-anjäm* (finally/end)
10	**ADVP**	Adverb-Place	Adverbs of place	*in-jä* (here)
11	**ADVPR**	Adverb-Preposition	Forms ambiguous between adverbs and prepositions	*birūn* (out/out of)
12	**ADVT**	Adverb-Time	Adverbs of time	*hälä* (now)
13	**ATD**	Attribute-Demonstrative	Demonstratives	*in* (this)
14	**ATD/A**	Attribute-Demonstrative-Accusative	Combination of demonstratives with *RA* the so-called direct object marker	*in-rä* *in-rä*
15	**ATD/K**	Attribute-Demonstrative+Subordinator	Combination of demonstratives and the subordinator *ke* appearing in a single form in orthography	*än-ke* (corresponding) to the relative pronoun who)
16	**ATE**	Attribute-Exclamation	Exclamations used in the specifier position of noun phrases	*ajab* in *ajab ketäb-i* (what a book!)
17	**ATI**	Attribute-Interrogative	Question words used in the specifier position of noun phrases	*kodäm* (which)
18	**ATU**	Attribute-Unspecified	Indefinite article	*har* (every)
19	**CONJ**	Conjunction	Any conjunction	*va* (and), *yä* (or)

No.	Tag	Complete Tag Name	Description	Example
20	**N**	Noun	Any distinct noun other than those specified in this tagset	*ketäb* (book)
21	**NPP**	Noun-Pronoun-Personal	Personal pronouns. These pronouns are used in subject and object position alike in addition to being used as possessive adjective and pronouns.	*man* (I)
22	**NPP/A**	Noun-Pronoun-Personal-Accusative	Combination of NPP and *RA* (the direct object marker) appearing as one unit in writing	*ma-ra* (me)
23	**NPREF**	Noun-Pronoun-Reflexive	Reflexive and emphatic pronouns	*xod-am* (myself)
24	**NPEM**	Noun-Pronoun-Emphatic	The emphatic form without the ending specifying the person	*xod* (self)
25	**NPKE**	Noun-Pronoun-KE	The relative pronoun *ke*	*ke* (that, who...)
26	**NPU**	Noun-Pronoun-Unspecified	Indefinite pronouns	*hame* (everyone)
27	**NPREC**	Noun-Pronoun-Reciprocal	Reciprocal pronouns	*hamdigar* (each other)
28	**NUMC**	Number-Cardinal	Cardinal numbers	*yek* (one)
29	**NUMC/**	Number-Cardinal-Unspecific	Unspecific numbers	*dah-hä* (tens)
30	**NUMO**	Number-Ordinal	Ordinal numbers	*avval* (first)
31	**NV/P**	Noun (Pronoun) + Verb	Combination of personal pronouns and verbs appearing as one unit in orthography	*u-st* (he-is)
32	**PART**	Past Participle	Past participle forms of verbs	*raft-e* (gone)
33	**PREP**	Preposition	Unambiguous prepositions	*be* (to)
34	**Prep/**	Preposition-Conjunction	The form *tä*, which is ambiguous between a preposition and a conjunction	*tä* (until, to, so that)
35	**PUNC**	Punctuation	Punctuation marks	. , ; " "
36	**RA**	Accusative Marker *RA*	The only postposition of standard Persian, the so-called direct object market *rä*	
37	**VAUX**	Verb-Auxiliary	Auxiliary verb	*bäyad* (must)
38	**VDEC**	Verb-Declarative	Any declarative verbs other than those specified	*gof-t-am* (I said)

No.	Tag	Complete Tag Name	Description	Example
39	VDECN	Verb-Declarative-Noun	Ambiguous forms between past tense third person singular declarative verbs and truncated infinitives functioning as nouns	*xar-id* (He bought, Shopping)
40	VINF	Verb-Infinitive	Infinitive form of verbs	*xar-id-an* (to buy)
41	VLINK	Verb-Linking	Linking verbs	*ast* (is)
42	VIMP	Verb-Imperative	Imperative forms of verbs	*bo-ro* (Go.)
43	VSUB	Verb-Subjunctive	Subjunctive forms of verbs	*be-rav-ad* ((if) he goes, he (must) go)
44	/LTR	Letter	Letters or mistyped partial words	
45	???	Unknown	Unknown items	

۲-۲-۲ روشها و ابزارهای تحلیل دستوری: زبان‌شناسی پیکره‌ای

به موازات پیشرفت و تحولات نظری زبان‌شناسی جدید و شکل‌گیری مکاتب گوناگون، روشهای تحلیل نیز تحول یافت. روشهای ساخت‌گرایانه که تا دههٔ چهل و پنجاه میلادی به‌اوج رسید، بیشتر به حوزهٔ ساخت‌واژه می‌پرداخت و از روش تجزیه به سازه‌های بیاپی[۱۶] بهره می‌گرفت. دستور زایشی با رویکردی نحوی به تکمیل روش یادشده پرداخت و تحلیل سازه‌ای[۱۷] را به وجود آورد و، با کمک گرفتن از نمودارهای ژرف‌ساختی، روساختی و گشتارها، تحلیل گشتاری[۱۸] را سامان داد. مکتبهای دیگر زبان‌شناسی نیز تحلیلهای متفاوتی ارائه کرده‌اند مانند تحلیل رابطه‌ای[۱۹] و تحلیل نقش‌گرا[۲۰] که در هریک از آنها مجموعه‌ای از قواعد، انگاره‌ها، نمودارها و نشانه‌ها برای توصیف نحوی زبان به کار گرفته می‌شود. با گسترش و اهمیت پیدا کردن رویکرد متن‌گرا و کاربرد عملی آن در حوزهٔ پردازش زبان و نیز بهبود و افزایش امکانات رایانشی برای ذخیره‌سازی، سامان‌دهی، پردازش، جستجو و دستیابی متنهای بزرگ زبانی، شاخهٔ جدیدی در زبان‌شناسی به صورت میان‌رشته‌ای با رایانه به نام **زبان‌شناسی پیکره‌ای** شکل گرفت.

در سال ۱۹۹۲، هلیدی، زبان‌شناس نامی، در همایش ویژه‌ای دربارهٔ زبان‌شناسی

16) immediate constituents analysis
17) phrase structure analysis
18) transformational analysis
19) relational analysis
20) functional analysis

پیکره‌ای گفت:

«از نخستین روزهایی که تصمیم گرفتم دستورنویس شوم، همواره می‌اندیشیدم که دستور موضوعی است با مقدار زیادی نظریه و مقدار ناچیزی داده. از این رو، برای دو نکته اهمیت قائل بوده‌ام: اول آنکه برای بررسی دستور نیاز به حجم بزرگی از داده‌های زبانی داریم، چراکه باور دارم دستور را باید به شکلی کمّی مطالعه کرد؛ دیگر آنکه باید چگونگی کاربرد روشهای کمّی را برای تعیین درجات ارتباط میان دستگاههای گوناگون دستوری نشان داد (کاری که در پایان‌نامهٔ دکتری خود کرده‌ام)». (Halliday 1992, p. 611)

بخش بزرگی از زبان‌شناسان دیدگاهی همانند هلیدی دارند. همیشه یکی از آرزوهای زبان‌شناسان کاربردی و حتی بسیاری از نظریه‌پردازان این بوده است که به مقادیر بزرگی از داده‌های زبانی دسترسی داشته باشند.

«در دانش زبان، پیکره مجموعه‌ای از متون نوشتاری یا گفتاری آوانویسی شده است که می‌توان آن را به عنوان مبنایی برای تحلیل و توصیف زبانی به‌کار برد» (Kennedy 1998, p. 1).

پیکرهٔ زبانی می‌تواند بسیار بزرگ، فراگیر و نمایندهٔ تمامی یک زبان یا گونه‌ای از آن باشد؛ به شکل برگه‌های یادداشت یا پرونده‌های رایانه‌ای شامل متنهای کامل یا گزیده‌هایی از آنها، بخشهای پیوسته‌ای از متون یا گزیده‌ای از نقل‌قولها و نکات و حتی فهرستهای واژگانی باشد. پیکره می‌تواند ویژهٔ بررسی خاصی فراهم آید و یا دربرگیرندهٔ مجموعهٔ عظیم و بی‌ساختاری از متون گوناگون باشد که برای منظورهای گوناگون به‌کار رود. زبان‌شناسیِ پیکره‌ای بنیادی روش‌شناختی برای پژوهشهای زبانی به شمار می‌آید. در اصل و عملاً زبان‌شناسی پیکره‌ای به‌آسانی با شاخه‌های دیگر زبان‌شناسی می‌آمیزد. می‌توان با کمک پیکره به بررسیهای آوایی، نحوی، اجتماعی یا دیگر زمینه‌های زبان پرداخت و در این صورت می‌گوییم که روشها و فنون زبان‌شناسی پیکره‌ای را با موضوعات آوایی، نحوی و اجتماعی زبان و مانند آن آمیخته‌ایم. (Leech 1992, p. 106)

تنها رشتهٔ دیگر زبان‌شناسی که، مانند این رشته، با ابزار و روشهای مطالعه و نه با موضوعی خاص سروکار دارد زبان‌شناسی رایانه‌ای است که به عنوان مطالعهٔ زبان با کمک رایانه تعریف شده است. امروزه به نظر می‌رسد که این دو رشته با یکدیگر پیوند یافته‌اند. یعنی می‌توان این حوزه را **زبان‌شناسی پیکره‌ای رایانه‌ای**[۲۱] نامید، که در این صورت

21) computer corpus linguistics (CCL)

نه‌تنها روش نوین بررسی زبان بلکه فعالیت پژوهشی تازه‌ای با رویکردی فلسفی در زبان‌شناسی به شمار می‌آید (Ibid). لیچ ویژگیهای مهم این رشته را چنین برمی‌شمارد:

۱. تمرکز بر کنش زبانی و نه توانش زبانی؛
۲. تمرکز بر توصیف زبانی و نه بر همگانیهای زبان؛
۳. تمرکز بر الگوهای کمّی زبانی همانند الگوهای کیفی آن؛
۴. تمرکز بر دیدگاههای تجربی (و نه عقلانی) در بررسیهای علمی زبان.

همان‌گونه که مشاهده می‌شود، این ویژگیها مجموعه‌ای را به وجود می‌آورد که توجه بیشتری به جنبه‌های رفتاری زبان و بروز طبیعی گفتار و نوشتار دارد و عملاً در مقابل دیدگاههای چامسکی و پیروان وی قرار می‌گیرد. (Ibid, p. 107)

تویبرت نیز نگرشی همسو با لیچ نشان می‌دهد:

«زبان‌شناسی پیکره‌ای بر پایهٔ این باور که زبان اساساً پدیده‌ای اجتماعی است بنا نهاده شده است؛ پدیده‌ای که پیش از هر چیز می‌توان آن را با داده‌های تجربی آماده، یعنی در کنشهای ارتباطی مشاهده و توصیف کرد. متنهای مورد مشاهده، در اصل، کنشهای ارتباطی گذرا هستند». (Teubert 1991, p. 1)

از سوی دیگر، وی بررسی این پدیدهٔ اجتماعی را مستلزم دانستن چگونگی درک گوینده یا شنونده از مطالب نمی‌داند، زیرا زبان، به عنوان یک پدیدهٔ اجتماعی، به صورت متنی متجلی می‌گردد که می‌توان آن را مشاهده، ضبط، توصیف و تحلیل کرد.

زبان‌شناسیِ پیکره‌ای به توصیف تک‌تک زبانهای طبیعی می‌پردازد و نه همگانیهای زبان. از آنجا که نمی‌توان به درون ذهن افراد رخنه کرد، تنها می‌توان قراردادهای زبانی را در کنشهای ارتباطی و متون یافت. گرچه فرهنگهای لغت، کتابهای دستور و کتابهای درسی زبان نیز جزئی از فضای کلامی هستند، اما نمونه‌های واقعی از فضای کلامی و متنها بهتر می‌توانند واقعیات زبان را نشان دهند. زبان‌شناسی پیکره‌ای، با آمیختن سه روش، به فراهم آوردن دانش تجربی زبانی کمک می‌کند:

الف) استخراج خودکار داده‌های زبانی از پیکره‌ها؛
ب) پردازش برون‌داد با روشهای عمدتاً آماری؛
پ) ارزیابی و تفسیر این‌گونه داده‌های پردازش‌شده.

مراحل اول و دوم را می‌توان و باید به‌طور کامل با برنامه و خودکار انجام داد، اما

مرحلهٔ سوم نیاز به تصمیم‌گیری و منطق انسانی دارد. (Ibid)

پیکره‌های زبانی را می‌توان برای منظورهای گوناگون به کار گرفت، ازجمله برای فرهنگ‌نگاری، معناشناسی، بررسیهای دستوری، آموزش زبان و مانند اینها. پیکره‌ها را می‌توان از نظر اندازه و گستره به دسته‌های محدود، متوسط و عظیم تقسیم کرد.

۱-۲-۳ نشانه‌گذاری پیکره‌ها[۲۲]

برای گویاتر شدن پیکره و کاربردهای خاص، کدهای متفاوتی به آن افزوده می‌شود. این نشانه‌گذاری از یک سو می‌تواند برای ارتباط دادن بخشهای یک پیکره به ساختار کلی آن باشد، مانند شمارهٔ سطر، صفحه، فصل و مانند اینها و یا بافت زبانی را مشخص نماید مانند شرایط تولید زبانی، گونهٔ زبانی، رسانه و مانند آن. از سوی دیگر، نشانه‌گذاری می‌تواند صرفاً زبانی باشد. یکی از معدود کارهایی که در زبان فارسی برای برچسب‌دهی پیکره‌های فارسی انجام شده است، طراحی و اجرای برنامه‌ای رایانه‌ای برای برچسب‌دهی دستوری خودکار متون فارسی است. (Assi and H. Abdolhosseini 2000)

اکنون تنها به برخی از کاربردهای پیکره‌های زبانی اشاره می‌کنیم:

ـ یکی از مهم‌ترین کاربردهای پیکره در پردازش زبانِ طبیعی است. مهم‌ترین دستاورد این حوزه درک و بازشناسی گفتار بوده که تنها با بهره‌گیری از پیکره‌های بزرگ امکان‌پذیر گشته است.

ـ اکنون هیچ پروژهٔ فرهنگ‌نگاری پیشرفته‌ای نمی‌توان یافت که از پیکره‌های زبانی پایگاههای داده‌های زبانی بهره‌گیری نکند. نمونهٔ چنین کاربردی در زبان فارسی واژگان گزیدهٔ زبان‌شناسی است که نرم‌افزار رایانه‌ای آن نیز با امکانات گسترده آماده شده است (عاصی و عبدعلی ۱۳۷۵) و نمونهٔ دیگر فرهنگ فارسی به انگلیسی پیشرو آریان‌پور (چهارجلدی) است که با همکاری این نگارنده و بر بنیاد یک پیکرهٔ بزرگ دوزبانه تدوین گردیده است. (آریان‌پور و عاصی ۱۳۸۲)

ـ ایجاد پایگاههای داده‌های زبانی نیز جنبه‌ای دیگر از کاربرد پیکره‌های زبانی است که نمونه‌های متعدد آن را هم‌اکنون در سراسر جهان، به صورت پیوسته یا ناپیوسته، در اختیار داریم. چنین پایگاهی را برای زبان فارسی نیز نگارنده در پژوهشگاه علوم

22) corpus annotation

انسانی ایجاد نموده است. (عاصی ۱۳۷۶)

ـ طرحهای بررسی واژه‌های همایند[۲۳] در زبانهای گوناگون با کمک پیکره‌های زبانی اجرا شده است. نمونهٔ مهم و موفق آن فرهنگ واژه‌های همایند BBI برای زبان انگلیسی است. هم‌اکنون، در پژوهشگاه علوم انسانی نیز طرحی برای تدوین فرهنگ واژه‌های همایند فارسی بر اساس پایگاه داده‌های زبان فارسی در دست اجراست.

ـ برنامه‌های پایشگری زبان برای پی‌گیری و ردگیری تحولات زبانی نیز از امکانات پیکره‌های زبانی سود می‌برند. این‌گونه پیکره‌ها را پیکرهٔ پویا یا پیکرهٔ پایشگر می‌نامند. (Kennedy 1998, p. 22)

ـ همهٔ طرحهای ترجمهٔ ماشینی به گونه‌ای از پیکره‌های زبانی سود می‌برند، به‌ویژه سیستمهای جدید که با رویکردی آماری و پیکره‌بنیاد به‌تازگی از راه می‌رسند.

نمونه‌ای از پیکرهٔ زبانی که برای زبان فارسی فراهم شده است و اکنون در مرحلهٔ گسترش و تکمیل است، پایگاه داده‌های زبان فارسی است که نگارنده در پژوهشگاه علوم انسانی طراحی و اجرا نموده است.

۲-۲-۴ پایگاه داده‌های زبان فارسی[۲۴]

هدف از ایجاد پایگاه داده‌های زبان فارسی فراهم کردن پیکره‌ای مطلوب و با حجم عظیمی از داده‌های زبانی با گستردگی و گوناگونیهای بسیار و با ساختاری بسامان و منطقی است، تا امکان هرگونه جستجو و دستیابی سریع به آگاهیهای مورد نیاز را در هر زمان فراهم نماید. چنین پیکره‌ای می‌تواند همواره روزآیند گردد و پاسخگوی نیاز کاربران گوناگون در همهٔ زمینه‌های نظری و کاربردی باشد.

در نخستین مرحله، با توجه به نیازهای گوناگون پژوهشی و کاربردی، از طیف دورانهای تاریخی زبان فارسی، برش فارسی معاصر برگزیده شد. همین برش نیز، که به‌طور قراردادی از آغاز قرن چهاردهم خورشیدی تا امروز را در بر می‌گیرد، خود دارای گونه‌های بسیاری است، از جمله گونهٔ رسمی نوشتاری یا به‌اصطلاح فارسی معیار و گونهٔ گفتاری آن، گونه‌های ادبی و سبکی فارسی، گونه‌های محاوره‌ای و عامیانهٔ آن، و

23) collocations 24) Persian Linguistic Database

گونه‌هایی که متغیرهای زبانی و اجتماعی دیگری مانند سن، جنس، سواد و تحصیل، طبقهٔ اجتماعی، و محیط‌های مختلف ارتباطی عامل تمایز آنها به شمار می‌روند.

داده‌ها به شکلها و قالب‌بندیها[۲۵]ی گوناگون در این پایگاه ذخیره می‌شوند: به صورت متنهای پیوستهٔ کامل آثار ادبی یا نوشته‌های مهم، به صورت فهرستهای واژه‌نما و بسامدی از همین متنها و متنهای دیگر، یعنی فهرست همهٔ واژگان آنها به‌همراه چند سطر از بافت زبانی آنها و بسامدشان، و نیز به صورت واژه‌نامه‌های تک‌زبانه و دوزبانه. همچنین، متنهای آوانویسی‌شدهٔ داده‌های گفتاری چه به صورت متن پیوسته و چه به صورت فهرستهای بسامدی در پیکره جای دارند و پیش‌بینی شده، با به کارگیری امکانات چندرسانه‌ای[۲۶]، فراگوئیِ آوائیِ داده‌ها نیز ارائه گردد. از اطلاعات این پایگاه به روشهای گوناگون می‌توان بهره گرفت: هرگونه جستجو در پیکره، چه به صورت هم‌زمان یا برخط و چه به صورت سفارش و برون‌خط، بر پایهٔ هریک از اقلام اطلاعاتی و یا ویژگیهای مربوط به آنها از جمله

ـ جستجوی واژگانی (بر پایهٔ یک یا چند کلیدواژه)؛
ـ جستجوی مفهومی (بر پایهٔ مفهوم یا معنای مورد نظر)؛
ـ جستجوی تلفظی (بر پایهٔ صورت تلفظی یک واژه)؛
ـ جستجوی هم‌بافت (بر پایهٔ واژه‌های همایند و یا بافتهای همسایه)؛
ـ گشت و گذار[۲۷] در متنها و واژه‌نامه‌ها.

این جستجوها را می‌توان در محدوده‌های دلخواه (مثلاً دورهٔ زمانی معیّن، یا نویسنده‌های مشخص، یا حجم معیّنی از پیکره) انجام داد.

گزارشهای پایگاه به گونه‌های صوری و محتوائیِ مختلفی طراحی شده‌اند تا پاسخگوی نیازهای گوناگون باشند:

ـ به شکل فهرستهای واژگانی، آماری و بسامدی؛
ـ به شکل اطلاعات موردی؛
ـ به شکل فرهنگ واژه‌نما (واژهٔ مورد نظر در شکل کاربردی آن همراه با اطلاعاتی دربارهٔ بافت زبانی آن مانند چند سطر جملهٔ شاهد، شمارهٔ سطر و صفحهٔ متن، نام

25) formats 26) multimedia 27) navigation

نویسنده و مشخصات اثر، تاریخ کاربرد، بسامد در پیکره و مانند آن)؛

ـ به شکل گزیده‌هایی از متنهای گوناگون.

این پایگاه برای استفادهٔ همگانی در نظر گرفته شده است، اما مراحل و سطوح دستیابی آن متفاوت است.

پایگاههای داده‌ها روز به روز اهمیت بیشتری می‌یابند و شمار، موضوع و زمینه‌های کاربردشان گسترده‌تر می‌گردد. اکنون، از پایگاههای معرفتی[۲۸] گفتگو می‌شود که بسیاری از رشته‌های دانش و فن به آنها مجهز می‌شوند و همه‌گونه آگاهیها و معارف، به صورت الکترونیک، در آنها نگهداری می‌شود (انواری و فتحیان‌پور ۱۳۷۳). در شبکه‌های اطلاعاتی گوناگونی که در سراسر جهان در دسترس همه است، پایگاههای داده‌های بی‌شماری وجود دارد که، اگر ما نیازمند گونه‌ای اطلاع باشیم و آن را به‌درستی ارزیابی نماییم، می‌توانیم به‌خوبی از آن بهره‌مند شویم. از جمله دربارهٔ بسیاری از زبانهای مهم جهان داده‌های فراوانی گردآوری شده است. اما، در این دریای بی‌کرانِ اطلاعاتی، داده‌های قابل استناد برای زبان فارسی یافت نمی‌شود.

پایگاه داده‌های زبان فارسی در ایران و، در وهلهٔ نخست، برای پاسخگویی به نیازهای پژوهندگان ایرانی ایجاد شده است و، در مرحلهٔ بعد، به عنوان یک بانک اطلاعاتی ایرانی در دسترس همهٔ کسانی است که دربارهٔ زبان فارسی در نقاط دیگر جهان پژوهش می‌کنند.

برخی از طرحها و پژوهشهای نحوی دیگر که به زبان فارسی مربوط می‌شوند به شرح زیر است.

رضائی (Rezaei 1999) در پایان‌نامهٔ دکتری، نتیجهٔ سه تحقیق خود را منعکس کرده است. اول برای تقطیع جملات سادهٔ زبان فارسی سیستمی مبتنی بر شبکهٔ انتقالی برافزوده[۲۹] طراحی کرد. این تقطیع‌گر توالیهای ممکن درون‌بند ساده را تبیین می‌کند، اما قادر به تقطیع بندهای درونه‌ای نیست. بنا بر تحقیق بعدی وی، تقطیع‌گر قلب نحوی را نیز در بر می‌گیرد. ایشان در تحقیق آخر، پدیده‌هایی از قبیل برجسته‌سازی و جابه‌جائیِ بندهای متمم به آخر جمله را مطرح می‌کند. پدیده‌های زبانی، در دو تقطیع‌گر آخر وی،

28) knowledge base 29) Augmented Transition Network

در قالب نظریهٔ حاکمیت و مرجع‌گزینی توصیف می‌شود.

کشاورزی (۱۳۷۸) تقطیع‌گری برای تقطیع جملات سادهٔ خبری، بر اساس دستور گروه ساختی هسته‌بنیاد[۳۰] و الگوریتمی بالا به پایین، ارائه داده است. این تقطیع‌گر قادر به شناسائیِ 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 و از میان ترکیبها فعل مرکب و پیشوندی را به اجزای آنها تقطیع می‌کند. قواعد ساخت ۴۵۰ جمله و واژگان، برای تقطیع، به تقطیع‌گر داده شده است. تقطیع‌گر، پس از دریافت جملهٔ ورودی، درختی ارائه می‌دهد که ساخت نحوی جمله را در شش مرحله مشخص می‌کند.

طیبی (۱۳۷۴) چندین تلکس دریافتی سازمان هواپیمائیِ کشوری را که ساختاری ساده و عاری از ابهام دارند و به زبان انگلیسی‌اند انتخاب کرده است. سپس، با رویکرد دستور واژگانی نقشمند[۳۱]، ساخت هرکدام از جملات و ترجمهٔ آنها را به کمک رایانه ارائه داده است.

یونسی‌فر (۱۳۷۳) نیز تحقیقی انجام داده که، در آن، جملات انگلیسی با شبکهٔ خودکار پیشرو تجزیه می‌شوند و سپس ترجمه بر اساس روشهای نحوی انجام می‌گیرد. این کار بر پایهٔ نظریهٔ وابستگیِ مفهومی انجام شده است.

۲-۳ معناشناسی فارسی

معناشناسی[۳۲]، که به بررسی و توصیف معنای واژه‌ها و جمله‌های زبان می‌پردازد، پیشینه‌ای بسیار طولانی دارد و بیرون از حوزهٔ زبان‌شناسی ــ مانند فلسفه و روان‌شناسی ــ نیز مطرح بوده است. واژه‌ها واحدهای منفرد معنایی به شمار می‌آیند که در شکل‌دهی معنای جمله با کمک روابط نحوی شرکت می‌کنند. از سوی دیگر، هر جنبه‌ای از معنای واژه نیز به صورت طرحی خاص از هنجارهای معنایی، در بافتهای مناسبِ دستوری، نمود می‌یابد. مجموعهٔ روابط بهنجاری که یک واحد واژگانی در همهٔ بافتهای ممکن به

30) head-driven phrase structure grammar (HPSG)

31) Lexical Functional Grammar 32) semantics

وجود می‌آورد روابط بافتی[۳۳] نامیده می‌شود. از این رو، می‌توان گفت معنای یک واژه در روابط بافتی آن منعکس است (Cruse 1989, pp. 15, 16). معنای واژه را به‌طور کلی در دو لایهٔ معنای ادراکی یا مفهومی[۳۴] و معنای متداعی یا ضمنی[۳۵] در نظر می‌گیرند. معنای مفهومی بخشهای اساسی و ضروری معنای واژه را در بر می‌گیرد و معانی ضمنی یا متداعی مانند هاله‌ای آن را فرا می‌گیرد.

۲-۳-۱ مؤلفه‌های معنایی

یکی از روشهای تحلیل معنا، مشابه روشی است که در تحلیل آوایی و ساخت‌واژی زبان به کار می‌رفت و به تجزیه به مؤلفه‌ها یا مشخصه‌های معنایی[۳۶] معروف است. در این رویکرد، با بررسی مجموعه‌ای از واژه‌های مرتبط (مانند اصطلاحات خویشاوندی)، مشخصه‌های مهم و تمایزدهندهٔ معنا شناسایی و دسته‌بندی می‌شود و در جدولهای تحلیل معنایی قرار می‌گیرد:

	مرد	زن	دختر	پسر	گاو نر	میز
جاندار	+	+	+	+	+	−
انسان	+	+	+	+	−	−
مذکر	+	−	−	+	+	−
بزرگ‌سال	+	+	−	−	+	−

۲-۳-۲ روابط معنائیِ واژه‌ها

یکی از راههای توصیف و تحلیل معنا بررسی روابط مفهومی واژه‌ها و مقایسهٔ آنها با یکدیگر است. مهم‌ترین روابط معنایی عبارت‌اند از:

ـ هم‌معنایی[۳۷]: دو صورت زبانی متفاوت با معنای یکسان، گرچه معمولاً گفته می‌شود که هم‌معنائیِ مطلق کمتر وجود دارد، مانند کامپیوتر و رایانه؛

33) contextual relations
34) conceptual meaning
35) associative meaning
36) semantic feature/components analysis
37) synonymy

ـ **تضاد معنایی**[۳۸]: دو صورت با دو معنای متضاد، مانند خوب و بد؛

ـ **شمول معنایی**[۳۹]: معنای یک صورت زبانی معنای دیگری را در بر می‌گیرد و معمولاً رابطه‌های شمول معنایی سلسله‌مراتبی هستند. مانند حیوان و اسب؛

ـ **هم‌آوایی**[۴۰]: دو واژه با صورت آوائیِ یکسان و معنی متفاوت (ممکن است صورت نوشتاری آنها متفاوت باشد)، مانند خوار و خار؛

ـ **همنامی**[۴۱]: دو واژه با معنی متفاوت که صورت آوایی و نوشتاری آنها یکسان است، مانند دوش (= دیشب) و دوش (وسیله‌ای در حمام)؛

ـ **چندمعنایی**[۴۲]: یک واژه که دارای چندین معنی مرتبط با یکدیگر است، مانند دل به معنی «قلب»، «مرکز»، «میان»، «جرأت»، «شکم»، ....

و بسیاری روابط فرعی دیگر.

از دیدگاه زبان‌شناسی، ساخت و معنای واژه‌های زبان به‌طور عام در حوزهٔ واژه‌شناسی[۴۳] بررسی می‌گردد و ساختار معنایی و مفهومی واژگان فنی رشته‌های علمی (اصطلاحات[۴۴]) در حیطهٔ اصطلاح‌شناسی[۴۵] مورد بررسی قرار می‌گیرد.

۲-۳-۳ برخی پژوهشهای معنایی

عظیمی (۱۳۷۵) تولید و درک گفتار فارسی را مورد بررسی قرار می‌دهد. به نظر وی، انسان چیزی را می‌شنود که انتظار شنیدن آن را دارد. انسان برای درک گفتار طرف مقابل به دنبال سرنخهایی می‌گردد و، اگر آنها را بیابد، از جزئیات کلام صرف نظر می‌کند و به یک نتیجه‌گیری کلی مبادرت می‌ورزد. فهمیدن جملات موقعی مشکل می‌شود که یا این راهکارهای ادراکی مؤثر نیفتد یا جمله متضمّن مسائلی چون پردازش جملات پیچیده‌تر باشد. اگر جملات پیچیده باشند، احتمالاً قدم به قدم پردازش می‌شوند. وی نظریهٔ خلأیابی را نیز بررسی می‌کند. در این نظریه شنونده الفاظی را در حافظه نگه می‌دارد تا در بخشهای بعدیِ جمله به یک خلأ برسد و آنگاه لفظ را وارد خلأ کند. به نظر وی، نه‌تنها ساخت نحوی جمله بلکه عناصر واژگانی نیز به درک گفتار کمک می‌کند. علاوه بر اینها، که همه جنبهٔ زبانی دارند، مسائلی غیرزبانی نیز در این روند مؤثرند.

---

38) antonymy 39) hyponymy 40) homophony 41) homonymy
42) polysemy 43) lexicology 44) terms 45) terminology

شمس فرد (۱۳۷۴) در پایان‌نامهٔ کارشناسی ارشد خود طرحی برای درک متن فارسی بر پایهٔ نظریهٔ وابستگی مفهومی ارائه داده است. باقری (۱۳۷۵) نیز، با استفاده از قواعد تولیدی، جملات حوزهٔ خاصی را بر پایهٔ نظریهٔ وابستگی مفهومی تقطیع کرده است. تقطیع‌گر رئیس‌قاسم (۱۳۷۰) از دو قسمت نحوی و معنایی تشکیل شده است. قسمت نحوی آن شامل تمام توالیهای ممکن موضوعهای بندهای ساده است. قسمت معنایی هم شبکهٔ وابستگی مفهومی جملات را به دست می‌دهد.

نمونه‌های یادشده تنها شمار کوچکی از بررسیهای انجام‌شده را در بر می‌گیرد و کارهای بسیاری در حال حاضر در دست انجام است که هریک نیاز به معرفی مفصّل دارد و نتایج آنها در آینده نمودار خواهد گردید.

منابع

آریان‌پور، منوچهر و مصطفی عاصی (۱۳۸۲)، فرهنگ فارسی به انگلیسی پیشرو آریان‌پور، جهان رایانه، تهران؛
اسلامی، محرم (۱۳۷۹)، شناخت نوای گفتار زبان فارسی و کاربرد آن در بازسازی و بازشناسی رایانه‌ای گفتار، پایان‌نامهٔ دکتری، دانشگاه تهران، تهران؛
امامی، شیلا (۱۳۸۰)، بررسی و طبقه‌بندی تکواژهای زبان فارسی (بخش دوم)، پایان‌نامهٔ کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران مرکزی؛
انواری، مرتضی و ملک‌آفاق فتحیان‌پور (۱۳۷۳)، «پایگاههای معرفتی در سیستمهای اطلاع‌رسانی»، اطلاع‌رسانی، دورهٔ ۱۱، شمارهٔ ۱، ص ۶۶-۶۸؛
اگرادی، ویلیام و دیگران (۱۳۸۰)، درآمدی بر زبان‌شناسی معاصر، ترجمهٔ علی درزی، سمت، تهران؛
باقری، مسعود (۱۳۷۵)، استنباط موضوعات مشترک از جملات مرتبط به هم، پایان‌نامهٔ کارشناسی ارشد، دانشگاه صنعتی شریف، تهران؛
بقایی، بهروز (۱۳۸۰)، بررسی و طبقه‌بندی تکواژهای زبان فارسی (بخش اول)، پایان‌نامهٔ کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران مرکزی؛
دانش‌کار آراسته، پویان (۱۳۸۱)، نرم‌افزار تشخیص فعل در زبان فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه علامه طباطبایی، تهران؛
رئیس‌قاسم، محسن (۱۳۷۰)، پردازش زبان طبیعی و پردازش زبان فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه صنعتی شریف، تهران؛
سمائی، سید مهدی (۱۳۷۷)، واژگان در دستورسنج، انگارهٔ نظری، پایان‌نامهٔ دکتری، دانشگاه تهران، تهران؛
شمس فرد، مهرنوش (۱۳۷۴)، درک متون فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه صنعتی شریف، تهران؛

طیبی، اکرم (۱۳۷۴)، کاربرد دستور واژگانی نقشمند در ترجمهٔ ماشینی پاره‌ای از متون فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه تهران، تهران؛

عاصی، مصطفی (۱۳۷۳)، «طرحی برای تهیهٔ فرهنگهای تخصصی با کمک کامپیوتر»، مجموعهٔ مقالات دومین کنفرانس زبان‌شناسی نظری و کاربردی، دانشگاه علامه طباطبایی، تهران، ص۲۶۷-۲۸۵؛

ـــــ (۱۳۷۶)، «پایگاه داده‌های زبان فارسی»، مجموعهٔ مقالات سومین کنفرانس زبان‌شناسی، دانشگاه علامه طباطبایی و پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ص۲۰۵-۲۱۱؛

ـــــ (۱۳۸۲)، «از پیکرهٔ زبانی تا زبان‌شناسی پیکره‌ای» مجموعهٔ مقالات پنجمین کنفرانس زبان‌شناسی، دانشگاه علامه طباطبایی، تهران، ص۴۸۴-۴۹۵؛

عاصی، مصطفی و محمد عبدعلی (۱۳۷۵)، واژگان گزیدهٔ زبان‌شناسی، تهران، شرکت انتشارات علمی و فرهنگی؛

عظیمی اکبریه، محسن (۱۳۷۵)، تولید و درک گفتار با توجه به داده‌های زبان فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه فردوسی مشهد، مشهد؛

غلام‌پور، ایمان (۱۳۷۹)، بازشناسی گفتار مستقل از گوینده، پایان‌نامهٔ دکتری، دانشگاه صنعتی شریف، تهران؛

فرخ، ماندانا (۱۳۸۱)، بررسی ساختمان افعال ساده و مرکب فارسی و تدوین روشهای سرواژه‌سازی به کمک رایانه، پایان‌نامهٔ کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد تهران مرکزی؛

قاسمی، سید ضیاءالدین (۱۳۷۷)، اصول آواشناختی سنز گفتار فارسی، پایان‌نامهٔ کارشناسی ارشد، تهران، دانشگاه تهران؛

کشاورزی، نیما (۱۳۷۸)، تقطیع نحوی جملات سادهٔ فارسی بر اساس دستور گروه ساختی هسته‌بنیاد، پایان‌نامهٔ کارشناسی ارشد، تهران، پژوهشگاه علوم انسانی و مطالعات فرهنگی؛

ماه‌جانی، بهزاد (۱۳۷۸)، ارائهٔ یک مدل جهت نمایش اطلاعات مرتبط با نحو در مدخلهای واژگانی، پایان‌نامهٔ کارشناسی ارشد، تهران، دانشگاه تهران؛

یونسی‌فر، فریبا (۱۳۷۳)، پیاده‌سازی یک مترجم ماشینی به روش نحوی، پایان‌نامهٔ کارشناسی ارشد، تهران، دانشگاه صنعتی شریف.

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

Cruse, D.A. (1989), *Lexical semantics: Cambridge Textbooks in Linguistics*, Cambridge, Cambridge University Press;

Halliday, M.A.K. (1992), "Language as System and Language as Instance: The Corpus as a Theoretical Construct", *Directions in Corpus Linguistics*, Svartvik (ed.), Berlin, Mouton de Gruyter;

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*, London, Longman;

Leech, G. (1992), *Corpora and Theories of Linguistic Performance*, *Directions in Corpus Linguistics*, Svartvik (ed.), Berlin, Mouton de Gruyter;

Rezaei, Siamak (1992), *Linguistic and Computational Analysis of Word Order and Scrambling in Persian*, Ph.D. Dissertation, Edinburgh, University of Edinburgh;

Schuetze, Hinrich (1995), "Distributional Part-of-Speech Tagging", *From Texts to Tags: Issues in*

Multilingual Language Analysis, Online Proceedings of the ACL SIDGAT Workshop. On the Internet at http://www.lanl.gov/find/cmp.lg;

Svartvik (ed.) (1992), *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82 (Stokholm, 4-8 August 1991), Berlin, Mouton de Gruyter;

Teubert, W. (1999), "Corpus Linguistics: A Partisan View", *International Journal of Corpus Linguistics*, Vol. 4 / No. 1., pp. 1-10.