فرهنگنامه‌های زبان‌شناسی ۱۸

فرهنگ توصیفی زبان‌شناسی پیکره‌ای

آزاده میرزائی

فرهنگ توصیفی زبان‌شناسی پیکره‌ای

دکتر آزاده میرزائی

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی
انتشارات

سرشناسه : میرزائی، آزاده،
عنوان و نام پدیدآور : فرهنگ توصیفی زبان‌شناسی پیکره‌ای/ آزاده میرزائی.
مشخصات نشر : تهران: نشر علمی، ۱۳۹۹.
مشخصات ظاهری : ۲۹۴ ص.: جدول.؛
شابک : ۹۷۸-۹۶۴-۴۰۴-۴۸۴-۷
وضعیت فهرست نویسی : فیپا
یادداشت : واژه‌نامه.
یادداشت : کتابنامه: ص. ۲۴۵ – ۲۵۰.
موضوع : زبان‌شناسی پیکره‌ای -- اصطلاح‌ها و تعبیرها .
موضوع : Corpora (Linguistics) -- Terminology
موضوع : زبان شناسی پیکره‌ای -- واژه‌نامه‌ها -- فارسی
موضوع : Corpora (Linguistics) -- Dictionaries -- Persian
موضوع : فارسی -- واژه‌نامه‌ها -- انگلیسی
موضوع : Persian language -- Dictionaries -- English
رده بندی کنگره : P۱۲۸
رده بندی دیویی : ۴۱۰/۱۸۸
شماره کتابشناسی ملی : ۷۴۴۳۷۴۹

خیابان انقلاب- خیابان۱۲ فروردین- خیابان شهدای ژاندارمری- پلاک۱۰۳
تلفن: ۱۲و۶۶۴۶۰۵۱۱-۶۶۴۶۳۰۷۲
www.elmipublications.com

فرهنگ توصیفی زبان‌شناسی پیکره‌ای

دکتر آزاده میرزائی
(دانشگاه علامه طباطبائی)

چاپ اول : ۱۴۰۰
تیراژ : ۴۰۰ نسخه
لیتوگرافی :کوثر
چاپ : مهارت
شابک: ۹۷۸-۹۶۴-۴۰۴-۴۸۴-۷
مرکز پخش: خیابان انقلاب- خیابان ۱۲فروردین- خیابان شهدای ژاندارمری- پلاک ۱۰۳
تلفن: ۱۲و ۶۶۴۶۰۵۱۱ -۶۶۴۶۳۰۷۲

یادداشت ناشر

بومی‌سازی هر دانشی، زمانی تحقق می‌یابد که واژگان فنی آن علم به صراحت توصیف شده باشند و از هیئتی برخوردار شوند که بتوان به‌گونه‌ای یکدست و هماهنگ به‌کارشان برد. درکنار دو مجموعه‌ی «کهن نامه‌های زبان‌شناسی» و «نگین‌های زبان‌شناسی»، سعی بر این بوده است تا با مجموعه‌ی تازه‌ای به نام «فرهنگنامه‌های زبان‌شناسی» به این هدف دیرینه، یعنی همانا دستیابی به آرمان تمامی زبان‌شناسان ایران که چیزی جز «زبان‌شناسی در خدمت زبان‌های ایران» نبوده و نیست، تحقق ببخشیم.

محمدعلی علمی

پیشگفتار

زبان‌شناسی پیکره‌ای حوزهٔ بینارشته‌ای است که از یک‌سو با زبان‌شناسی و حوزه‌های وابستهٔ آن در ارتباط است و از سوی دیگر با علوم رایانه، هوش مصنوعی، پردازش زبان طبیعی و علم آمار مرتبط می‌شود. در این حوزهٔ بینارشته‌ای، گردآوری، آماده‌سازی و بهره‌برداری از داده‌های طبیعی زبان با استفاده از امکانات رایانه و محاسبات آماری مختلف، با ایجاد امکانات پژوهشی و یا پاسخ به پرسش‌های زبان‌شناسی نظری، زبان‌شناسی کاربردی، آموزش زبان، ترجمه، فرهنگ‌نویسی، سبک‌شناسی، تحلیل گفتمان و محتوای علوم انسانی، زبان‌شناسی اجتماعی، روان‌شناسی زبان و همچنین هوشمندسازی ماشین در جهت پردازش و تولید زبان طبیعی انسان با اهداف مختلفی چون متن‌کاوی، داده‌کاوی، خلاصه‌سازی، تحلیل محتوا، سامانه‌های پرسش و پاسخ و مانند آن صورت می‌گیرد. بدیهی است که در یک حوزهٔ بینارشته‌ای، ایجاد زبانی مشترک میان پژوهشگران حوزه‌های وابسته از اهمیت ویژه‌ای برخوردار است. بر این اساس تهیه فرهنگ توصیفی زبان‌شناسی پیکره‌ای از یک‌جهت برای ایجاد زبان مشترک میان پژوهشگران مختلف بسیاری ضروری است و از سوی دیگر کاری دشوار می‌نماید.

در این فرهنگ تلاش شده است مفاهیم مرتبط زبان‌شناسی نظری، کلیدواژه‌های اختصاصی زبان‌شناسی پیکره‌ای و برخی اصطلاحات مهم زبان‌شناسی رایانشی که در

ارتباط با این حوزه قرار می‌گیرد، پوشش داده شود. از سوی دیگر در موضوع زبان‌شناسی پیکره‌ای تلاش شده است تا منابع و داده‌های زبانی، ابزارهای مختلف پیکره‌ای چه در زمینهٔ تولید و آماده‌سازی پیکره‌ها و چه در بحث تحلیل پیکره‌ها جهت آشنایی پژوهشگران، هم به‌منظور بهره‌برداری از امکانات موجود در این حوزه و هم به‌منظور الگوبرداری در پژوهش‌های مشابه، معرفی شوند. همچنین تلاش شده است که علاوه بر معرفی مهم‌ترین منابع غیرفارسی، پوشش خوبی از پیکره‌ها و داده‌های زبان فارسی هم به دست داده شود.

مدخل‌های این فرهنگ به ترتیب الفبای فارسی مرتب شده‌اند. برای دسترسی به فهرست مدخل‌ها و معادل‌های آن در زبان انگلیسی، واژه‌نامه‌ای از فارسی به انگلیس و بالعکس در پایان ارائه شده است. همچنین ازآنجاکه بسیاری از ابزارها، منابع و پیکره‌های زبانی با واژه‌های سرنام و اختصار معرفی و شناخته می‌شوند، در ابتدا اختصارات و معادل‌های فارسی و انگلیسی آن در جدولی معرفی شده‌اند.

در این فرهنگ، هر جا مدخلی از زبان انگلیسی بیش از یک معادل در زبان فارسی داشته است، رایج‌ترین معادل فارسی معرفی شده و در ترتیب الفبایی، معادل‌های دیگر بدون آنکه دوباره تعریف شوند، در جای خود آمده و به مدخل تعریف‌شده ارجاع شده‌اند.

در مواردی که تعریف یک مدخل وابسته به تعریف مدخل دیگر است و یا با تعریف آن، تعریفش کامل‌تر می‌شود، ذیل مدخل حاضر، به مدخل مرتبط، با علامت پیکان (←) ارجاع داده شده است.

برای گردآوری مدخل‌های این فرهنگ از منابع مرجع زبان‌شناسی پیکره‌ای و در بسیاری از موارد از فرهنگ توصیفی زبان‌شناسی پیکره‌ای و راهنمای مرجع زبان‌شناسی پیکره‌ای استفاده شده است، همچنین معرفی کلیدواژه‌های زبان‌شناسی رایانشی بر اساس برخی منابع مرتبط صورت گرفته است. در معرفی پیکره‌ها و منابع زبان فارسی از وبگاه‌های آن منابع و همچنین سه دوره مجموعه مقالات همایش‌های زبان‌شناسی پیکره‌ای استفاده شده است. در معرفی ابزارها و منابع زبان‌های دیگر هم با مراجعه به وبگاه‌های مرجع اطلاع‌رسانی ابتدا مدخل‌های مرتبط آماده و سپس با رجوع به وبگاه هر منبع یا ابزار، معرفی آن صورت گرفته است. فهرست کتاب‌های منبع در پایان فرهنگ و آدرس وبگاه‌ها ذیل هر مدخل آمده است.

تهیه این فرهنگ به پیشنهاد استاد بزرگوار جناب آقای دکتر کورش صفوی صورت گرفت. از ایشان صمیمانه سپاسگزارم. همچنین از جناب آقای محمدعلی علمی که امکان انتشار این مجموعه را فراهم کردند تشکر می‌کنم.

آزاده میرزائی

پاییز ۱۳۹۹

فهرست اختصارات

ACE	Australian Corpus of English	پیکره استرالیایی زبان انگلیسی
ACL	Association for Computational Linguistics	انجمن زبان‌شناسی رایانشی
ANC	American National Corpus	پیکره ملی انگلیسی آمریکایی
ASCII	American Standard Code for Information Exchange	کد استاندارد آمریکایی برای تبادل اطلاعات (اسکی)
BAS	Bavarian Archive for Speech Signals	آرشیو باواریایی برای سیگنال صوتی
BNC	British National Corpus	پیکره ملی انگلیسی بریتانیایی
BoE	Bank of English	پیکره بانک انگلیسی
CADS	Corpus-assisted discourse studies	پژوهش‌های گفتمانی پیکره‌یار
CAI	Computer-Aided Instruction	دستورالعمل‌های رایانه‌یار

CALI	Computer-Aided Language Instruction	دستورالعمل‌های زبانی رایانه‌یار
CALL	Computer Assisted Language Learning	آموزش زبان با کمک رایانه (کال)
CANCODE	Cambridge and Nottingham Corpus of Discourse in English	پیکره گفتمانی زبان انگلیسی کمبریج و ناتینگهام
CIDE	Collaborative International Dictionary of English	فرهنگ لغت بین‌المللی مشارکتی زبان انگلیسی
CLARIN	Common Language Resources and Technology Infrastructure	منابع زبانی مشترک و زیرساخت‌های فناورانه
COBUILD	Collins Birmingham University International Language Database	پیکره کوبیلد
COCA	Corpus of Contemporary American English	پیکره زبان انگلیسی آمریکایی معاصر (پیکره کوکا)
COLT	Bergen Corpus of London Teenage English	پیکره برگن از نوجوانان انگلیسی‌زبان
CoNLL	Conference on Natural Language Learning	کنفرانس یادگیری زبان طبیعی
CORE	Corpus of Online Registers of English	پیکره کُر

CSLU	Center for Spoken Language Understanding	مرکز درک زبان گفتار
CSTR	Centre for Speech Technology Research	مرکز تحقیقات فناوری‌های گفتار
DDL	data-driven learning	یادگیری مبتنی‌بر داده/ یادگیری داده‌محور/ یادگیری اکتشافی
DIRNDL	Discourse Information Radio News Database for Linguistic Analysis (DIRNDL) corpus	پیکره درندل
dotIR	WebIR (dotIR)	مجموعه وب دات‌آی‌آر
EACL	European Chapter of the Association for Computational Linguistics	بخش اروپایی انجمن زبان‌شناسی رایانشی
EAGLES	Expert Advisory Group on Language Engineering Standards	گروه مشورتی تخصصی استانداردهای مهندسی زبان
ECI	European Corpus Initiative	نهاد پیکره‌ای اروپا
ElAN	Eudico Linguistic Annotator	نرم‌افزار الن
ELDA	Evaluations and Language Resources Distribution Agency	آژانس ارزیابی و انتشار منابع زبانی (الدا)
ELRA	European Language Resources Association	انجمن منابع زبانی اروپا

ESFSLD	European Science Foundation Second Language Databank	بانک دادهٔ زبان دوم بنیاد علوم اروپا
FLEx	Fieldworks Language Explorer	نرم‌افزار فلکس
GATE	General Architecture for Text Engineering	نرم‌افزار معماری عمومی مهندسی متن (گِیت)
GloWbE	Corpus of Global Web-Based English	پیکره مبتنی بر وب جهانی زبان انگلیسی
HTML	Hypertext Markup Language	زبان نشانه‌گذاری فرامتن
ICAME	International Computer Archive of Modern and Medieval English	آرشیو بین‌المللی رایانه‌ای انگلیسی مدرن و قرون وسطایی
ICE	International Corpus of English	پیکره بین‌المللی زبان انگلیسی
irBlogs	irBlogs	مجموعه داده استاندارد وبلاگ‌های ایران
IviE	Intonational Variation in English	پیکره تنوعات آهنگین زبان انگلیسی
KWIC	key word in context	واژهٔ کلیدی در بافت
LCPW	Lancaster Corpus of Children's Project Writing	پیکره نوشتار کودکان لنکستر
LDC	Linguistic Data Consortium	کنسرسیوم داده‌های زبان‌شناختی
LLC	London–Lund Corpus	پیکره لاندن-لوند
LOB	Lancaster–Oslo/Bergen	پیکره لنکستر-اسلو/برگن

MICASE	Michigan Corpus of Academic Spoken English	پیکره گفتاری آکادمیک زبان انگلیسی میشیگان
NER	named entity recognition	تشخیص موجودیت نامدار
NLP	natural language processing	پردازش زبان طبیعی
NOW	News on the Web corpus	پیکره خبرهای محتوای وب
OCR	optical character recognition	نویسه‌خوان نوری
OEC	Oxford English Corpus	پیکره انگلیسی آکسفورد
OLAC	Open Language Archives Community	انجمن آرشیو زبانی آزاد
OTA	Oxford Text Archive	آرشیو متنی آکسفورد
PCC	Potsdam Commentary Corpus	پیکره معاصر پوتسدام
PDTB	Penn Discourse Treebank	درخت‌بانک گفتمان پن
PerDTB	Persian discourse treebank	درخت‌بانک گفتمان زبان فارسی
PEPC	Parallel English-Persian Corpus Extracted from Wikipedia	پیکره موازی فارسی-انگلیسی ویکی‌پدیا
PerPB	Persian Proposition Bank	پیکره گزاره‌های معنایی زبان فارسی
PLC	Persian Learner Corpus	پیکره زبان‌آموز فارسی
PLDB	Persian Linguistic Database	پایگاه داده‌های زبان فارسی

POS	part-of-speech tagging	برچسب‌گذاری مقولهٔ دستوری
SEU	Survey of English Usage	پیکره بررسی کاربرد زبان انگلیسی
SFLC	Salam Farsi Learner Corpus	پیکره سلام فارسی
SGML	Standard Generalised Markup Language	زبان نشانه‌گذاری تعمیم‌یافتهٔ استاندارد
SUSANE	Surface and Underlying Structural Analyses of Naturalistic English	پیکره تحلیل روساختی و زیرساختی ساختار دادهٔ طبیعی انگلیسی
T2K-SWAL	TOEFL 2000 Spoken and Written Academic Language	پیکره زبان نوشتاری و گفتاری تافل ۲۰۰۰
TELC	Thai English Learner Corpus	پیکره انگلیسی‌آموزان تایلندی
TF-IDF	Term Frequency-Inverse Document Frequency	معیار وزن‌دهی تی‌اِف-آی‌دی‌اِف
TLG	Thesaurus Linguae Graecae	پیکره تی‌ال‌جی
TnT	Trigrams'n'Tags	برچسب‌های سه‌گرمی
UCREL	University Centre for Computer Corpus Research on Language	مرکز دانشگاهی پژوهش‌های رایانه‌ای پیکره‌ای زبانی
UTPECC	University of Tehran Persian-English Comparable Corpus	پیکره تطبیقی فارسی-انگلیسی دانشگاه تهران
XML	Extensible Markup Language	زبان نشانه‌گذاری گسترش‌پذیر

YCOE	York–Toronto–Helsinki Corpus of Old English Prose	پیکره منثور زبان انگلیسی قدیم هلسینکی-تورنتو-یورک
ZEN	Zürich English Newspaper Corpus	پیکره اخبار انگلیسی زوریخ
حدف		بانک اطلاعات حروف گسسته دست‌نویس فارسی

آ

آرشیو archive

مجموعه‌ای از متون و یا اسنادی است که در حجمی انبوه در یک انباره نگهداری می‌شوند. پیکره را از این نظر می‌توان با آرشیو هم‌معنی دانست اما تفاوت اصلی این دو در آن است که پیکره‌ها داده‌های ساختاریافته‌ای هستند که از ویژگی نمایندگی و توازن برخوردارند اما بایگانی‌ها و آرشیوها تنها انباره‌ای از داده‌های زبانی هستند.

آرشیو باواریایی برای سیگنال صوتی

Bavarian Archive for Speech Signals (BAS)

آرشیوی از پیکره‌ها، داده‌های گفتاری و ابزارها تحلیلی است که در دانشگاه مونیخ آلمان نگهداری می‌شود. هدف اصلی گردآوری این آرشیو ارائهٔ داده‌های گفتاری به پژوهشگران و علاقه‌مندان این حوزه است.

آرشیو بین‌المللی رایانه‌ای انگلیسی مدرن و قرون وسطایی

International Computer Archive of Modern and Medieval English (ICAME)

ICAME سازمانی بین‌المللی، متشکل از زبان‌شناسان و دانشمندان علم اطلاعات است که بر روی متون ماشین‌خوان زبان انگلیسی متمرکز هستند. هدف این سازمان توسعهٔ تمامی شاخه‌های زبان‌شناسی پیکره‌ای در ارتباط با زبان انگلیسی است و در

همین راستا تولید پیکره‌ها و داده‌های زبان انگلیسی، گردآوری و توزیع داده‌های موجود و حمایت مالی و معنوی از پروژه‌های در دست انجام را در دستور کار خود قرار داده است. از سوی دیگر ایجاد زمینه‌های لازم جهت انجام تحقیقات زبان‌شناسی در خصوص منابع زبان انگلیسی و بهره‌گیری از آن‌ها در پژوهش‌های مرتبط با پردازش زبان طبیعی از دیگر فعالیت‌های این سازمان است. همچنین برگزاری کنفرانس‌های سالیانه، انتشار نشریهٔ علمی و راه‌اندازی تالارهای گفتگو در جهت ایجاد بستر مناسب برای تحقیق، بحث و بررسی زبان‌شناسی پیکره‌ای با محوریت زبان انگلیسی از دیگر اقدامات این سازمان به شمار می‌رود. نهایتاً هدف غایی و نهایی این سازمان ایجاد آرشیوی غنی از پیکره‌ها و داده‌های زبان انگلیسی و عرضهٔ آن‌ها به مؤسسات تحقیقاتی و پژوهشگران علاقه‌مندان این حوزهٔ پژوهشی است. اطلاعات بیشتر در http://icame.uib.no.

آرشیو متنی آکسفورد **Oxford Text Archive (OTA)**

آرشیوی از حدود ۲۵۰۰ متن الکترونیکی به ۲۵ زبان مختلف است که برخی از آن‌ها حاشیه‌نویسی شده‌اند. تولید، ذخیره‌سازی و انتشار متون دیجیتالی این مجموعه و دیگر منابع زبانیِ متعلق به آن، با هدف فراهم‌آوری منابع موردنیاز برای پژوهش‌های زبان‌شناختی و ادبی صورت گرفته است. اطلاعات بیشتر در https://ota.bodleian.ox.ac.uk/repository/xmlui

آرشیو متون الکترونیکِ الکس **Alex Catalogue of Electronic Texts**

آرشیوی از متون الکترونیکی در دسترس است که حق کپی‌رایت آزاد دارند. متون این مجموعه شامل آثار ادبی انگلیسی بریتانیایی، آمریکایی و برخی متون فلسفی است. تعداد اسناد این آرشیو در حال حاضر ۶۰۰ متن را شامل می‌شود.

آزمون ناپارامتریک **non-parametric test**

اگر توزیع جامعهٔ آماری، نامشخص و یا حجم نمونه، کوچک باشد از آزمون آماری ناپارامتریک استفاده می‌شود. آمار پارامتریک نسبت به آمار ناپارامتریک از قدرت بیشتری برخوردار است اما در شرایط مورداشاره لازم است که به آزمون ناپارامتریک

اتکا شود. در زبان‌شناسی پیکره‌ای به دلایل مختلف ممکن است داده‌های موجود در پیکره از حجم قابل قبولی برخوردار نباشند. در این حالت برای تحلیل داده‌ها استفاده از آزمون ناپارامتریک کارآمدتر است.

آژانس ارزیابی و انتشار منابع زبانی (اِلدا)
Evaluations and Language Resources Distribution Agency (ELDA)

آژانس ارزیابی و انتشار منابع زبانی (اِلدا) به‌عنوان واحد عملیاتی و شریک تجاری الرا در سال ۱۹۹۵ و هم‌زمان با تأسیس اِلرا، شروع به کار کرده است. رسالت اِلدا شناسایی، دسته‌بندی، جمع‌آوری، ارزیابی و تولید منابع زبانیِ موردنیاز فناوری‌های وابسته به زبان است. درواقع اِلدا به‌عنوان یک شرکت، وظیفه برآورده کردن اهداف اِلرا را بر عهده دارد و کلیه فعالیت‌های تجاری و مالی شرکت را انجام می‌دهد.

آموزش دستور رایانه‌یار computer aided grammar instruction

نرم‌افزارهای پیکره‌خوان می‌توانند در آموزش دستور زبان مفید باشند. در آموزش‌های رایانه‌یارِ دستور زبان، از نرم‌افزارهایی استفاده می‌شود که می‌توانند در پیکره‌های برچسب‌خورده کاوش کنند و نتایج موردنظر کاربر را معرفی کنند. این پیکره‌ها معمولاً دارای برچسب مقولهٔ دستوری کلمه و روابط نحوی هستند. نمونه‌ای از این نرم‌افزارها در زبان فارسی نرم‌افزار تحت وبی، به آدرس search.dadegan.ir است که امکان جستجو در پیکرهٔ وابستگی نحوی زبان فارسی را فراهم آورده است. این پیکره با حجمی حدود نیم میلیون کلمه، حاوی اطلاعات مربوط به مقولهٔ دستوری کلمات و همچنین نقش دستوری واژه‌های موجود در جملات پیکره است.

آموزش زبان با کمک رایانه (کال)
Computer Assisted Language Learning (CALL)

به بهره‌گیری از رایانه و فناوری‌های مرتبط با آن در آموزش زبان، شامل یادگیری زبان با رایانهٔ شخصی، با گوشی‌های تلفن همراه، در محیط‌های مجازی، آموزش از راه دور و تحت وب، بهره‌گیری از پیکره‌های زبانی و فهرست‌های بافت‌نما در آموزش زبان و موارد دیگری مانند آن گفته می‌شود. تمرکز این رویکرد آموزشی به تربیت

زبان‌آموزانِ زبان‌پژوه است و مدرس زبان را به‌عنوان تسهیل‌گر (و نه رهبر و کنترل‌کننده کلاس) در کنار زبان‌آموز قرار می‌دهد. روش CAI و CALI نسبت به کال قدیمی‌تر و پیشتازتر هستند اما چون همچنان بر روی نقش محوری مدرس در کلاس تأیید دارند نسبت به کال با استقبال کمتری روبرو بوده‌اند. CAI و CALI بیشتر بر دستورالعمل‌ها تمرکز دارند و کال به خودِ آموزش و تهیه مواد آموزشی زبان‌آموزمحور توجه می‌کند. بهره‌گیری از رویکردهای نظری مختلف آموزش زبان، تهیه و تولید نرم‌افزارهای آموزشی و آموزش نرم‌افزارها به زبان‌آموزان مهم‌ترین زیربخش‌های فعالیت‌های مرتبط با کال است.

آهنگ گفتمانی **discourse prosody**

به هم‌وقوعی برخی از واژه‌ها در یک پیکرهٔ زبانی با دسته‌های واژگانی دیگر و یا القای برخی مفاهیم، تفکرات و ایده‌های درون‌متنی گفته می‌شود. این هم‌وقوعی حکایت از نظم مشخصی دارد که نشان می‌دهد اطلاعات و نگرش‌های پنهان درون متن‌ها با بررسی محتوای واژگانی آن متون قابل کشف و بررسی است. برای مثال واژه‌ای مانند «وقوع» با مجموعه واژه‌هایی چون «زلزله»، «حادثه»، «جرم»، «خطا»، «انفجار»، «بهمن» و مانند آن همراه می‌شود که همگی از اتفاقاتی غیرمنتظره و ناگوار خبر می‌دهند. در مقابل واژه‌ای مانند «تجلی» بیشتر با دسته واژه‌هایی هم‌وقوعی دارد که بار معنایی مثبت دارند، برای نمونه بررسی داده‌ها نشان می‌دهد واژه‌ها و عبارت‌هایی مانند «شهادت»، «آثار هنری»، «نوروز»، «شعر»، «روح انسانی» و مانند آن با واژهٔ تجلی هم‌وقوعی بالایی دارند.

ا

ابزار Tool

منظور از ابزار در زبان‌شناسی پیکره‌ای مجموعه نرم‌افزارهایی است که با اهداف مختلفِ ساخت، تجهیز و تحلیل پیکره‌های زبانی مورد بهره‌برداری قرار می‌گیرند. ابزارهای پیکره‌ای را می‌توان در چهار گروه دسته‌بندی کرد.

گروه اول ابزارهای پیکره‌ساز هستند. این ابزارها که خزشگر نامیده می‌شوند می‌توانند با خزش در صفحات وب، خبرگزاری‌های آنلاین، وبلاگ‌ها، محیط‌های مجازی گفتگو و مانند آن، داده‌های زبانی را گردآوری و ذخیره کنند.

گروه دوم ابزارهای برچسب‌گذار هستند و به‌عنوان واسط میان انسان و رایانه، برچسب‌های موردنظر کاربر انسانی را به زبان قابل‌فهم توسط ماشین تبدیل می‌کنند. نرم‌افزارهای برچسب‌زن به دو شکل خودکار و غیرخودکار عمل می‌کنند. برچسب‌گذارهای خودکار با روش‌های مختلفی چون یادگیری ماشینی، بهره‌گیری از قواعد زبانی و مانند آن، به‌صورت خودکار داده‌های زبانی را برچسب‌گذاری می‌کنند، اما برنامه‌های غیرخودکار صرفاً محیطی را فراهم می‌آورند که در آن با ارائهٔ فهرست برچسب‌ها و نمایش بخش‌های مختلف پیکره بتوان به‌صورت دستی برچسب‌های مناسب را به لایه‌های مختلف زبان اعطا کرد.

گروه سوم ابزارها، نرم‌افزارهای تحلیل محتوا هستند. این نرم‌افزارها که به نرم‌افزارهای واژه‌نما معروف‌اند با کاوش در داده‌های زبانی و ارائهٔ اطلاعات آماری از موجودیِ واژه‌ایِ پیکره‌ها به استخراج فهرست بسامدی واژه‌ها و کلیدواژه‌ها، فهرست همایندها، فهرست بافت‌نما و مانند آن می‌پردازند و به‌این‌ترتیب از یک‌سو در تحلیل

متن و گفتمان کارآمد هستند و از سوی دیگر در آموزش زبان و یا فرهنگ‌نویسی، فهرست‌های لازم جهت تدریس و یا مدخل‌گزینی و ارائهٔ شاهدمثال را به دست می‌دهند. این نرم‌افزارها در کاربرد اخیر با عنوان ابزارهای واژه‌پرداز شناخته می‌شوند.

گروه چهارم که می‌توان آن‌ها را با عنوان نرم‌افزار پیکره‌خوان معرفی کرد، صرفاً جهت نمایش اطلاعات درون پیکره‌ها و برچسب‌های زبانی ساخته می‌شوند. برخی پیکره‌های برچسب‌خورده، واسط یا رابط کاربری را هم در کنار پیکره عرضه می‌کنند تا زبان‌شناس بتواند بدون نیاز به رجوع به پایگاه داده و بی‌نیاز از کدنویسی‌های ضروری جهت استخراج اطلاعات موردنظرش با استفاده از یک ظاهر ساده، در پیکره کاوش کند و روابط زبانی مختلف را با توجه به برچسب‌های موجود در پیکره مشاهده کند. مانند واسطی که برای پیکرهٔ دادگان وابستگی نحوی زبان فارسی و به آدرس search.dadegan.ir تهیه شده است. لازم به ذکر است که برخی نرم‌افزارها نمایش بصری دارند و با ارائهٔ نتایج به شکل ابر واژگان و یا شبکه‌های معنایی، تصویر بهتری از نتایج بررسی به دست می‌دهند. به لحاظ بستر کاری هم می‌توان ابزارهای پیکره‌ای را به دو دسته تقسیم کرد، الف- آن‌هایی که تحت وب کار می‌کنند، ب- آن‌هایی که باید بارگیری شوند و تحت ویندوز، مک و سیستم‌عامل‌های دیگر، به‌صورت نصبی و یا اجرایی، مورد بهره‌برداری قرار گیرند. برای اطلاعات بیشتر در خصوص برخی ابزارهای موجودِ زبان‌شناسی پیکره‌ای می‌توان به آدرس https://www.corpus-analysis.com مراجعه کرد.

ابزار تحلیل صوت **Speech Analysis Tool**

به مجموعه نرم‌افزارهایی اشاره می‌کند که امکان مشاهدهٔ موج صوتی و درج برچسب‌های لازم بر روی بخش‌های مختلف آن را فراهم می‌آورد.

ابهام **ambiguity**

ابهام و ایهام در لغت به مفاهیمی اشاره می‌کنند که بیش از یک تعبیر و تفسیر دارند. تفاوت این دو اصطلاح در این است که با وجود بافت می‌توان از ابهام، رفع ابهام کرد اما در ارتباط با ایهام، بافت، کمکی به شفافیت تعبیر نمی‌کند.

ابهام در زبان‌شناسی پیکره‌ای اصطلاح عامی است که در فرایند برچسب‌گذاری مطرح می‌شود. این وضعیت به شرایطی اشاره می‌کند که کاربر انسانی نمی‌تواند برای

برخی واحدهای زبانی میان چند برچسب به تصمیم واحدی برسد. اولین راهکار در این وضعیت توجه به بافت است. اگر با وجود بافت همچنان ابهام بر سر جایش باقی باشد، با توجه به رویکرد پیکره‌سازانِ پیکرهٔ موردنظر یا می‌توان از برچسبی با نام «**برچسب چندوجهی**» استفاده کرد و یا قرارداد یا قراردادهایی را برای پوشش دادن موارد ابهام در طول کار وضع کرد. بدیهی است که در هر حالت در هنگام انتشار پیکره، در خصوص قراردادها و تصمیم‌ها اطلاع‌رسانی صورت می‌گیرد.

ابهام‌زدایی **disambiguation**

مجموعه فعالیت‌هایی است که به دنبال آن، از زبان رفع ابهام می‌شود. برای این منظور دو روش وجود دارد. اول آنکه پیکره‌های زبانی توسط کاربران آگاه زبان، برچسب‌گذاری شوند و در خصوص موارد اختلاف و ابهام، توسط کاربر انسانی تصمیم‌گیری شود و پس از آن بر اساس پیکرهٔ برچسب‌خورده، یادگیری ماشینی صورت بگیرد و ماشین بتواند با توجه به ویژگی‌های مختلف، در برخورد با جملهٔ جدید، از موارد ابهام‌زا، ابهام‌زدایی کند. برچسب‌گذاری در این سطح، انواع متنوعی دارد که برای نمونه می‌توان به دو مورد آن اشاره کرد. در ابهام‌زدایی از نقش‌های دستوری، از برچسب مقولهٔ سخن استفاده می‌شود تا مثلاً بتوان میان «جوان» به‌عنوان صفت و «جوان» به‌عنوان اسم، تمایز قائل شد. در ابهام‌زدایی معنایی و رفع ابهام از هم‌نگاره‌ها، تلاش می‌شود از ابهام‌های حاصل از چندمعناها، هم‌نویسه‌ها و همنام‌ها رفع ابهام شود و برای مثال میان «ایران» به‌عنوان اسم خانم، «ایران» به‌عنوان یک سرزمین و «ایران» به‌عنوان یک حکومت تمایز گذاشته شود. ابهام‌زدایی در هر سطحی که باشد با توجه به بافت صورت می‌گیرد.

روش‌های دیگر به پیکره‌های برچسب‌خورده متکی نیستند و این فرایند را یا به‌صورت قاعده‌بنیاد یا با توجه به روابط دودویی یا بیشتر و یا روش‌های دیگری مانند آن صورت می‌دهند.

ابهام‌زدایی احتمالاتی **probabilistic disambiguation**

یکی از روش‌های برچسب‌گذاری خودکار پیکره‌های زبانی آن است که ماشین براساس پیکره‌های برچسب‌خوردهٔ موجود، به برچسب‌گذاری داده‌های جدید بپردازد. برای این منظور روش‌های مختلفی وجود دارد. در یکی از این روش‌ها