

پیکره‌های زبان فارسی

دکتر حیات عامری

معاون 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: دبیرخانه تربیت مدرس. مرکز تحقیقات زبان و ادبیات فارسی

رده بندی کنگره: ۲۱۰۷/۳۷۸

رده بندی دیویی: ۴۱۰

شماره کتابشناسی ملی: ۸۲۰۱۲۲۳

انتشارات دولت علم

نام کتاب: پیکره‌های زبان فارسی

پدیدآور: دکتر حیات عامری

ویراستار: محمدعلی عبدالعظیمی

صفحه آرایی 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

Email: dolateelm@gmail.com

تلفن: ۶۶۴۷۷۶۲۵

واحد ۴۱۲

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	ترکیب‌های
۴۰	حاشیه‌نویس
۴۰	مجموعه برچسب EAGLE
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 سنتی پرس.....
۵۷	دادگان.....
۵۷	نشان گذاری اسناد پیکره.....
۵۷	نمره دهی.....
۵۸	اجزای اسناد پیکره.....
۵۹	ایزار نشانه گذاری.....
۶۰	پیکره فارسی ارزیابی سامانه های تقلب دار.....
۶۰	مرحله پیش پردازش.....
۶۱	دسته بندی اسناد.....
۶۱	استخراج متون.....
۶۱	ایجاد تقلب.....
۶۱	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.....

۶۹	تنزیل (پیکره قرآنی)
۷۰	خطاهای نگارشی در برخی متون موجود
۷۰	مراحل توسعه پیکره تنزیل
۷۱	پیکره نور
۷۲	مدل زبانی آماری
۷۲	آمارهای مربوط به پیکره نور
۷۳	تاریخچه بانی پیکره
۷۳	طبقه بندی کتب پیک
۷۳	مجموعه داده عروض (۲، ۱۰، ۲۰)
۷۴	وزن عروضی
۷۵	کاربرد سیستم
۷۵	به دست آوردن وزن عروضی یک شعر
۷۷	پیشنیازهای طرح
۷۸	مراحل سیستم
۷۸	الگوریتم های انطباق رشته
۷۹	ترکیب الگوریتم های انطباق رشته
۸۰	آزمایش های تجربی
۸۱	فصل سوم: پیکره های گفتاری
۸۳	فارس دات

تهیه دادگان	۸۴
تقطیع و برجسبدهی	۸۵
انسداد چاکنایی	۸۶
آمار کلی و نرم افزارهای پایگاه دادگان	۸۷
فارس دات تلفنی	۸۷
کاربرد	۸۹
ریش سناسی پیکره	۸۹
دادگان پیک	۹۰
گویشوران	۹۰
محل ضبط صدا و تجهیزات	۹۱
حاشیه نویسی	۹۱
واژگان	۹۲
نرم افزار کاربر	۹۲
دادگان صوتی هجاهای فارسی	۹۳
مرز هجا	۹۵
نحوه تعیین هجاهای دادگان هجایی	۹۵
شیوه تعیین جایگاه مناسب استخراج هجاها	۹۶
ضبط کلمات حاوی عناصر دادگان و ملزومات رایانه ای	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 سه‌د	۱۰۳
گریند	۱۰۴
ضبط داده‌ها	۱۰۴
آزمون شنیداری	۱۰۶
دادگان گفتار لهجه‌دار سه‌د	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 سیستم

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

فصل پنجم: مجموعه دادگان	۱۳۵
فارسانت	۱۳۷
وردنت	۱۳۸
مشخصات و شمول	۱۳۸
روش کلی	۱۳۹
منابع کلمات	۱۴۰
سام	۱۴۱
صفات	۱۴۴
افعال	۱۴۵
ابزارهای ایجادشده	۱۴۶
واژگان زبانی فارسی	۱۴۷
مشخصه‌های واژگان زبا	۱۴۷
تصرف 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

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

- 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 الف
- 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 زبانی 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. از 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، گرامر و کاربردشناسی. پس از طراحی چنین پیکره‌ای از ابزارهایی برای تحلیل زبانی، این پیکره‌ها استفاده می‌شود. این ابزارها نرم‌افزارهایی 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 مبتنی بر یک سری 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 جبران ناپذیری 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، ساختار منطقی و سازمان‌یافته باشد تا امکان هرگونه جست‌وجو و دستیابی سریع به آگاهانه مورد نیاز را در هر زمان فراهم سازد. چنین پیکره‌ای می‌تواند همواره روزآمد 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 ماشین 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 فراوان 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 به‌وجودآمده برای زبان فارسی و مراحل ساخت، ویژگی‌ها، امکانات، کاربردهای آن است و نیز ناکارآمدی‌ها و نواقص هرکدام 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 مابلی استفاده‌ای که در آن انواع پیکره‌های موجود برای زبان فارسی با ذکر ویژگی‌ها و کارآمدها یا نواقص هریک از این پیکره‌ها، به خوبی طبقه‌بندی 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 هریک را شرح می‌دهیم.

نتیجه این پژوهش می‌تواند برای دانشجویان رشته‌های زبان‌شناسی، زبان انگلیسی، زبان روسی، زبان و ادبیات فارسی، زبان و ادبیات عربی و به‌طورکلی 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.