واژه‌نامه زبان‌شناسی پیکره‌ای

پاول بیکر، اندرو هاردی، و تونی مک‌انری

مترجمان:
سعیده قندی
محمدحسن ترابی

با مقدمه و نظارت علمی دکتر سیدمصطفی عاصی

سازمان اسناد و کتابخانه ملی
جمهوری اسلامی ایران

سرشناسه	پاول بیکر، اندرو هاردی، و تونی مک‌انری **Paul Baker, Andrew Hardie and Tony McEnery**
عنوان و نام پدیدآور	واژه‌نامه زبان‌شناسی پیکره‌ای / نویسندگان پاول بیکر، اندرو هاردی، و تونی مک‌انری ؛ مترجمان: سعیده قندی، محمدحسن ترابی؛ ویراستار پریسا بخشنده
مشخصات نشر	تهران : نشر نویسه پارسی، ۱۳۹۹
مشخصات ظاهری	۳۲۷ صفحه
شابک	۹۷۸-۶۲۲-۶۶۴۹-۴۳-۸
وضعیت فهرست نویسی	فیپا
یادداشت	عنوان اصلی کتاب: **A Glossary of Corpus Linguistics**
موضوع	زبان‌شناسی پیکره‌ای -- واژه‌نامه‌ها -- فارسی **Corpus Linguistics-- Dictionaries -- Persian**
موضوع	فارسی -- واژه‌نامه‌ها ـ انگلیسی **Persian language -- Dictionaries -- English**
موضوع	زبان انگلیسی -- واژه‌نامه‌ها ـ فارسی **English language -- Dictionaries -- Persian**
شناسه افزوده	قندی، سعیده، ۱۳۶۶ - ، مترجم
شناسه افزوده	ترابی، محمدحسن، ۱۳۶۶ - ، مترجم
شناسه افزوده	بخشنده، پریسا، ۱۳۶۵ - ، ویراستار
رده بندی کنگره	**QP۲۱۶**
رده بندی دیویی	۸۲۳۶/۶۱۲
شماره کتابشناسی ملی	۶۱۱۹۱۴۲

این کتاب ترجمه‌ای است از:

A Glossary of Corpus Linguistics

Paul Baker, Andrew Hardie, and Tony McEnery

© Paul Baker, Andrew Hardie and Tony McEnery, 2006

نویسندگان: پاول بیکر، اندرو هاردی، و تونی مک‌انری

مترجمان: سعیده قندی، محمدحسن ترابی

ویراستار: پریسا بخشنده

آتلیه نشرنویسه پارسی: STUDIO FIVE

ناشر: نشر نویسه پارسی

دفتر انتشارات: ۰۲۱-۷۷۰۵۳۲۴۶

نماینده فروش: کتابفروشی توس: ۶۶۴۶۱۰۰۷

سامانه پیام کوتاه: ۳۰۰۰۴۵۵۴۵۵۴۱۴۲

وبگاه: www.neveeseh.com

نوبت چاپ: اول، ۱۳۹۹

شمارگان: ۳۰۰ نسخه

شابک: ۹۷۸-۶۲۲-۶۶۴۹-۴۳-۸

چاپ و صحافی: روز

کلیه حقوق محفوظ و متعلق به «نشرنویسه پارسی» است.
تکثیر و انتشار این اثر یا قسمتی از آن به هر شیوه، بدون مجوز قبلی و کتبی ممنوع و مورد پیگیری قانونی قرار خواهد گرفت.

فرهنگ‌نامه‌ها	شماره مسلسل انتشارات
۷	۱۰۸

فهرست مطالب

مقدمه	۷
پیشگفتار مترجمان	۹
یادداشت‌های مقدماتی	۱۳
آدرس وبگاه‌ها	۱۳
فهرست اختصارات	۱۳
اصطلاحات	۲۳
منابع	۲۵۷
واژه‌نامه فارسی-انگلیسی	۲۹۳

مقدمه

زبان پدیده‌ای پویا و همواره درحال دگرگونی است و بیشترین تغییرات در واژه‌ها و معنی آن‌ها نمود می‌یابد. عوامل گوناگونی در این دگرگونی‌ها دخالت دارند اما شاید مهم‌ترین آن‌ها را بتوان عامل زمان دانست. هرچه از گذشته به دوران کنونی نزدیک می‌شویم، به‌دلایل مختلف ازجمله افزایش ارتباطات و کاربرد رسانه‌های گوناگون و توسعه فراگیر دانش و فن، سرعت دگرگونی‌ها بیشتر می‌شود و با پیدایش و گسترش برق‌آسای فناوری اطلاعات این سرعت به اوج می‌رسد. در حوزه علم، بیان و توصیف دقیق مفاهیم از مهم‌ترین بایستگی‌هاست که آن هم به کاربرد مجموعه‌ای از واژه‌ها و اصطلاحاتی[1] وابسته است که با دقت از سوی دانشوران و کارشناسان تعیین و تعریف می‌شوند. با نگاهی به واژگان تخصصی زبان‌شناسی درمی‌یابیم که بسیاری از واژه‌های این حوزه با معانی و تعریف‌های تازه‌ای در متن‌ها کاربرد یافته‌اند. به‌ویژه در شاخه تازه‌تأسیس و روبه‌رشدی مانند زبان‌شناسی پیکره‌ای که با مفاهیم و فرایندهای نوینی سروکار دارد.

تنها کافی است واژه **پیکر** (یا در مواردی خاص، پیکره) را در زبان فارسی درنظر بگیریم که تا حدود پنج-شش دهه پیش (زمان انتشار فرهنگ معین) به‌معنی تصویر، نقش و یا مجسمه و تندیس کاربرد داشت و با آغاز به‌کارِ رشته دانشگاهیِ زبان‌شناسی و به‌میان آمدن روش‌های تازه بررسی زبان با بهره‌گیری از مواد گردآوری شده، اصطلاح **پیکره**[2] یا **پیکره زبانی**[3] برای اشاره به هرگونه داده زبانی با هر حجم و شیوه گردآوری (یادداشت‌برداری، برگه‌نویسی دستی یا ضبط آوایی) بازتعریف شد. با این تعریف، پیکره می‌توانست دربرگیرنده چند جمله، فهرستی از چند واژه یا عبارت، شماری برگه یادداشت‌شده از منابعی پراکنده یا صورت نوشتاری چند ساعت مطلب ضبط‌شده باشد. این پیکره‌ها اغلب کاربردی موردی، شخصی، نادقیق و محدود داشتند و پس از بررسی به کناری نهاده می‌شدند و به‌اصطلاح یکبارمصرف بودند.

---

[1] terminology
[2] corpus
[3] linguistic corpus

همین محدودیت‌ها و نارسایی‌ها احساس نیازی فزاینده به پیکره‌هایی بزرگتر، دقیق‌تر و با کارایی بیشتر را در میان پژوهندگان زبان ایجاد کرده بود.

نقطه عطف هنگامی فرارسید که با به‌کارگیری امکانات رایانه‌ای و فناوریِ اطلاعات، درونداد، ذخیره‌سازی، پردازش، جستجو و بازیابی داده‌های زبانی در حجم‌های عظیم، با سرعتی شگفت‌انگیز و دقتی فراتر از انتظار ممکن گشت. امروز واژه **پیکره** با تعریفی تازه، بار معنایی بسیار متفاوت و گسترده‌ای یافته است و به‌عنوان اصطلاحی فنی و تخصصی در زبان‌شناسی شناخته می‌شود. شاید برای اولین‌بار با این معنی در نخستین کنفرانس زبان‌شناسی (عاصی، ۱۳۶۹) معرفی شد. از همین‌جا نام حوزه‌ای جدید یعنی میان‌رشته **زبان‌شناسیِ پیکره‌ای**[1] نیز ساخته‌شد (عاصی، ۱۳۷۲).

اکنون زبان‌شناسی پیکره‌ای به‌اندازه‌ای گسترش یافته که مجموعه واژه‌ها و اصطلاحات فنی آن واژگانی تخصصی را پدید آورده است و بخشی بنیادی از این شاخه علمی، سودمند و پرکاربرد به‌شمار می‌آید. فرهنگ پیشِ رو نخستین گام در معرفی واژگان زبان‌شناسی پیکره‌ای و ارائه برابرهای فارسی آن است. مترجمان برابرهای مناسبی برای بسیاری از اصطلاح‌ها انتخاب کرده‌اند، با این‌حال هنوز در عرصه واژه‌گزینیِ این حوزه تلاش بیشتری باید کرد.

مصطفی عاصی
بهمن ماه ۱۳۹۸

عاصی، مصطفی. ([۱۳۶۹]، ۱۳۷۲). کاربرد کامپیوتر در زبان‌شناسی و فرهنگ‌نگاری. در دبیرمقدم، محمد (گردآورنده)، مجموعه مقالات نخستین کنفرانس زبان‌شناسی نظری و کاربردی (صص. ۱۶۳–۱۷۸). تهران، دانشگاه علامه طباطبایی.

[1] corpus linguistics

پیشگفتار مترجمان

کتابی که در دست دارید، فرهنگ توصیفی زبان‌شناسی پیکره‌ای، ترجمه‌ای است از کتابی نسبتاً قدیمی (نسخه انگلیسی کتاب در سال ۲۰۰۶ منتشر شده است)، اما فعلاً، البته تا جایی که ما اطلاع داریم، تنها فرهنگ توصیفی زبان‌شناسی پیکره‌ای در دسترس است و ما کتاب جدیدتری نیافتیم. ما در اولین تلاش برای انتشار چنین کتابی دست به ترجمه آن زدیم تا راهی باز شود برای تألیف کتابی جدیدتر با تأکید بر پیکره‌ها و ابزارهای در دسترس برای زبان فارسی. اما به هر حال کتاب حاضر هنوز منبع ارزشمندی است که می‌تواند به نیازهای بسیاری از دانشجویان و فعالان حوزه زبان‌شناسی پیکره‌ای و حوزه‌های وابسته به آن پاسخ دهد. کتاب برای خود ما بسیار آموزنده بود و مطمئنیم برای خوانندگان هم آموزنده خواهد بود.

در متن کتاب پیوندهایی به وبگاه‌های مختلف بود، که همان طور که در یادداشت‌های مقدماتی خود کتاب آمده است، ممکن بود آن صفحه‌های اینترنتی دیگر وجود نداشته باشند. ما پیوندها را بار دیگر کنترل کرده و جاهایی که آدرس پیوندها تغییر کرده‌اند، پیوندهای جدید (یعنی تا سال ۲۰۲۰/۱۳۹۸) را جایگزین کردیم. چند مورد از وبگاه‌ها و یا ابزارها کاملاً از دسترس خارج شده بودند، در این صورت پیوند را تغییر ندادیم، فقط در پانویس توضیح دادیم که این صفحه دیگر در دسترس نیست. البته برای اطمینان از اینکه صفحه‌ای را از دست نداده باشیم، آن موضوع خاص را در سه جویشگر www.google.com، www.bing.com، و www.ecosia.org جستجو کردیم و اکنون تقریباً مطمئن هستیم که آن چند وبگاه یا ابزار دیگر وجود خارجی ندارند. اما ما نیز، مانند نویسنده‌های اصلی کتاب، پیشنهاد می‌کنیم که هر وقت به یکی از پیوندها مراجعه کردید و آن صفحه را نیافتید، دوباره آن را جستجو کنید، شاید آن صفحه مورد نظر را در آدرسی جدید یافتید. دو نکته دیگر را هم باید اضافه کنیم، یکی اینکه در متن برای ارجاع‌های متقابل بین مدخل‌ها از حروف **پررنگ** استفاده کرده‌ایم و دوم اینکه برای سهولت دسترسی به اصطلاحات، در پایان کتاب یک واژه‌نامه فارسی–انگلیسی هم قرار داده‌ایم.

مطالعه و استفاده از کتاب واژه‌نامه زبان‌شناسی پیکره‌ای را به همه علاقه‌مندان توصیه می‌کنیم، به‌خصوص به فعالان حوزه زبان‌شناسی پیکره‌ای، زبان‌شناسی رایانشی، فرهنگ‌نگاری، واژه‌سازی، آموزش زبان فارسی به فارسی‌زبانان، متخصصین آموزش زبان‌های خارجی، مطالعات ترجمه، و البته کسانی که در حوزه مطالعات زبان اول، روان‌شناسی زبان و زبان کودک پژوهش می‌کنند.

برای انتخاب برابرنهادهای اصطلاحات، از چند کتاب استفاده کردیم که هم به رسم ادب و هم برای آشنایی خوانندگان، فهرست آن‌ها را در اینجا می‌آوریم:

خسروی‌زاده، پروانه، شیخ‌زاده، مارال، مرادی، مهدی، و مواجی، وحید. (۱۳۹۱). فرهنگ توصیفی زبان‌شناسیِ رایانشی. تهران: پردیس دانش.

قطره، فریبا. (۱۳۹۶). واژه‌نامه توصیفی فرهنگ‌نویسی. تهران: نویسه پارسی.

گروه مؤلفان. (۱۳۹۷). هزار واژه زبان‌شناسی ۱. تهران: فرهنگستان زبان و ادب فارسی.

مندیک، پیت. (۲۰۱۰). کلیدواژه‌های فلسفه ذهن. ترجمه محمدحسن ترابی، ۱۳۹۵. تهران: نویسه پارسی.

میرزائی، آزاده. (۱۳۹۶). آشنایی با زبان‌شناسی پیکره‌ای. تهران: دانشگاه علامه طباطبائی.

همایون، همادخت. (۱۳۷۹). واژه‌نامه زبان‌شناسی و علوم وابسته. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

در مواردی هم که در سایر منابع برابرنهادی نیافتیم، از استاد گرامی‌مان، دکتر سیدمصطفی عاصی - که نام ایشان برای همه فعالان حوزه زبان‌شناسی پیکره‌ای آشناست - کمک گرفتیم و یا اگر اصطلاحی را خودمان وضع کردیم، نظر ایشان را نیز جویا شدیم. از ایشان به دلیل کمک‌هایشان برای انتخاب برابرنهادها، مقدمه ارزشمندی که بر کتاب نوشتند، و نیز به این دلیل که وقت ارزشمند خود را برای نظارت علمی بر ترجمه این کتاب در اختیار ما قرار دادند صمیمانه سپاسگزاریم. متشکریم از آقای مصطفی کبیری، که اگر کمک‌های ایشان نبود، کتاب حاضر ترجمه نمی‌شد. همچنین قدردانیم از خانم پریسا بخشنده، ویراستار کتاب، که با ریزبینی و

دقت‌نظر مثال‌زدنی خود، ترجمه فارسی این کتاب را خواندنی‌تر و خوانش‌پذیرتر کردند. و در نهایت، بی‌نهایت سپاسگزاریم از آقای امیر احمدی، مدیر نشر نویسه پارسی که با صبر و تلاش ستودنی ایشان، کتاب منتشر شد و در اختیار خوانندگان قرار گرفت.

امیدواریم که ترجمه خوب و قابل‌قبولی در اختیار علاقه‌مندان قرار داده باشیم؛ می‌دانیم که این ترجمه بدون نقص نیست و حتماً کاستی‌هایی دارد و به عنوان مترجمان اثر، مسئولیت همه این کاستی‌ها را می‌پذیریم و قدردان نظرات و انتقادات همه مخاطبان گرامی‌مان خواهیم بود.

مترجمان

اردیبهشت ۱۳۹۹

یادداشت‌های مقدماتی

آدرس وبگاه‌ها

تلاش کرده‌ایم که تا جای ممکن از ارجاع دادن به وبگاه‌ها خودداری کنیم، زیرا دریافتیم که برخی از وبگاه‌هایی که در ابتدای نوشتن کتاب در متن کتاب وارد کردیم، در زمان رسیدن به مراحل پایانی دیگر وجود نداشتند. ما وبگاه‌های برخی از سازمان‌ها، گروه‌ها، پیکره‌ها یا نرم‌افزارها را در صورتی در متن می‌آوردیم که احتمال می‌دادیم بسته نمی‌شوند. اما نمی‌توانیم ماندگاری همه وبگاه‌هایی که در اینجا آورده‌ایم را تضمین کنیم. اگر خوانندگان خواستند که اطلاعات خاصی را در اینترنت دنبال کنند، اما با پیوندی بسته مواجه شدند، عذرخواهی ما را بپذیرند و سپس از یک جویشگر معتبر مانند www.google.com کمک بگیرند (اگر گوگل هنوز وجود داشته باشد!).

فهرست اختصارات

زبان‌شناسی پیکره‌ای رشته‌ای است که اختصارات زیادی دارد. این ازنظر وحدت رویه مشکلاتی را به وجود می‌آورد: برخی اصطلاحات با صورت اختصاری‌شان و برخی با نام کاملشان بهتر شناخته می‌شوند. ما می‌خواهیم که در ترتیب‌بندی مدخل‌های فرهنگ وحدت رویه داشته باشیم، اما می‌خواهیم یافتن مدخل‌ها آسان هم باشد؛ بنابراین، تصمیم گرفتیم در ترتیب‌بندی مدخل‌های فرهنگ، صورت کامل همه اختصارات را بنویسیم، اما فهرستی از همه اختصارات را در ابتدای فرهنگ همراه با نام کاملشان بیاوریم. بنابراین، خوانندگانی که می‌خواهند با استفاده از این فرهنگ اطلاعاتی درباره BNC به دست آورند، می‌توانند عنوان کامل آن را در فهرست اختصارات در ابتدای کتاب بیابند، و بعد به مدخل British National Corpus در فرهنگ مراجعه کنند.

صورت اختصاری	مدخل انگلیسی	برابرنهاد فارسی
ACASD	Automatic Content Analysis of Spoken Discourse word sense tagging system	سیستم برچسب‌دهی معنای واژه برای تحلیل محتوایی خودکار متون گفتاری
ACE	Australian Corpus of English	پیکره انگلیسی استرالیا
ACH	Association for Computers and the Humanities	انجمن رایانه و علوم انسانی
ACL	Association for Computational Linguistics	انجمن زبان‌شناسی رایانشی
ACLDCI	Association for Computational Linguistics Data Collection Initiative	کارگروه گردآوری داده انجمن زبان‌شناسی رایانشی
AGTK	Annotation Graph Toolkit	انوتیشن گرف تولکیت
AHI	American Heritage Intermediate Corpus	پیکره میانی امریکن هریتیج
ALLC	Association for Literary and Linguistic Computing	انجمن رایانش ادبی و زبانی
AMALGAM	Automatic Mapping Among Lexico-Grammatical Annotation Models Tagger	برچسب‌زن نگاشت خودکار در انگاره‌های نشانه‌گذاری واژگانی-دستوری
ANC	American National Corpus	پیکره ملی آمریکا
ANLT	Alvey Natural Language Tools	ابزارهای زبان طبیعی آلوِی
AP	Associated Press Treebank	درخت‌بانک آسوشیتدپرس
APHB	American Printing House for the Blind Treebank	چاپ‌خانه آمریکا برای درخت‌بانک نابینایان
ARCHER	Representative Corpus of Historical English Registers Corpus	پیکره نماینده سیاق‌های تاریخی انگلیسی (پیکره آرچر)
ASCII	American Standard Code for Information Exchange	کد استاندارد آمریکا برای تبادل اطلاعات
ATC	Air Traffic Control Corpus	پیکره کنترل ترافیک هوایی

صورت اختصاری	مدخل انگلیسی	برابرنهاد فارسی
AUTASYS	Automatic Text Annotation System Tagger	برچسب‌زن سیستم نشانه‌گذاری متنی خودکار
BAS	Bavarian Archive for Speech Signals	آرشیو باواریایی برای سیگنال‌های گفتاری
BASE	British Academic Spoken English Corpus	پیکره انگلیسی گفتاری دانشگاهی بریتانیایی
BNC	British National Corpus	پیکره ملی بریتانیا
BoE	Bank of English	بانک زبان انگلیسی
CALL	Computer Assisted Language Learning	زبان‌آموزی رایانه‌یار
CAMET	Computer Archive of Modern English Texts	آرشیو رایانه‌ای از متون انگلیسی جدید
CANCODE	Cambridge and Nottingham Corpus of Discourse in English	پیکره گفتاری کمبریج و ناتینگهام از زبان انگلیسی
CEEC	Corpus of Early English Correspondence	پیکره مکاتبات انگلیسی متقدم
CEG	Cronfa Electroneg o Gymraeg	پیکره گنجینه الکترونیک زبان ولزی
CELEX	Centre for Lexical Information Relational Database	دادگان رابطه‌ای مرکز اطلاعات واژگانی
CES	Corpus Encoding Standard	استاندارد کدگذاری پیکره
CETH	Centre for Electronic Texts in the Humanities	مرکز متون الکترونیکی حوزه علوم انسانی
CHAT	Codes for the Human Analysis of Transcripts System	سیستم کد برای تحلیل انسانی ترانوشت‌ها
CHILDES	Child Language Data Exchange System	سیستم تبادل داده‌های زبانی کودک
CIDE	Collaborative International Dictionary of English	فرهنگ همکارانه بین‌المللی برای زبان انگلیسی

برابرنهاد فارسی	مدخل انگلیسی	صورت اختصاری
سیستم تحلیل زبان رایانه‌ای شده	Computerized Language Analysis System	CLAN
سیستم برچسب دهی واژگانی خودکار احتمال سازه	Constituent Likelihood Automatic Word-tagging System	CLAWS
پیکره زبان‌آموز انگلیسی از گویشوران چینی	Chinese Learner English Corpus	CLEC
کنسرسیوم پژوهش‌های واژگانی	Consortium for Lexical Research	CLR
جعبه‌ابزار مدل‌سازی زبانی آماری دانشگاه کارنجی ملون-کمبریج	Carnegie Mellon University–Cambridge Statistical Language Modeling Toolkit	CMU SLM
پیکره پاره‌گفتارهای زبان طبیعی همکارانه و همپایه	Cooperative, Coordinated Natural Language Utterances Corpus	Coconut
پیکره برگن از انگلیسی نوجوانان لندن	Bergen Corpus of London Teenage English	COLT
پیکره استخراج منابع پیکره‌ای و اصطلاحات	Corpus Resources and Terminology Extraction	CRATER
پیکره انگلیسی آمریکایی گفتاری	Corpus of Spoken American English	CSAE
مرکز پیکره‌های گفتاری ایجادشده به‌منظور درک زبان گفتاری	Centre for Spoken Language Understanding Speech Corpora	CSLU
مرکز تحقیقات فناوری گفتار	Centre for Speech Technology Research	CSTR
پیکره نیم‌زبان بریتانیایی نوشتاری	Corpus of Written British Creole	CWBC
ابزار نشانه‌گذاری گفتگو	Dialogue Annotation Tool	DAT

برابرنهاد فارسی	مدخل انگلیسی	صورت اختصاری
پیکره درزمانی از انگلیسی گفتاری امروز	Diachronic Corpus of Present-day Spoken English	DCPSE
تعریف سندی	document type definition	DTD
شعبه اروپایی انجمن زبان‌شناسی پیکره‌ای	European Chapter of the Association for Computational Linguistics	EACL
گروه مشاورین نخبه برای استانداردهای مهندسی زبان	Expert Advisory Group on Language Engineering Standards	EAGLES
کارگروه پیکره اروپایی	European Corpus Initiative	ECI
برچسب‌زن زبانی یودیکو	Eudico Linguistic Annotator	ELAN
شبکه فعالیت زبانی اروپا	European Language Activity Network	ELAN
آژانس ارزیابی‌ها و توزیع منابع زبانی	Evaluations and Language Resources Distribution Agency	ELDA
انجمن منابع زبانی اروپا	European Language Resources Association	ELRA
قطب علمی اروپایی فناوری های زبان انسان	European Network of Excellence in Human Language Technologies	ELSNET
پیکره فعال‌سازی مهندسی زبان اقلیت	Enabling Minority Language Engineering Corpus	EMILLE
تجزیه‌گر دستور محدودیت برای زبان انگلیسی	Constraint Grammar Parser of English	ENGCG
داده‌بانک زبان دوم بنیاد علوم اروپا	European Science Foundation Second Language Databank	ESFSLD
پیکره انگلیسی بریتانیایی فرایبورگ-لوب (FLOB)	Freiburg-LOB Corpus of British English	FLOB
دادگان زبان میانی فرانسوی	French Interlanguage Database	FRIDA
پیکره فرایبورگ-براون از انگلیسی آمریکایی	Freiburg–Brown Corpus of American English	FROWN

برابرنهاد فارسی	مدخل انگلیسی	صورت اختصاری
تکه‌های درخت نامعین	Fuzzy Tree Fragments	FTF
معماری کلی برای مهندسی متن	General Architecture for Text Engineering	GATE
پیکره نفت گوانگژو	Guangzhou Petroleum English Corpus	GPEC
مرکز پژوهش‌های ارتباطات انسانی	Human Communication Research Centre	HCRC
پیکره دانشگاه علم و فناوری هنگ‌کنگ	Hong Kong University Of Science And Technology Corpus	HKUST
فناوری زبان انسان	human language technology	HLT
زبان نشانه‌گذاری ابرمتن	Hypertext Markup Language	HTML
آرشیو رایانه‌ای بین‌المللی از انگلیسی میانه و معاصر	International Computer Archive of Modern and Medieval English	ICAME
پیکره بین‌المللی انگلیسی	International Corpus of English	ICE
پیکره بین‌المللی از برنامه به‌کارگیری پیکره انگلیسی	International Corpus of English Corpus Utility Program	ICECUP
پیکره بین‌المللی انگلیسی آموز	International Corpus of Learner English	ICLE
میزکار پیکره‌ای موسسه پردازش ماشینی زبان	Institut für Maschinelle Sprachverarbeitung	IMS
پیکره تعاملی زبان گفتاری برای آموزش	Interactive Spoken Language Education Corpus	ISLE
پیکره تغییرات آهنگی در زبان انگلیسی	Intonational Variation in English Corpus	IviE
کلیدواژه در متن	key word in context	KWIC
پیکره لنکستر از ماندارین چینی	Lancaster Corpus of Mandarin Chinese	LCMC

صورت اختصاری	مدخل انگلیسی	برابرنهاد فارسی
LCPW	Lancaster Corpus of Children's Project Writing	پیکره پروژه‌نویسی کودکان لنکستر
LDB	Linguistic DataBase	دادگان زبانی
LDC	Linguistic Data Consortium	کنسرسیوم داده‌های زبانی
LeaP	Learning the Prosody of a Foreign Language Corpus	پیکره فراگیری ویژگی‌های نوایی زبان خارجی
Lindsei	Louvain International Database of Spoken English Interlanguage	دادگان بین‌المللی از زبان انگلیسی میانه گفتاری لووین
LLC	London–Lund Corpus	پیکره لندن–لوند
LOB	Lancaster–Oslo/Bergen Corpus	پیکره لنکستر–اوسلو/برگن
MARSEC	Machine-Readable Spoken English Corpus	پیکره انگلیسی گفتاری رایانه‌خوانا
MBT	Memory Based Tagger	برچسب‌زن حافظه‌بنیاد
METER	Measuring Text Reuse Corpus	پیکره اندازه‌گیری تکرار متن
MICASE	Michigan Corpus of Academic Spoken English	پیکره میشیگان از انگلیسی
MTP	Münster Tagging Project	پروژه برچسب‌دهی مونستر
MXPOST	Maximum Entropy Part-of-Speech Tagger	برچسب‌زن اجزای کلام آنتروپی بیشینه
NECTE	Newcastle Electronic Corpus of Tyneside English	پیکره الکترونیکی انگلیسی تاینساید نیوکاسل
NEET	Network of Early Eighteenth Century English Texts	شبکه متون انگلیسی اوایل قرن هجدهم
NITCS	Northern Ireland Transcribed Corpus of Speech	پیکره ترانویسی‌شده گفتاری ایرلند شمالی
NLP	natural language processing	پردازش زبان طبیعی
OCP	Oxford Concordance Programme	برنامه واژه‌نمای آکسفورد
OCR	optical character recognition	بازشناسی نوری کارکتر

صورت اختصاری	مدخل انگلیسی	برابرنهاد فارسی
OLAC	Open Language Archives Community	جمعیت فعال در آرشیو زبان باز
OTA	Oxford Text Archive	آرشیو متن آکسفورد
POS	part-of-speech tagging	برچسب‌دهی اجزای کلام
POW	Polytechnic of Wales corpus	پیکره پلی‌تکنیک ولز
SARA	SGML-Aware Retrieval Application	اپلیکیشن بازیابی SGML آگاه
ScoSE	Saarbrücken Corpus of Spoken English	پیکره انگلیسی گفتاری ساربروکن
SEC	Lancaster/IBM Spoken English Corpus	پیکره انگلیسی گفتاری لنکستر/IBM
SEU	Survey of English Usage Corpus	پیکره بررسی کاربرد زبان انگلیسی
SGML	Standard Generalised Markup Language	زبان نشانه‌گذاری عمومی استاندارد
SPAAC	Speech Act Annotated Corpus for Dialogue Systems	پیکره نشانه‌گذاری‌شده کارگفت برای سیستم‌های گفتگو
SUSANNE	Surface and Underlying Structural Analyses of Naturalistic English Corpus	پیکره تحلیل ساختاری روساختی و ژرف‌ساختی انگلیسی طبیعت‌گرایانه
TEI	Text Encoding Initiative	کارگروه نشانه‌گذاری متن
TELC	Thai English Learner Corpus	پیکره زبان‌آموز تایلندی انگلیسی
TESS	Text Segmentation for Speech Project	پروژه پاربندی متن برای گفتار
TLFi	Trésor de la Langue Française Informatisé	گنجینه رایانشی زبان فرانسه
TLG	Thesaurus Linguae Graecae	گنجینه زبان یونانی

صورت اختصاری	مدخل انگلیسی	برابرنهاد فارسی
TnT	Trigrams‘n’Tags	برچسب‌زن ترایگرام ن تگز
TOSCA	Tools for Syntactic Corpus Analysis Corpus	پیکره ابزارهای پیکره تحلیل نحوی (توسکا)
T2K-SWAL	TOEFL 2000 Spoken and Written Academic Language Corpus	پیکره زبان گفتاری و نوشتاری دانشگاهی تافل ۲۰۰۰
UAM	Universidad Autónoma de Madrid Spanish Treebank	درخت‌بانک زبان اسپانیایی دانشگاه آزاد مادرید
UCREL	University Centre for Computer Corpus Research on Language	مرکز دانشگاهی پژوهش‌های پیکره‌ای رایانه‌ای زبان
USAS	UCREL Semantic Analysis System	سیستم تحلیل معنایی UCREL
WBE	Wolverhampton Business English Corpus	پیکره انگلیسی تجاری وُلورهمپتون
WSC	Wellington Corpus of Spoken New Zealand English	پیکره ولینگتون از انگلیسی گفتاری نیوزلندی
WWC	Wellington Corpus of Written New Zealand English	پیکره ولینگتون از انگلیسی نوشتاری نیوزلندی
Xaira	XML Aware Indexing and Retrieval Architecture	معماری نمایه‌سازی و بازیابی XML آگاه
XML	Extensible Markup Language	زبان نشانه‌گذاری گسترش یابنده
YCOE	York–Toronto–Helsinki Corpus of Old English Prose	پیکره نثر انگلیسی باستان یورک-تورنتو-هلسینکی
ZEN	Zürich English Newspaper Corpus	پیکره روزنامه‌های انگلیسی زبان زوریخ

A-a

کارکترهای اکسان‌دار **accented characters**

به‌منظور اطمینان از اینکه متن در یک پیکره در پلتفورم‌های[1] مختلف به‌شیوه یکسانی پیاده‌سازی[2] شود، پیشنهاد می‌شود که برای کارکترهای اکسان‌دار از یک سیستم کدگذاری[3] شناخته‌شده استفاده شود. دستورالعمل‌های **کارگروه نشانه‌گذاری متن (TEI)** پیشنهاد می‌کند که کارکترهای اکسان‌دار به‌عنوان هستینه[4] و با استفاده از & و ; برای نشان دادن آغاز و پایان هستینه، کدگذاری شوند.

جدول ۱. نمونه‌ای از ارجاع‌های هستینه‌ها برای کارکترهای اکسان‌دار، کسر، و نشانه‌های ارزها

کد	توضیح کارکتر	کارکتر
ä	حرف کوچک a با اوملات[1] (یا دیارسیس[2])	ä
á	حرف کوچک a با اکسان اکیوت[3]	á
è	حرف کوچک e با اکسان گریو[4]	è
ô	حرف کوچک o با اکسان سیرکومفلکس[5]	ô
ã	حرف کوچک a با تیلد[6]	ã
æ	لیگاتور[7] کوچک ae	æ
¼	علامت کسر: یک‌چهارم	1/4
£	علامت پوند	£

[1]umlaut [2] diaersis [3] acute accent [4] grave [5] circumflex accent [6]tilde [7] ligature

[1] platforms
[2] render
[3] encoding system
[4] entity

جدول ۱ تعدادی از کارکترهای اکسان‌دار و کدگذاری متناظر آن‌ها و همچنین دو نمونه ارجاع هستینه را برای کسر و ارز نشان می‌دهد (نک. **علائم نگارشی**).

دقت **accuracy**

معیاری اساسی برای ارزیابی **ابزارهای نشانه‌گذاری زبانی** خودکار مانند **تجزیه‌گرها** یا **برچسب‌زن‌های اجزای کلام**. دقت برابر است با تعداد **نمونه‌های** درست برچسب‌خورده تقسیم‌بر کل تعداد نمونه‌ها. دقت را معمولاً به‌صورت درصد نشان می‌دهند. دقتِ به‌روزترین برچسب‌زن‌های دستوری زبان انگلیسی بین ۹۵ تا ۹۷ درصد است (نک. **دقیق‌سازی و بازخوانی**).

پروژه‌های اکوئیلکس **Acquilex Projects**

بودجه دو پروژه اکوئیلکس ازطرف کمیسیون اروپا تأمین شد و مقر آن در دانشگاه کمبریج بود. پروژه اول سودمندیِ ایجاد یک پایگاه دانش واژگانی[1] چندزبانه از نسخه‌های[2] رایانه‌خوانای[3] **فرهنگ‌های** متداول را بررسی کرد. پروژه دوم به بررسی سودمندی **پیکره‌های** متنی رایانه‌خوانا به‌مثابه منبعی از آن دسته از اطلاعات واژگانی‌ای پرداخت که در فرهنگ‌های متداول کدگذاری نشده بودند. این پروژه همچنین به شُرکای چاپ‌ونشر فرهنگ به‌مثابه گزینه ممکنی نگاه کرد تا از آن دسته از **دادگان** واژگانی و نرم‌افزار استخراج پیکره‌ای بهره‌برداری کند که به‌وسیله پروژه‌هایی طراحی شده بود که برای فرهنگ‌نگاری متداول مورد استفاده قرار می‌گرفت. نک. http://www.cl.cam.ac.uk/Research/NL/acquilex/acqhome.html

[1] lexical knowledge base
[2] versions
[3] machine readable