

پیکره‌های زبان فارسی

دکتر حیات عامری
عضو هیئت علمی دانشگاه تربیت مدرس

پیکره‌های زبان فارسی

دکتر حیات عامری

عضو هیئت علمی دانشگاه تربیت مدرس

۱۳۹۷



مرکز تحقیقات زبان و ادبیات فارسی

سرشناسه: عامری، حیات، ۱۳۵۱
عنوان و نام پدیدآور: پیکره‌های زبان فارسی؛ [برای] مرکز تحقیقات زبان و ادبیات فارسی
دانشگاه تربیت مدرس.
مشخصات نشر: تهران: دولت علم، ۱۳۹۷.
مشخصات ظاهری: ۱۹۸ ص.: جدول، نمودار.
شابک: 978-600-98180-4-4 ۲۷۰۰۰۰ ریال
وضعیت فهرست نویسی: فیپا
یادداشت: کتابنامه.
موضوع: زبان‌شناسی پیکره‌ای
موضوع: Corpora (Linguistics)
موضوع: فارسی
موضوع: Persian language
موضوع: پردازش زبان طبیعی
موضوع: Natural language processing (Computer science)
شناسه افزوده: دانشگاه تربیت مدرس. مرکز تحقیقات زبان و ادبیات فارسی
رده بندی کنگره: ۱۳۹۷ ۱۲۴۲/ز۲۸ P۱۲۸
رده بندی دیویی: ۴۱۰
شماره کتابشناسی ملی: ۵۲۰۱۲۲۳

انتشارات دولت علم

نام کتاب: پیکره‌های زبان فارسی
پدیدآور: دکتر حیات عامری
ویراستار: محمدعلی عبدالعظیمی
صفحه‌آرایی و گرافیک: حمیدرضا باباخانی
تاریخ انتشار: بهار ۱۳۹۷
شمارگان: ۵۰۰
قیمت: ۲۷۰۰۰۰ ریال
شابک: 978-600-98180-4-4
نشانی: تهران - خیابان انقلاب - روبروی دانشگاه تهران - پاساژ فروزنده - واحد ۴۱۲
Email: dolateelm@gmail.com
شماره تماس: ۶۶۴۷۷۶۲۵

کلیه حقوق این کتاب نزد ناشر محفوظ است.

فهرست

پیشگفتار.....	۱۵
فصل اول: کلیات.....	۱۹
پیکره.....	۲۱
حاشیه‌نویسی.....	۲۲
برچسب‌دهی پیکره.....	۲۲
انواع پیکره.....	۲۴
پیکره‌های متنی فارسی.....	۲۵
پیکره‌های گفتاری.....	۲۶
پیکره‌های نحوی.....	۲۶
مجموعه دادگان.....	۲۷
پیکره‌های موازی.....	۲۸
پیکره‌های سیستم‌های نویسه‌خوان نوری.....	۲۸
فصل دوم: پیکره‌های متنی فارسی.....	۲۹
مجموعه همشهری.....	۳۱
اسناد مجموعه.....	۳۲

۳۳	مقایسه نسخه‌های ۱ و ۲ مجموعه همشهری
۳۵	کاربردها
۳۵	پیکره بی جن خان
۳۶	پیکره متنی زبان فارسی
۳۶	مشخصات زبانی پیکره
۳۶	ویژگی‌های پیکره
۳۸	توکن‌بندی
۴۰	حاشیه‌نویسی
۴۰	مجموعه برچسب EAGLES
۴۱	برچسب‌دهی POS نیمه خودکار
۴۱	پایگاه دادگان زبان فارسی
۴۲	مشخصات زبانی
۴۳	منابع گردآوری داده‌ها
۴۴	ساختار زبانی پیکره
۴۴	ساختار رایانه‌ای
۴۵	ویژگی‌های سیستم جدید پایگاه داده‌ها
۴۶	موارد استفاده از پایگاه داده‌های زبان فارسی
۴۶	انواع جست‌وجو
۴۶	انواع گزارش‌ها
۴۶	کاربران پایگاه

۴۷	پرسیکا (پیکره متون خبری).....
۴۸	مراحل تهیه پیکره.....
۵۱	پاسخ (پیکره استاندارد سامانه های خلاصه ساز).....
۵۲	فرایند ساخت پیکره.....
۵۶	پیکره فارسی تحلیل احساس سستی پرس.....
۵۷	دادگان.....
۵۷	نشانه گذاری اسناد پیکره.....
۵۷	نمردهی.....
۵۸	اجزای اسناد پیکره.....
۵۹	ابزار نشانه گذاری.....
۶۰	پیکره فارسی ارزیابی سامانه های تقلب یاب.....
۶۰	مرحله پیش پردازش.....
۶۱	دسته بندی اسناد.....
۶۱	استخراج متون.....
۶۱	ایجاد تقلب.....
۶۲	وارد نمودن موارد تقلب در اسناد مشکوک.....
۶۳	مجموعه داده استاندارد وبلاگ های ایران.....
۶۳	مراحل ساخت مجموعه.....
۶۵	مجموعه محک وب دات آی آر.....
۶۷	مراحل ساخت پیکره استاندارد.....

تنزیل (پیکره قرآنی)	۶۹
خطاهای نگارشی در برخی متون موجود	۷۰
مراحل توسعه پیکره تنزیل	۷۰
پیکره نور	۷۱
مدل زبانی آماری	۷۲
آمارهای مربوط به پیکره نور	۷۲
توزیع زبانی پیکره	۷۳
طبقه‌بندی کتب پیکره	۷۳
مجموعه داده عروض (نسخه ۲,۰۰)	۷۳
وزن عروضی	۷۴
کاربرد سیستم	۷۵
به دست آوردن وزن عروضی یک شعر	۷۵
پیشنیازهای طرح	۷۷
مراحل سیستم	۷۸
الگوریتم‌های انطباق رشته	۷۸
ترکیب الگوریتم‌های انطباق رشته	۷۹
آزمایش‌های تجربی	۸۰
فصل سوم: پیکره‌های گفتاری	۸۱
فارسی‌دات	۸۳

تهیه دادگان	۸۴
تقطیع و برچسب‌دهی	۸۵
انسداد چاکنایی	۸۶
آمار کلی و نرم‌افزارهای پایگاه دادگان	۸۷
فارس‌دات تلفنی	۸۷
کاربرد	۸۹
گویش‌شناسی پیکره	۸۹
دادگان پیکره	۹۰
گویشوران	۹۰
محل ضبط صدا و تجهیزات	۹۱
حاشیه‌نویسی	۹۱
واژگان	۹۲
نرم‌افزار کاربر	۹۲
دادگان صوتی هجاهای فارسی	۹۳
مرز هجا	۹۵
نحوه تعیین هجاهای دادگان هجایی	۹۶
شیوه تعیین جایگاه مناسب استخراج هجاها	۹۶
ضبط کلمات حاوی عناصر دادگان و ملزومات رایانه‌ای	۹۷
هم‌زمانی سیگنال صحبت با منحنی ارتعاش تارآواها	۹۸
تقطیع کلمات و ایجاد بانک هجاها	۹۸

دادگان دایفونی فارسی	۹۹
دایفون	۱۰۰
تهیه پیکره جهت استخراج دایفون‌ها	۱۰۰
ضبط کلمات حاوی دایفون و ملزومات رایانه‌ای آن	۱۰۲
طراحی اصول استخراج (تقطیع) دایفون‌ها	۱۰۲
دادگان گفتار احساسی سه‌ند	۱۰۳
گویندگان	۱۰۴
ضبط داده‌ها	۱۰۴
آزمون شنیداری	۱۰۶
دادگان گفتار لهجه‌دار سه‌ند	۱۰۶
گویندگان	۱۰۷
ضبط داده‌ها	۱۰۷
فرمت نام فایل‌ها	۱۰۸
ارزیابی پاره‌گفتارها توسط گویندگان	۱۰۹
آزمون شنیداری	۱۰۹
دادگان تلفنی اعداد متصل	۱۰۹
گویندگان	۱۱۰
واژگان	۱۱۰
جمع‌آوری دادگان	۱۱۱
تصدیق صحت دادگان و آمایش آن	۱۱۱

آزمایش‌های انجام‌شده روی دادگان به وسیله سیستم‌بازشناسی گفتار مبتنی بر کلمه	۱۱۲
فصل چهارم: پیکره‌های نحوی	۱۱۳
پیکره نحوی وابستگی زبان فارسی (نسخه ۱, ۱, ۱)	۱۱۵
قالب‌بندی داده‌ها	۱۱۶
منابع متنی مورد استفاده در پیکره	۱۱۷
دستور وابستگی	۱۱۷
تجزیه وابستگی	۱۱۸
مجموعه برجسب‌های وابستگی موجود در پیکره	۱۱۹
برجسب اجزای سخن	۱۲۱
ابزارهای ایجادشده	۱۲۵
دادگان درختی فارسی در چارچوب دستور ساخت سازه‌ای هسته‌بنیان	۱۲۶
منبع داده‌ها و ابزار مورد نیاز	۱۲۷
فرایند خودراه‌اندازی	۱۲۸
روش توسعه بانک درختی	۱۲۸
پیکره درختی وابستگی فارسی اوپسالا	۱۳۰
انتخاب دادگان	۱۳۰
حاشیه‌نویسی نحوی	۱۳۱
تجزیه و راه‌اندازی	۱۳۱
مجموعه برجسب تعمیم یافته استنفورد	۱۳۲

فصل پنجم: مجموعه دادگان	۱۳۵
فارس‌نت	۱۳۷
وردنت	۱۳۸
مشخصات و شمول	۱۳۸
روش کلی	۱۳۹
منابع کلمات	۱۴۰
اسامی	۱۴۱
صفات	۱۴۴
افعال	۱۴۵
ابزارهای ایجادشده	۱۴۶
واژگان زبانی زبان فارسی	۱۴۷
مشخصه‌های واژگان زبانی	۱۴۷
تصریف کلمه در زبان فارسی	۱۴۹
برنامه واحدساز صرفی	۱۵۲
فرهنگ طیفی زبان فارسی	۱۵۲
یادداشت نمایه (ایندکس)	۱۵۴
فرهنگ جامع واژگان مترادف و متضاد زبان فارسی	۱۵۵
شیوه کار	۱۵۵
فرهنگ برخط واژگان مترادف و متضاد زبان فارسی	۱۵۸
فرهنگ ظرفیت نحوی افعال فارسی (نسخه ۳,۰)	۱۵۹

انواع ظرفیت‌های فعل در زبان فارسی	۱۶۱
مراحل تولید فرهنگ ظرفیت فعل	۱۶۲
فرهنگ املائی خط فارسی	۱۶۲
واژگان نحوی و معنایی افعال مرکب فارسی (نسخه ۱,۰)	۱۶۳
اطلاعات نحوی	۱۶۴
اطلاعات ریشه	۱۶۴
ساخت موضوعی و ساختار نحوی	۱۶۶
قالب زیرمقوله‌ای	۱۶۷
تغییرات ظرفیتی و هممعنایی	۱۶۸
اطلاعات معنایی	۱۶۸
فصل ششم: پیکره‌های موازی	۱۷۱
پیکره موازی انگلیسی - فارسی پیام	۱۷۳
جمع‌آوری داده‌ها	۱۷۴
آماده‌سازی پیکره	۱۷۵
هماهنگ‌سازی پیکره موازی	۱۷۶
یافتن معادل‌های ترجمه	۱۷۷
کاربردهای دیگر پیکره موازی	۱۷۷
پیکره فارسی ۱۹۸۴	۱۷۸
ویژگی‌های صرفی - نحوی	۱۸۰

۱۸۱	 پیکره
۱۸۲	 پیکره موازی انگلیسی - فارسی تهران
۱۸۳	 پیکره موازی انگلیسی - فارسی میزان
۱۸۳	 پیکره تطبیقی فارسی - انگلیسی دانشگاه تهران
۱۸۵	 فصل هفتم: پیکره‌های سیستم‌های نویسه‌خوان نوری
۱۸۷	 بانک اطلاعات حروف گسسته دست‌نویس فارسی
۱۸۸	 بخش‌های پیکره حذف
۱۸۸	 طراحی و اجرای پیکره
۱۹۱	 مجموعه ارقام دست‌نویس هدی
۱۹۱	 جمع‌آوری داده‌ها
۱۹۲	 نحوه استخراج ارقام
۱۹۲	 اصلاح دستی
۱۹۳	 انتخاب مجموعه آموزش و آزمون
۱۹۴	 منابع و مآخذ
۱۹۴	 منابع فارسی
۱۹۷	 منابع انگلیسی

پیشگفتار

طراحی پیکره یکی از زیرساخت‌های ضروری برای انجام تحقیقات زبانی و پردازش زبان طبیعی است. از کاربردهای پردازش زبان طبیعی می‌توان به ترجمه ماشینی، بازیابی اطلاعات، استخراج اطلاعات، خلاصه‌سازی خودکار، نویسه‌خوان نوری و بسیاری از کاربردهای دیگر اشاره کرد. برای هر یک از این اهداف به طراحی یک پیکره زبانی خاص در یکی از سطوح زبانی نیاز داریم. سطوح تحلیل زبان عبارت‌اند از سطح آواشناسی، واژگانی، نحوی، معناشناسی، گفتمان و کاربردشناسی. پس از طراحی چنین پیکره‌ای از ابزارهایی برای تحلیل زبانی این پیکره‌ها استفاده می‌شود. این ابزارها نرم-افزارهایی هستند که کار تحلیل زبانی را سریع‌تر و با دقت نسبتاً بالایی انجام می‌دهند.

امروزه زبان فارسی به عنوان زبان معیار در کشور ایران در تمام محافل ارتباطی با چالش‌هایی روبه‌روست. یکی از این چالش‌ها استفاده از زبان فارسی در فضای مجازی و تأثیری است که فضای مجازی بر آن می‌گذارد. زبان به‌مثابه یکی از ویژگی‌های خاص بشر همواره ماهیتی پویا و متغیر داشته و دارد. با تغییر فرهنگ و سبک زندگی بشر در طول تاریخ همواره زبان وی نیز دستخوش تغییر و دگرگونی بوده است. زبان فارسی هم در طی زندگی طولانی خود دچار دگرگونی‌های اساسی شده است؛ اما آنچه در این مجال باعث نگرانی است، تغییرات زبانی است که در فضای مجازی از روند طبیعی و تدریجی خود خارج شده است. در فضای مجازی عواملی در گسترش و تسریع این تغییرات دخیل‌اند که جنبه‌های مختلفی از فرهنگ و منش کاربران اینترنت را پوشش می‌دهند. خط و زبان فارسی به دلیل ویژگی‌های خاصی که دارد در فضای مجازی با مسائلی روبه‌روست که عدم توجه به آن‌ها ممکن است آسیب‌های جبران‌ناپذیری به این زبان وارد نماید. با بررسی مشکلات زبان فارسی در فضای مجازی،

ریشه‌یابی آن‌ها و ارائه راهکارهایی می‌توان تا حد زیادی بر این مشکلات غلبه و از زبان فارسی در برابر تغییرات ناگهانی، نادرست و سهل‌انگارانه محافظت کرد. یکی از راهکارهای اصلی برای به حداقل رساندن این آسیب‌ها تقویت و گسترش پیکره‌های زبان فارسی است.

دکتر عاصی از اوایل سال ۱۳۷۲، کار ایجاد پایگاه داده‌هایی را برای زبان فارسی در پژوهشگاه علوم انسانی آغاز کردند و تا سال ۱۳۷۸ دو مرحله (فاز) آن اجرا شد و مرحله سوم که مهم‌ترین فاز - یعنی گسترش و افزایش حجم داده‌ها و دگرگونی اساسی در نرم‌افزار و ایجاد امکانات نوین شبکه‌ای برای ارائه خدمات و اطلاعات آن در شبکه جهانی اینترنت بود - به دلیل نبود منابع مالی، چند سالی از اجرا بازماند تا اینکه با کمک مالی وزارت ارتباطات و فناوری اطلاعات اجرای فاز سوم طرح آغاز شد و به پایان رسید.

هدف از ایجاد پایگاه داده‌های زبان فارسی فراهم کردن پیکره‌ای مطلوب به دور از نارسایی است. پیکره‌ای که با وجود حجم عظیمی از داده‌های زبانی با گستردگی و گوناگونی‌های بسیار، ساختاری بسامان و منطقی داشته باشد تا امکان هرگونه جست‌وجو و دستیابی سریع به آگاهی‌های مورد نیاز را در هر زمان فراهم سازد. چنین پیکره‌ای می‌تواند همواره روزآیند شود و پاسخگوی نیاز همه پژوهندگان زبان فارسی و کاربران گوناگون در همه زمینه‌های نظری و کاربردی باشد (عاصی، ۱۳۸۴).

امروزه روش‌های آماری و مبتنی بر یادگیری ماشینی در پردازش زبان طبیعی و ایجاد سامانه‌هایی چون سامانه‌های ترجمه ماشینی، پرسش و پاسخ خودکار، تبدیل رایانه‌ای متن به گفتار و بالعکس، بازیابی اطلاعات و... کاربرد فراوانی یافته‌اند. یکی از ملزومات استفاده از روش‌های آماری در پردازش زبان طبیعی، دسترسی به داده‌های زبانی شامل پیکره‌های متنی، پیکره‌های درختی، واژگان، بانک‌های صوتی و... است و عدم دسترسی مناسب به چنین داده‌هایی مشکلات فراوانی را پیش پای پژوهشگران قرار می‌دهد.

از سوی دیگر بسیاری از زبان‌شناسان در پژوهش‌های خود از پیکره‌های زبانی بهره می‌گیرند و به بررسی ویژگی‌ها و کشف قواعد زبان از طریق اطلاعات موجود در داده‌های زبانی می‌پردازند.

محتوای الکترونیک زبان فارسی به سرعت در شبکه‌های مجازی و فضای وب در حال گسترش است و ضرورت دارد برای ساماندهی و نظارت بر آن‌ها اقدام مقتضی صورت گیرد.

این نوشتار مروری بر پیکره‌های به وجود آمده برای زبان فارسی و مراحل ساخت، ویژگی‌ها، امکانات، کاربردهای آن است و نیز ناکارآمدی‌ها و نواقص هر کدام از این پیکره‌ها را مورد بررسی قرار می‌دهد.

ضرورت انجام این پژوهش را زمانی احساس نمودم که در برخورد با دانشجویان رشته‌های مرتبط با زبان و نیز بسیاری از پژوهشگران و استادان این رشته‌ها، دریافت‌م که هیچ منبع قابل استفاده‌ای که در آن انواع پیکره‌های موجود برای زبان فارسی با ذکر ویژگی‌ها و کارآمدی‌ها یا نواقص هریک از این پیکره‌ها، به خوبی طبقه بندی و توصیف شده باشد، وجود ندارد. لذا این پژوهشگران برای انجام هر کار پژوهشی در این زمینه بدون آگاهی از وضعیت پژوهش‌های موجود، آغاز به کار می‌کنند. این امر بدون تردید آفت‌های زیادی برای علم ایجاد می‌کند و سبب پنهان ماندن حقایق بسیاری در مورد موضوع مورد پژوهش می‌شود. به علاوه این خطر نیز وجود دارد که پژوهشهای تکراری زیادی انجام پذیرد.

در فصل اول به تعریف پیکره و مسایل نظری مربوط به تهیه پیکره‌ها می‌پردازیم و در فصل‌های بعدی انواع پیکره‌های ایجاد شده برای زبان فارسی را تا سال ۱۳۹۴ به تفکیک نوع معرفی می‌کنیم و ویژگی‌های هر یک را شرح می‌دهیم.

نتیجه این پژوهش می‌تواند برای دانشجویان رشته‌های زبان‌شناسی، زبان انگلیسی، زبان روسی، زبان و ادبیات فارسی، زبان و ادبیات عربی و به طور کلی همه پژوهشگرانی که به نحوی به زبان فارسی در پژوهش‌های خویش توجه دارند، مفید باشد. در انتها لازم است از سرکار خانم فرزانه بختیاری به خاطر زحمات بسیاری که برای این پژوهش متحمل شدند، صمیمانه قدردانی کنم.

فصل اول

کلیات

پیکره

پیکره یکی از پروژه‌های زیرساختی در زمینه تحلیل زبانی و همچنین پردازش زبان طبیعی محسوب می‌شود و به حجم عظیمی از داده‌های زبانی گفته می‌شود که بر اساس معیارهای مشخص برای هدف معینی جمع‌آوری و ذخیره شده باشند، به طوری که نماینده زبان یا گویش مورد مطالعه‌اند (Atkins & Clear, 1992). معیارهای انتخاب زبانی که در طراحی یک پیکره در نظر گرفته می‌شوند عبارت‌اند از نوع متن (گفتاری، نوشتاری یا حتی الکترونیکی)، زبان یا گونه زبانی (گونه معیار^۱، فوق معیار^۲، زیرمعیار^۳)، نوع رسانه^۴ (کتاب، نشریه، آگهی و...)، محل تولید متن و بازه زمانی تولید متن. پیکره‌ها بر اساس اهداف پردازش زبان طبیعی در سطوح مختلف زبانی (آواشناسی^۵، تکواژی^۶، نحوی^۷، گفتمان^۸ و کاربردشناسی^۹) انجام می‌پذیرند و هر یک حجم متفاوتی از متون را دربرمی‌گیرد. پیکره‌های زبانی بر اساس هدف غایی خود انواع مختلفی دارند: پیکره نوشتاری، پیکره گفتاری، پیکره تاریخی، پیکره زبان کودک، پیکره چندزبانه^{۱۰}، پیکره زبان‌آموز^{۱۱}، پیکره موازی^{۱۲} و پیکره نحوی (بانک درختی^{۱۳}).

-
1. Standard
 2. Super-Standard
 3. Sub-Standard
 4. Medium
 5. Phonology
 6. Morphology
 7. Syntax
 8. Discourse
 9. Pragmatics
 10. Multilingual Corpus
 11. Learner Corpus
 12. Parallel Corpus
 13. Treebank

حاشیه‌نویسی^۱

برای مناسب‌تر کردن پیکره زبانی برای کاربرد مفیدتر در حوزه پردازش زبانی به حاشیه‌نویسی پیکره و کدگذاری^۲ پیکره می‌پردازند. این نشانه‌گذاری به دو صورت می‌تواند انجام شود: یا به شکل نشانه‌گذاری برای ارتباط دادن بخش‌های یک پیکره به ساختار کلی آن، یا نشانه‌گذاری زبانی که این امر عبارت است از افزودن برخی اطلاعات در مورد نقش کلمات^۳ یا ریشه کلمات^۴ موجود در یک زبان.

نشانه‌گذاری زبانی ممکن است در سه سطح انجام پذیرد:

۱. در سطح صورت‌های کاربردی^۵: در این مورد برای هر صورت کاربردی یک نشانه (کد) در متن افزوده می‌شود.

۲. در سطح پاره‌های زبانی^۶: که نشانه مربوط به توالی چند صورت کاربردی در متن است.

۳. در سطح تداعی^۷: که نشانه‌های ارتباطی و تداعی میان صورت‌های زبانی و پاره‌های زبانی را در متن مشخص می‌سازد.

به پیکره‌هایی که در سطح صورت‌های کاربردی نشانه‌گذاری می‌شوند، برچسب دار^۸ یا برچسب‌دهی شده می‌گویند و این نشانه‌ها می‌توانند نوع دستوری واژه یا برخی ویژگی‌های معنایی آن را نشان دهند (عاصی، ۱۳۸۲).

برچسب‌دهی پیکره

نشانه‌گذاری در سطح پاره زبانی را برچسب‌دهی نحوی می‌گویند و می‌توان ارتباط نحوی یا هم‌آیی یا لانه‌گیری واحدها را با آن‌ها نمایش داد. پیکره‌هایی را که تجزیه نحوی شده‌اند بانک‌های درختی می‌نامند و در امر آموزش ابزارهای پردازش زبان طبیعی کاربرد دارند.

-
1. Annotation
 2. Encoding
 3. Pos Tagging
 4. Lemma
 5. Tokens
 6. Segments
 7. Associative Level
 8. Tagged Corpus

بنابراین پیکره‌های به‌وجودآمده برای هر زبانی دو گونه‌اند: پیکره‌های خام و پیکره‌های برچسب‌خورده. پیکره‌های برچسب‌خورده دارای اطلاعات نحوی، صرفی، معنایی و گفتمان می‌باشند و بیشتر در امر آموزش ابزارهای پردازش زبان طبیعی کاربرد دارند. برچسب‌گذاری پیکره‌ها یکی از روش‌های حاشیه‌نویسی محسوب می‌شود. با استفاده از برچسب‌گذاری، قواعد و دستور زبان را می‌توان به مدلی آماری تبدیل کرد و آن را قابل استفاده در زبان‌شناسی رایانه‌ای نمود.

یکی از زیرساخت‌های لازم برای طراحی چنین پیکره‌های حاشیه‌نویسی‌شده و کامل، تعریف مجموعه برچسب‌هایی است که در جهت هدف نهایی یک پیکره به وجود آمده باشند. این مجموعه برچسب‌ها باید دو ویژگی عمده داشته باشند: اول اینکه باید مطابق با استانداردهای برچسب‌گذاری سایر زبان‌ها باشد؛ دیگر اینکه بتواند موارد خاص زبانی زبان مورد تحلیل را نیز پوشش دهد.

برچسب‌گذاری در چهار سطح انجام می‌شود:

۱. تعیین برچسب مقوله کلمه؛
 ۲. برچسب‌گذاری نحوی؛
 ۳. برچسب‌گذاری معنایی؛
 ۴. برچسب‌گذاری کاربردشناختی.
- مجموعه برچسب‌ها چند نوع‌اند:

۱. برچسب‌های نحوی - ساختوازی^۱ که شامل مقوله‌های نحوی اصلی همچون فعل، اسم، صفت، قید، حرف اضافه و... می‌باشد.

۲. برچسب‌های خاص^۲ که خاص کلماتی است که در طبقه مقوله‌های اصلی قرار نمی‌گیرد اما تعیین برچسب آن‌ها در استخراج اطلاعات زبانی از پیکره حائز اهمیت است.

۳. برچسب‌های متفرقه^۳ که در برچسب‌های طبقه اصلی قرار نمی‌گیرند و به عنوان کلمات مجزا، کلمات خارجی، نشانه‌ها، علائم ریاضی و اختصاری در پیکره وجود دارند (Cloern, 1999).

1. Morphosyntactic Tagging

2. Unique Tags

3. Residual / Miscellaneous Tags