استاندارد ملی ایران
۱-۱۶۹۳۹
تجدیدنظر اول
آذر ۱۳۹۲

جمهوری اسلامی ایران
Islamic Republic of Iran

سازمان ملی استاندارد ایران
Iranian National Standardization Organization

INSO
16939-1
1st.Revision
Dec.2013

مدیریت منبع زبان- تقطیع واژگانی متن‌های نوشتاری- قسمت ۱: مفاهیم اساسی و اصول کلی

Language resource management - Word segmentation of written texts- Part 1: Basic concepts and general principles

ICS: 01.140.10

به نام خدا

آشنایی با سازمان ملی استاندارد ایران

مؤسسهٔ استاندارد و تحقیقات صنعتی ایران به موجب بند یک مادهٔ۳ قانون اصلاح قوانین و مقررات مؤسسهٔ استاندارد و تحقیقات صنعتی ایران، مصوب بهمن ماه ۱۳۷۱ تنها مرجع رسمی کشور است که وظیفه تعیین، تدوین و نشر استانداردهای ملی (رسمی) ایران را به عهده دارد.

نام موسسه استاندارد و تحقیقات صنعتی ایران به موجب یکصد و پنجاه و دومین جلسه شورای عالی اداری مورخ ۹۰/۶/۲۹ به سازمان ملی استاندارد ایران تغییر و طی نامه شماره ۲۰۶/۳۵۸۳۸ مورخ ۹۰/۷/۲۴ جهت اجرا ابلاغ شده است .

تدوین استاندارد در حوزه های مختلف در کمیسیون های فنی مرکب از کارشناسان سازمان ، صاحب نظران مراکز و مؤسسات علمی، پژوهشی، تولیدی و اقتصادی آگاه و مرتبط انجام می شود وکوششی همگام با مصالح ملی و با توجه به شرایط تولیدی، فناوری و تجاری است که از مشارکت آگاهانه و منصفانهٔ صاحبان حق و نفع، شامل تولیدکنندگان، مصرف‌کنندگان، صادرکنندگان و وارد کنندگان، مراکز علمی و تخصصی، نهادها، سازمان های دولتی و غیر دولتی حاصل می شود .پیش نویس استانداردهای ملی ایران برای نظرخواهی به مراجع ذی نفع و اعضای کمیسیون های فنی مربوط ارسال می‌شود و پس از دریافت نظرها و پیشنهادها در کمیتهٔ ملی مرتبط با آن رشته طرح و در صورت تصویب به عنوان استاندارد ملی (رسمی) ایران چاپ و منتشر می شود.

پیش نویس استانداردهایی که مؤسسات و سازمان های علاقه مند و ذی صلاح نیز با رعایت ضوابط تعیین شده تهیه می کنند درکمیتهٔ ملی طرح و بررسی و درصورت تصویب، به عنوان استاندارد ملی ایران چاپ و منتشر می شود .بدین ترتیب، استانداردهایی ملی تلقی می شوند که بر اساس مفاد نوشته شده در استاندارد ملی ایران شمارهٔ ۵ تدوین و در کمیتهٔ ملی استاندارد مربوط که سازمان ملی استاندارد ایران تشکیل می‌دهد به تصویب رسیده باشد.

سازمان ملی استاندارد ایران از اعضای اصلی سازمان بین المللی استاندارد (ISO)[1] ،کمیسیون بین المللی الکتروتکنیک (IEC)[2] و سازمان بین المللی اندازه شناسی قانونی (OIML)[3] است و به عنوان تنها رابط[4] کمیسیون کدکس غذایی (CAC)[5] در کشور فعالیت می کند. در تدوین استانداردهای ملی ایران ضمن توجه به شرایط کلی و نیازمندی های خاص کشور ، از آخرین پیشرفت های علمی، فنی و صنعتی جهان و استانداردهای بین‌المللی بهره‌گیری می‌شود .

سازمان ملی استاندارد ایران می تواند با رعایت موازین پیش بینی شده در قانون، برای حمایت از مصرف کنندگان، حفظ سلامت و ایمنی فردی و عمومی، حصول اطمینان از کیفیت محصولات و ملاحظات زیست محیطی و اقتصادی، اجرای بعضی از استانداردهای ملی ایران را برای محصولات تولیدی داخل کشور و/یا اقلام وارداتی، با تصویب شورای عالی استاندارد، اجباری نماید. سازمان می تواند به منظور حفظ بازارهای بین المللی برای محصولات کشور، اجرای استاندارد کالاهای صادراتی و درجه‌بندی آن را اجباری نماید. همچنین برای اطمینان بخشیدن به استفاده کنندگان از خدمات سازمان ها و مؤسسات فعال در زمینهٔ مشاوره، آموزش، بازرسی، ممیزی و صدورگواهی سیستم های مدیریت کیفیت و مدیریت زیست‌محیطی، آزمایشگاه ها و مراکز کالیبراسیون (واسنجی) وسایل سنجش، سازمان ملی استاندارد ایران این گونه سازمان ها و مؤسسات را بر اساس ضوابط نظام تأیید صلاحیت ایران ارزیابی می کند و در صورت احراز شرایط لازم، گواهینامهٔ تأیید صلاحیت به آن ها اعطا و بر عملکرد آن ها نظارت می کند. ترویج دستگاه بین المللی یکاها، کالیبراسیون (واسنجی) وسایل سنجش، تعیین عیار فلزات گرانبها و انجام تحقیقات کاربردی برای ارتقای سطح استانداردهای ملی ایران از دیگر وظایف این سازمان است.

1- International Organization for Standardization
2 - International Electrotechnical Commission
3- International Organization of Legal Metrology (Organisation Internationale de Metrologie Legale)
4 - Contact point
5 - Codex Alimentarius Commission

کمیسیون فنی تدوین استاندارد

"مدیریت منبع زبان- تقطیع واژگانی متن‌های نوشتاری

قسمت ۱: مفاهیم اساسی و اصول کلی"

رئیس :	**سمت و/ یا نمایندگی**
احمدی، عباس (کارشناس ارشد زبان شناسی)	مدرس دانشگاه آزاداسلامی واحد ایلام
دبیر:	
علی بیگی، محمود (کارشناسی مترجمی زبان انگلیسی)	مدرس جهاد دانشگاهی واحد استان ایلام
اعضاء: (به ترتیب حروف الفبا)	
ابراهیمی، محمد (کارشناسی مترجمی زبان انگلیسی)	کارشناس بانک صادرات ایلام
بسطامی، مهدی (کارشناسی مترجمی زبان انگلیسی)	کارشناس بانک سپه ایلام
خانی، بهزاد (کارشناس زبان و ادبیات فارسی)	کارشناس جهاددانشگاهی واحد استان ایلام
مظلوم، فاطمه (کارشناس میکروبیوشمی)	کارشناس اداره کل استانداردوتحقیقاتصنعتیاستان ایلام
یاری، بهروز (کارشناس زبان انگلیسی)	کارشناس اداره کل کار، رفاه و امور اجتماعی استان ایلام

فهرست مندرجات

پیش گفتار

استاندارد «مدیریت منبع زبان-تقطیع واژگانی متن‌های نوشتاری- قسمت ۱: مفاهیم اساسی و اصول کلی» که پیش‌نویس آن در کمیسیون‌های مربوط تهیه و تدوین شده و در یکصد و سومین اجلاس کمیته ملّی استاندارد اسناد و تجهیزات اداری و آموزشی مورخ۹۱/۱۱/۲۹ مورد تصویب قرار گرفته است، اینک به استناد بند یک ماده ۳ قانون اصلاح قوانین و مقررات مؤسسه استاندارد و تحقیقات صنعتی ایران، مصوببهمن ماه ۱۳۷۱، به عنوان استاندارد ملّی ایران منتشر می شود.

برای حفظ همگامی و هماهنگی با تحولات و پیشرفت‌های ملّی و جهانی در زمینه صنایع، علوم و خدمات، استانداردهای ملّی ایران در مواقع لزوم تجدید نظر خواهد شد و هر پیشنهادی که برای اصلاح و تکمیل این استانداردها ارائه شود، هنگام تجدید نظر در کمیسیون فنی مربوط مورد توجه قرار خواهد گرفت. بنابراین، باید همواره از آخرین تجدید نظر استانداردهای ملّی استفاده کرد.

منبع و مأخذی که برای تهیه این استاندارد مورد استفاده قرار گرفته به شرح زیر است:

ISO24614-1:2012,Language resource management - Word segmentation of written texts – Part 1: Basic concepts and general principles

مقدمه

این استاندارد ملی تقطیع واژگانی در زبان‌های نوشتاری را هدف قرار داده است.این استاندارد بر مفاهیم اساسی و اصول کلی تقطیع واژگانی تمرکز دارد که به زبان‌ها به طور کلی می‌پردازد.

تقطیع واژگانی، تقطیع متن به واحدهای زبانی است که دربردارند معنا می باشند. به عنوان مثال، «کاخ سفید» را می توان به دو واحد معنی دار تقسیم کرد «کاخ» و «سفید» که به کاخی که سفید است اشاره دارد در حالی که «کاخ سفید» مربوط به یکواحد معنی دار است که به محل اقامت رئیس جمهور ایالات متحده آمریکا اشاره دارد.در این استاندارد،این‌گونه واحدهای زبانیِ معنادار واحدهای تقطیع واژگانی(WSU)[1] نامیده می شوند . همانطور که در مثال قبلی نشان داده شد، واحدهای تقطیع واژگانی می تواند از بیش از یک واژه تشکیل شود. یکWSUمی‌تواند حالت‌های مختلفی داشته باشد از جمله: از یک ستاک و وندهایی تشکیل شود (مانند:«بی +کار»)،یک واژه مرکب (مانند: «تخته سیاه»)، اسم خاص باشد (مانند:«خلیج فارس»)، یک اصطلاح باشد(مانند: مثال «دم اسبی باران باریدن»)، یا عبارتی چند واژه‌ای باشد (مانند:«مراقب ۰۰۰ بودن»). برای زبان‌هایی که بین کلمات فاصله می‌گذارند، به عنوان مثال انگلیسی، تقطیع متن به واحدهای تقطیع واژگانی با استفاده از فضاهای موجود به عنوان پایه‌ای برای ایجاد مرزهای واحدهای تقطیع واژگانی ساده استفاده می‌شود ، هرچند که برای ادارهٔ کوته‌نوشت‌ها، نشان‌گذاری و واحدهای چندواژه‌ایِ معنا در بین دیگر موارد باید ملاحظات بیشتری مد نظر قرار گیرد. برای زبان‌هایی که بین واژها فاصله دارند، مثل چینی، ژاپنی یا برای زبان‌هایی که تا حدودی بین واژه‌ها فاصله دارند، مثل تایلندی و کره‌ای تقطیع متن به واحدهای تقطیع واژگانی به رویکرد دیگری نیازمند است.

علاوه بر این، تقطیع واژه‌ها برای زبان‌هایی مانند چینی که ترکیبات گسترده‌ای دارند،و زبان‌هایی که چسپندگی گسترده دارند مانند ژاپنی، کره‌ای و مجارستانی پیچیده است. از سوی دیگر، این واقعیت است کـه زبـان ژاپنـی دستخط‌های چندگانه راپشتیبانیمی‌کند، برای تقطیع واژگان مفید است.

با این حال، فضای خالی به تنهایی برای تقطیع متن کافی نیست. برای مثال «پایِ سیب» بـه عنـوان یـک نـوع کلوچهٔ میوه‌دارِ ساخته شده از سیب درک شده است، پس «پای» و«سیب» به عنوان دوواحد تقطیـع واژهٔ مجـزا در نظر گرفته شده‌اند. از سویی دیگر، با توجه به خواص ترکیبی[2]و اصطلاحی می‌توان آن را به عنوان یک ماهیت

1 -Word segementation units

2-Collocation

واحد و یک واحد تقطیع مجزا آن را تلقی کرد. قواعد تقسیم بندی می‌تواند در میان زبان‌ها متفاوت باشد، حتی وقتی که برای اصطلاحات معادل به کار روند).

تشریح استانداردها برایقواعد و روش‌ها جهت تقطیع واژه‌ها می‌تواند نوآوری و توسعه در زمینه‌هایی مانند یادگیری زبان و ترجمه را تسهیل کند .این امر می‌تواند فن‌آوری‌های مربوط به زبان، از جمله کنترل املا ، کنترلدستور زبان، مراجعه به فرهنگ لغت، مدیریت اصطلاحات، حافظه ترجمه، بازیابی اطلاعات، استخراج اطلاعات و ترجمه ماشینی را بهبود بخشد. به عنوان مثال عدم شناسایی «غزل خداحافظی را خواندن» به عنوان یک واحد تقطیع واژهٔ تک، در فن آوری ترجمه‌های ماشینی، به جای ترجمه اصطلاحی یک ترجمهٔ تحت اللفظی از آن می‌کنند.

مدیریت منبع زبان- تقطیع واژگانی متن‌های نوشتاری

قسمت ۱: مفاهیم اساسی و اصول کلی

۱هدف و دامنه کاربرد

هدف از تدوین این استاندارد تعیین مفاهیم اساسی و اصول کلی تقطیع واژگان و دستورالعمل های مستقل از زبان است تا تقطیع متون نوشتاری را به شیوه‌ای معتبر و تجدیدپذیر، به واحدهای تقطیع واژه (WSU)مقدور سازد.

یادآوری_در تحقیقات مربوط به زبان و صنعت، واژه، مفهومی اساسی و ضروری است.بنابراین برای تقطیع یک متن به واژه‌ها ضروری است از آنچه تشکیل دهندهٔ واژه‌ها هستند یک تعریف جهانی داشته باشیم. کسی نمی‌تواند برای تعیین حدود واژه‌ها تنها از قواعد فاصله ها و نشان‌گذاری استفاده کند. چنین قواعدی برای موقعیت‌هایی مانند اسم‌های مرکب که با خط فاصله به هم پیوسته‌اند، کوته نوشت ها، اصطلاحات و یا واژه‌های عبارت مانند که شامل نماد ها و یا شماره هستند، توضیحی ندارند.تقطیع واژگانی حتی برای زبان‌هایی مانند چینی و ژاپنی که جهت جداسازی واژه‌ها از فاصله استفاده نمی‌کنند، و برای زبان‌های چسپیده مانند کره‌ای که برخی رده‌های کلمات دستوری را به عنوان وندها استفاده می‌کنند، مشکل‌تر است.

بسیاری از برنامه‌های کاربردی و زمینه‌هایی که نیاز به تقطیع متن به واژهدارند،این استاندارد در موارد زیر کاربرد دارد:

ترجمه

واژه شماری روش اصلی برای محاسبه هزینه‌های ترجمه می باشد. تقطیع واژگان تابعی استاندارد در سامانه‌های حافظهٔ ترجمه و ابزار ترجمه به کمک رایانه (CAT)[1] است .تقطیع واژگان به‌وسیلهٔ ابزار استخراج اصطلاح انجام می‌شود، که گاهی اوقات در سامانه های مدیریت اصطلاحات و ابزارهای ترجمه به کمک رایانه(CAT) صورت می‌گیرد.

1-Computer-Assisted Translation

مدیریت محتوا

اکثر سامانه‌ها و پایگاه دادگان‌های مدیریت محتوا اجازهٔ جستجوی کلمات تکی را می‌دهند. محتوایی که جستجو می‌شود باید تقطیع شود تا اجازه انطباق با واژهٔ مورد جستجو را بدهد.علاوه بر این، عملیات جستجو نیاز به دانش مرزهای کلمات دارد.

فن‌آوری های گفتار

سامانه‌های تبدیل متن به گفتار، گفتار را بر اساس کلمات تولید می‌کنند و در نتیجه برای جستجوی کلمات، تخصیص فشار صدا[1]، تخصیص الگوی عروضی[2] و غیره به تقطیع واژگان نیاز دارند.

زبانشناسی رایانه‌ای

سامانه‌های متنوع پردازش زبان طبیعی(NLP)[3]باید متن را به واژگان تقطیع کنند تا وظایف خود را انجام دهند.
سامانه های NLP عبارتند از:

- پردازنده‌های صرفی-نحوی،
- تجزیه‌کننده های نحوی،
- کنترل کننده‌های املا،
- سامانه‌های طبقه بندی متن ، و
- حاشیه نویسان مجموعهٔ پیکرهٔ زبانی.

فرهنگ نویسی

منابع واژگانی اغلب بر اساس اندازه، معمولا با مراجعه به تعدادکلمات، ارزیابی می‌شوند.

یادآوری_اندازه منابع زبان، معیار اساسی برای مدیریت آن‌ها است .تعیین اندازهٔ منابع زبان به طور معمول با شمارش کلمات به دست می‌آید. با این حال، چون برنامه های کاربردی NLP از روش‌های تقطیع مختلفی استفاده می‌کنند، هر یک به صورتی متفاوت تعداد واژه‌ها را محاسبه می‌کنند و برای یک متن یکسان به مجموع کلمات متفاوتی می‌رسند. یک اندازه‌گیری استاندارد، قابل اعتماد، تجدیدپذیر، نتایجی سنجیدنی را فراهم می‌سازد. این بدان معنا نیست که برنامه های کاربردی ممکن نیست

1- Stress assignment

2- Prosodic pattern assignment

3- Natural language process

ازشیوه‌های تقطیع منحصر به نرم افزار خود استفاده کنند. به عنوان مثال، یک برنامه کاربردی ترکیب گفتار ممکن است در مقایسه با برنامه دیگری متن را به واحدهای کوچکتر یا بزرگتر تقطیع کند.

۲ مراجع الزامی

مدارک الزامی زیر حاوی مقرراتی است که در متن این استاندارد ملّی ایران به آن‌ها ارجاع شده است. بدین ترتیب آن مقررات جزئی از این استاندارد ملّی ایران محسوب می شود. در صورتی که به مدرکی با ذکر تاریخ انتشار ارجاع داده شده باشد، اصلاحیه‌ها و تجدید نظرهای بعدی آن مورد نظر این استاندارد ملّی ایران نیست. در مورد مدارکی که بدون ذکر تاریخ انتشار به آن‌ها اشاره شده است، همواره آخرین تجدید نظر و اصلاحیه‌های بعدی آن‌ها مورد نظر است.

استفاده از مراجع زیر برای این استاندارد الزامی است:

2.1- ISO 1087-1:2000, Terminology work — Vocabulary — Part 1: Theory and application

2.2- ISO 1087-2:2000, Terminology work — Vocabulary — Part 2: Computer applications

2.3- ISO 24611, Language resource management — Morpho-syntactic annotation framework)

2.4- ISO 24612, Language resource management — Linguistic annotation framework (LAF))

2.5- ISO 24613:2008, Language resource management — Lexical markup framework (LMF)

2.6- ISO 12620, Computer applications in terminology — Data categories

2.7- ISO 16642:2003, Computer applications in terminology — Terminological markup framework

2.8- ISO 30042:2008, Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)

۳اصطلاحات و تعاریف

در این استاندارد، اصطلاحات و تعاریف زیر به کار می‌رود:

۳-۱

کوته نوشت[1]

1- Abbreviation

تعیین کلامی است که با حذف کلمات یا حروف از یک فرم دیگر شکل گرفته و مفهوم یکسانی را مشخص می‌کند.
[ISO 1087-1:2000]

۲-۳

وند[1]

تک‌واژ مقید[2] (۳-۵) که ممکن است به یک ستاک (۳-۲۲) یا یک لغت (۳-۱۴)اضافه شود.

یادآوری_وندها را می‌توان به چندین نوع فرعی مانند پیشوند، پسوند، میانوند ودوروند[3]طبقه‌بندی نمود. وندها می‌توانند اشتقاقی باشند یا می‌توانند صرفی یا چسپیدهباشند.

۳-۳

پیوند[4]

روند به هم پیوستن یک یا چندوند (۳-۲)بهیک ستاک (۳-۲۲) می‌باشد.
[ISO 24613:2008]

یادآوری_اینگونه وندی را در فارسی نداریم

۴-۳

عاریه‌گیری

روند تشکیل واژه است که در آن یک عبارت زبانی از زبان دیگر اقتباس می‌شود، معمولا زمانی‌که هیچ واژه‌ای برای شی یا مفهوم جدید وجود ندارد.

1- Affix

2- Bound morpheme

3- Circumfix-۳ وندی است که از دو قسمت مجزا تشکیل شده، اطراف ستاک را فرا گرفته و به آن می‌چسپد

4- Agglutination

۵-۳

تک‌واژ مقید[1]

تک واژی (۳-۱۸) که فقط با یک یا چندین تک واژ دیگر می‌آید.
[ISO 24613:2008]

مثال ۱:فارسی: «گاه» نمی تواند به عنوان یک واژه در متن استفاده شود.در عوض، آن رابه عنوان یکی از عناصر تشکیل دهنده بسیاری از کلمات، مانند «درگاه»، «دادگاه»، «گاه‌شمار» به‌کار می‌برند.

مثال ۲:فارسی: پیشوند«بی» که معادل پیشوندهای « un, non, ir, a, in» در زبان انگلیسی می‌باشد. بی‌صدا = Unvoiced، بی‌تاثیر =non-functioning، بی‌قاعده=irregular، بی‌نواخت = atonic، بی‌اثر= inactive.

۶-۳

ترکیب[2]

واژه‌ای(۳-۲۳)که از دو یا چندلغت (۳-۱۴) ساخته می‌شود.

یادآوری۱-ازاستاندارد ISO 24613:2008، تعریف ۳-۱۰ اقتباس شده است.

یادآوری۲- ترکیب ممکن استدرون‌مرکز[3] باشداگر یک هسته[4] (یعنی بخش اساسی که حاوی معنای اصلی کل ترکیب) و پیراینده هایی (که این معنا را محدود می کنند)داشته باشد، یا برون مرکز[5]باشد اگر که هسته نداشته باشد. یک ترکیب می تواند طولانی باشد.با توجه به درجه واژه‌بندی[6]، دو زیر گروه از ترکیب وجود دارد: واژهمرکب و عبارت مرکب.

۷-۳

ترکیب کردن[1]

1- Bound morpheme

2-Compound

3- Endocentric

4- Head

5- Exocentric

6- Lexicalization

واژه‌سازیی است که در آنواژه جدید از کنار هم قرار گرفتن حداقل دو لغت (۳-۱۴)، در شکل اصلی خود و یا با دگرگونی های کمی، تشکیل می‌شود.
[ISO 24613:2008]

۸-۳

اشتقاق[۲]

تغییر در شکل یک کلمه (۳-۲۳) است برای ایجاد واژه جدید (۳-۲۳)، معمولا با تغییر ستاک (۳-۲۲) و یا با افزایش وند.
[ISO 24613:2008]

۹-۳

تک‌واژ آزاد[۳]

تک‌واژی (۳-۱۸) است که می‌تواند به خودی خود به عنوان یک واژه (۳-۲۳) مورد استفاده قرار گیرد.

مثال ۱:واژه داده شدهٔ«خوبی»، «خوب» یک تک‌واژ آزاد است، در حالی که «ی» نیست.دومی، تکواژمقید است.

۱۰-۳

هم نویسه[۴]

هر یک از دو یا چند اشکال واژه (۳-۲۴) یا واژگان (۳-۲۳) با املا یکسان هستند که نمایانگر مفاهیم (هم‌نویسهٔ معنایی[۵]) یا عملکردهای نحوی (هم‌نویسهٔ‌نحوی[۶]) مختلف می‌باشند.
[ISO 1087-2:2000]

1- Compounding

2- Derivation

3- Free morpheme

4- Homograph

5 - Semantic homography

6 - Syntactic homography

۱۱-۳

تصریف[1]

فرایندی که در آن یک شکل واژه (۳-۲۴) با اضافه کردن وند (۳-۲) به ستاک (۳-۲۲) به وجود می‌آید.

یادآوری۱_ تصریف فرایندیدستوریاست و نه واژگانی.

۱۲-۳

مدخل واژه‌نامه‌ای[2]

شکل قراردادی انتخاب شده برای نشان دادن یکلغت (۳-۱۴) است.

[ISO 24613:2008]

مثال:در مجموعه شکل‌های واژگانیمعین به عنوان مثال«یافت»، «می‌یابد»،«یافتن»،درزبان فارسی، شکل«یافتن» به عنوان مدخل واژه‌نامه‌ای به نمایندگی از این گروه از شکل‌های واژگانی انتخاب می‌شود.

۱۳-۳

مدخل واژه‌نامه‌ای ساختن[3]

روند تعیین مدخل واژه‌نامه‌ای (۳-۱۲) برای یک شکل واژگانی (۳-۲۴) معین در یک متن می‌باشد.

مثال: واژه معین «یافت» در فارسی، در مدخل واژه‌نامه‌ای ساختن منجر می‌شود به «یافتن» به عنوان مدخل واژه‌نامه‌ای آن.

یادآوری- برگرفته از ISO 1087-2:2000، تعریف ۳-۱۹ و ISO 30042:2008، تعریف ۳-۱۴.

1- Inflection

2- Lemma

3 - Lemmatization

۱۴-۳

لغت[1]

واحدی انتزاعی که عموماً در دسته‌ای از شکل‌ها شرکت می‌کند که معنای عمومی یکسانی دارند.
[ISO 24613:2008]

یادآوری۱-یک‌لغت ممکن است بخشی ازلغتی دیگر باشد، در نتیجهٔیک اشتقاق و ترکیب است.
یادآوری۲-«شکل» در استاندارد ISO 24613 به عنوان «زنجیره‌ای از واژک‌ها» تعریف شده است.

۱۵-۳

قاموسی کردن[2]

روند ساختن یک واحد زبانیکه به عنوان یک واژهعمل کند.

یادآوری-چنین واحد زبانی می‌تواند یک واژک باشدبه عنوان مثال «خندیدن»،زنجیره‌ای از واژک‌هاباشدبه عنوان مثال «گوجه فرنگی» و یا حتی یک عبارتمانند « ریق رحمت را سر کشیدن»باشدکه یک عبارت اصطلاحی را تشکیل می‌دهد.

۱۶-۳

قاموس[3]

فهرستی از مدخل‌ها که عمدتاً با مدخل‌های واژه‌نامه‌ای (۳-۱۲) و اطلاعات مربوط به آن‌ها آغاز می‌شوند.

۱۷-۳

واژک[4]

شکل سطحی که به وسیلهٔ یک تک‌واژ (۳-۱۸) منحصر به فرد نشان داده می‌شود.

مثال:در زبان فارسی، واژک‌های تک‌واژهای جمع‌ساز «ها» و «ان» شامل «ها»، «ان»، و «تهی» هستند (به عنوان مثال «میزها»، «پرندگان» و «اسرار») که «تهی» هیچ شکل سطحی منحصربه‌فردی ندارد. از این‌رو واژهٔ «میزها» شامل دو واژک «میز» و «ها»است در حالی‌که تک‌واژهای مربوط به واژک‌های «پرنده» و «گان» به ترتیب «پرنده» و «ان» هستند.

1- Lexeme

2-Lexicalization

3- Lexican

4-Morph

۱۸-۳

تک‌واژ[1]

کوچکترین واحد معنایی که به‌وسیلهٔ زنجیره‌ای از واج‌ها یا زنجیره‌ای از حروف بیان می‌شود.
[ISO 24613:2008]

یادآوری–دو نوع زیرمجموعه برای تک‌واژ وجود دارد: تک‌واژ آزاد و تکواژ مقید.

۱۹-۳

عبارت چند واژه‌ای[2]

لغتی (۳-۱۴) که از زنجیره‌ای از لغت‌ها تشکیل شده و خواصی دارد که از خواص تک تک لغات یا حالت طبیعیِ ترکیب آنها قابل پیش‌بینی نیست.
[ISO 24613:2008]

یادآوری- عبارت چند واژه‌ای می‌تواند یک عبارت مرکب باشد [ترکیبی واژه‌ایبا ترکیبی عبارتی، یک اصطلاح، بخشی از یک جمله و یا یک جمله (به عنوان مثال یک ضرب المثل یا نقل قول آشنا)]. همیشه تعیین ادات سخن برای تمام محدودهٔ عبارت چند واژه‌ایممکن نیست.

۲۰-۳

ترکیب عبارتی[3]

واژه‌ای (۳-۲۳) متشکل از دو یا چند لغت(۳-۱۴)، که معنای آن از عناصر تشکیل دهنده آن قابل پیش بینی است.

مثال: «درخت سیب»در زبان فارسی یک ترکیب عبارتیمتشکل از دو لغت«درخت» و «سیب» است، که معانی آن در معنای این ترکیب حفظ شده است.

یادآوری۱- اصطلاحات از دو یا چند عنصر واژگانی استفاده می‌کنند، امایک ترکیب عبارتی را تشکیل نمی‌دهند.
یادآوری۲-ترکیب عبارتی ممکن است از نظر برخی زبان‌شناسان به عنوان یک عبارت به حساب آید. با این‌حال به سبب ابهام در پیش‌بینی پذیری معنایی و درجهٔ فرآیند واژه‌سازی، در عمل همیشه تمایزی روشنی بین یک ترکیب واژه‌ای و یک عبارت ترکیبی،

1- Morpheme

2- Multiword expression

3- Phrasal compound

یا بین یک ترکیب عبارتی و یک عبارت وجود ندارد. آمار واژگانی[1]– در بسامد واژگانی به طور اخص- نقش مهمی را در این رابطـه ایفا می‌کند.

۳- ۲۱

تکرار[2]

فرایندی که در آن تمامی واژه (۳-۲۳)، و یا بخشی از آن، تکرار شده است.

۳- ۲۲

ستاک

واحد زبانی است که شکل آن کوچکتر یا برابر با شکل یک لغت (۳-۱۴) واحد استو ممکن است تحت تاثیر روندتصریفی، پیوندی[3]، ترکیبی و یا اشتقاقی قرار گیرد.
[ISO 24613:2008]

۳- ۲۳

واژه

لغتی (۳-۱۴)است که ، به عنوان یک ویژگی حداقل، بخشی از اجزا کلام را دارد.
[ISO 24613:2008]

۳- ۲۴

شکل واژه[4]

گونه‌ای صرفی-نحوی از یک واژه (۳-۲۳) معین است.
[ISO 1087-2:2000]

مثال: در زبان فارسی، رشتهٔ «پیدا کرد»،«می یابد» و «پیدا می‌کند»شکل‌هایی از واژهٔ«پیدا کردن» هستند.

۱- تکنیکی آماری است که در بحث سیر تکامل زبان‌های مختلف مورد استفاده قرار می‌گیرید 1- Lexico-statistics:

2- Reduplication

3 -Agglutinative

4-Word form

۳- ۲۵

تقطیع واژگانی[1]

فرآیند تقطیع متن به زنجیره‌ای از واحدهای تقطیع واژگانی (۳- ۲۶) می‌باشد.

۳- ۲۶

واحدهای تقطیع واژگانی(WSU)[2]

شکل واژه (۳-۲۴) و یا یک رشته علامت از نوعی دیگر است که به عنوان یک واحد با آن برخورد می‌شود.

یادآوری- یک رشته علامت که به شکل یک واژه نیست ممکن است از علائم عددی، علائم خارجی، علائم نقطه گذاری و یا برخی از علائم متفرقه دیگر مانند رادیکال‌های چینی، نمادهای شیمیایی مانندH2O، یا ترکیبی از علائم لاتین و عددی تشکیل شده باشد مانندF16.

۳- ۲۷

ساختار واژه[3]

ساختار داخلی یک واژه (۳-۲۳) حاصل از تجزیه و تحلیل واژه‌شناسی است.

یادآوری-در زبان‌هاپیوستهمانند کره‌ای، ژاپنی و ترکی واژه ممکن است رشته‌ای ازتک‌واژها ، با ضریب نسبتا بالایی از تک‌واژ در واژه، باشد که در آن هر وند مورد بحثی (به هر دو صورت نحوی و اشتقاقی)به طور معمول بیانگر یک معنای خاص دستوری روشن، و یک به یک است. ساختار یک واژه در این زبان‌ها، باتک‌واژهایآزاد و وندهای جداگانه به عنوان عناصر تشکیل دهندهٔ آن،می توانند بسیار پیچیده باشد.

۳- ۲۸

واژهٔ مرکب[4]

یک ترکیب (۳-۶) که معنای کلی آن از بخش‌های تشکیل دهندهٔ آن کاملا قابل پیش‌بینی نیست.

1- Word segementation

2-Word segementation units

3-Word structure

4-Word compound

مثال:« گره‌گشا»، « همدم»

۴چارچوبی اساسی برای تقطیع واژگان

۴-۱ مفاهیم اساسی مربوط به تقطیع واژگان

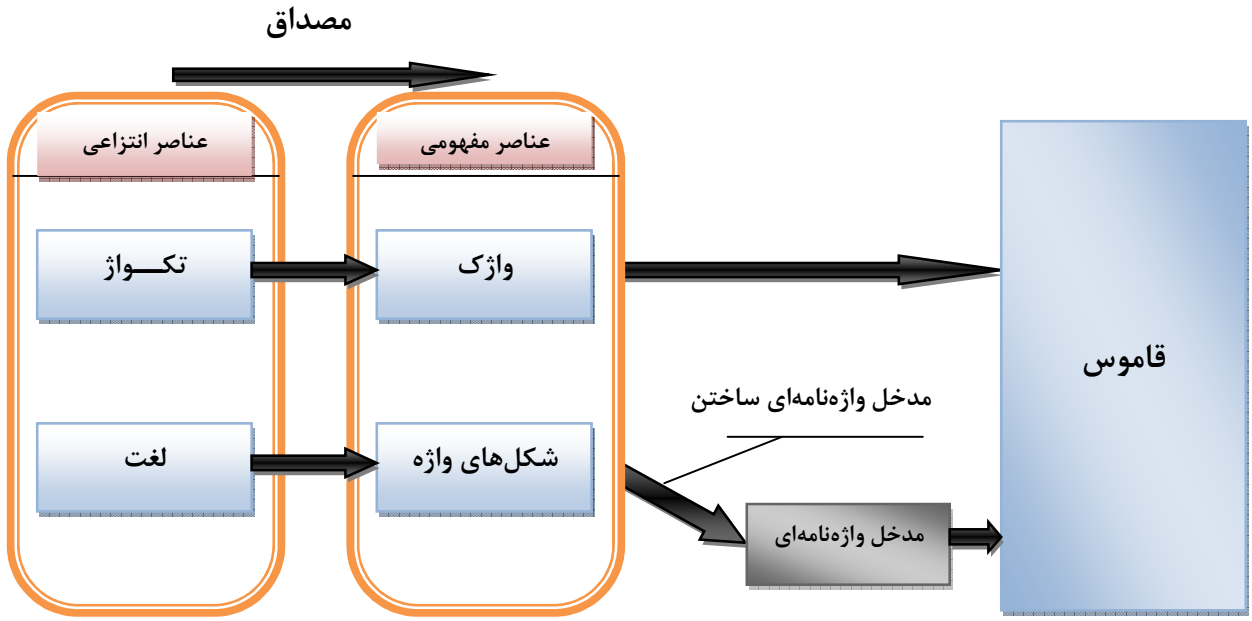
مفاهیم شرح داده شده در این بند در درک اصول تقطیع واژگان ضروری است.

شکل ۱ارتباط میان عناصر انتزاعیِ «تک‌واژ» و« لغت» و عناصرمفهومیِ « واژک»، «شکل واژه» و «قاموس»را نشان می دهد. شکل مفهومی یک ت‌تک‌واژ یک واژک است. شکل مفهومی یک لغت یک شکل واژه است. یک قاموس عموماً از مدخل‌های واژه‌نامه‌ای تشکیل می‌شود که از طریق فرآیند مدخل واژه نامه‌ای ساختن از شکل‌های واژه به‌دست می‌آیند.

یادآوری ۱–اصطلاحاتی همچون « تک‌واژ» و «واژه» دارای معانی مختلف در زمینه های زبان‌شناسی و اصطلاحات هستند.این‌ها و دیگر اصطلاحاتی که در بند ۲ شرح داده شدند، با توجه به معانی زبان‌شناسی آن‌هاست.

واژه‌شناسی مطالعه واحدهای معنی‌دار زبان است و اینکه برای شکل دادن به واژه‌ها چگونه می توان آنها را با هم ترکیب کرد. واژه‌شناسی را می توان به واژه‌شناسی لغوی، که عمدتا با شکل‌گیری واژه بر اساسلغات سرکار دارد، و یاواژه‌شناسی تصریفی یا واژه‌شناسی پیوندی[1] (بر اساس گونهٔ زبان)، که عمدتا با شکل گیری واژه بر اساس تک واژه اسروکار دارد تقسیم کرد. واژه‌شناسی لغوی شامل فرآیندهای اشتقاق، ترکیب، کوته نوشت، عاریه‌گیری و تکرار است.

1-Agglutinative morphology



شکل ۱ - ارتباط بین عناصر انتزاعی و مفهومی در ساخت یک قاموس

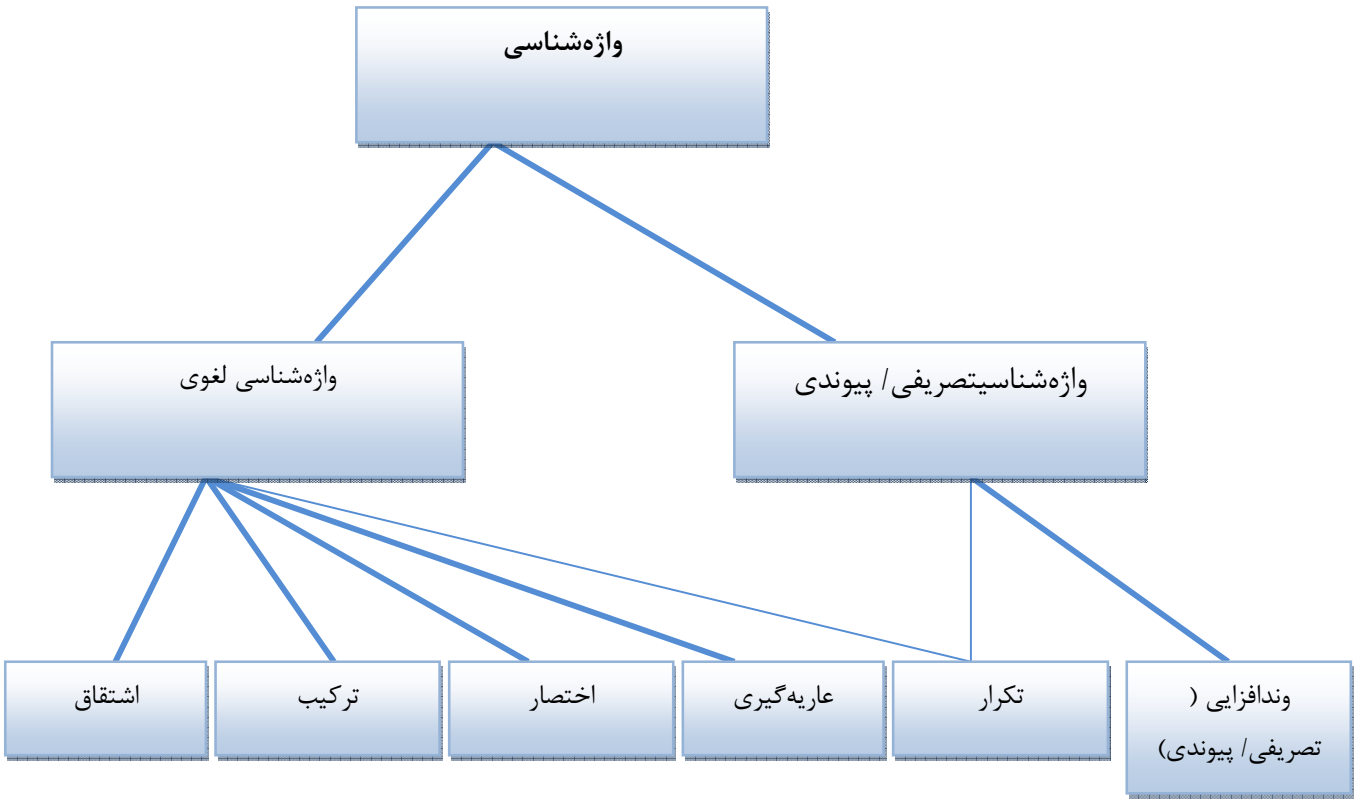
یادآوری۲- بیشتر اصطلاح «واژه‌شناسی لغوی»استفاده می‌شود تا « واژه‌شناسی اشتقاقی» چرا که اشتقاق تنها یک فرآیند شکل گیری واژه است.

واژه‌شناسی تصریفی یا پیوندیشامل دو نوع متفاوت ازنه‌تنها وندافزاییبلکه تکرار می‌باشند. تکرار ممکن است باعث شکل‌گیریواژگان جدید شود، به همین دلیل است که آن را نیز یک فرایند واژه‌شناسی لغوی در نظر می‌گیرند. به عنوان مثال، در آفریکانس[1] از تکرار برای تاکید بر مفهوم واژهٔ تکراری استفاده می‌کنند، مانند«krap» به معنی « خراشیدن»است در حالی‌که «krap krap-krap »به معنی «به شدت خراشیدن» است. برای زبانهای پیوندی، که در آن‌ها وندها به ستاک می‌چسپند، یک مجموعه خاص از قواعدواژه‌شناسی به منظور تقطیع واژگانی مورد نیـاز است.

یادآوری۳-این قواعد در استاندارد ISO 24614-2معین شده‌اند.

1-Afrikaans: زبان آفریقای جنوبی

شکل ۲ را ببینید.

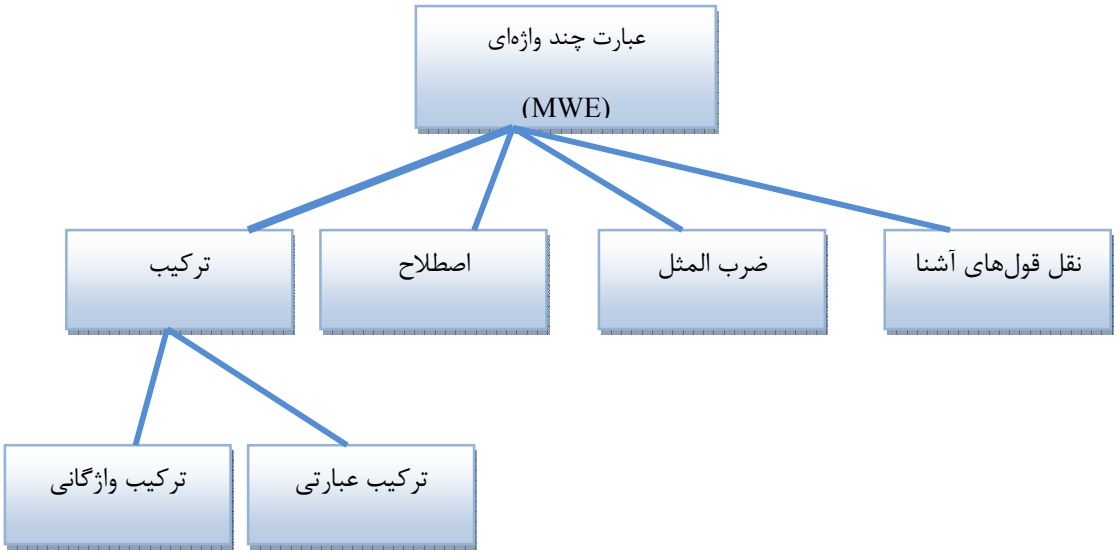


شکل ۲- سامانه واژه‌شناسی در زبان‌ها

عبارات چند واژه‌ای (WWE)[1] شامل ترکیبات، اصطلاحات، ضرب المثل‌ها و نقل‌قول‌های آشنا است. به شکل ۳رجوع کنید.ترکیبات شامل ترکیبات واژگانی وترکیبات عبارتی می‌باشند.معنی ترکیب واژه‌ای نمی تواندازمعنای تک تک بخش‌های آن گرفته شده باشد. به عنوان مثال،«کاخ سفید»را به عنوان محل اقامت ریاست جمهوری ایالات متحده، اشاره به یک مفهوم منحصر به فرد دارد، نه فقط یک کاخی که سفید است.بااین حال،معنای ترکیب عبارتی می تواند از معنای تک تک بخش‌های آن گرفته شده باشد. به عنوان مثال«بادمجان بم» یک بادمجانیاست که در بم می‌روید. اگر چه«بادمجان اصفهان» هم یکبادمجان است که در اصفهان می‌روید، مورد پیشینیک ترکیبعبارتیبه حساب می‌آید (نگاه کنیدبه مقدمه ومثال مورد ۵- ۲۰)، و در نتیجه یک عبارت چند

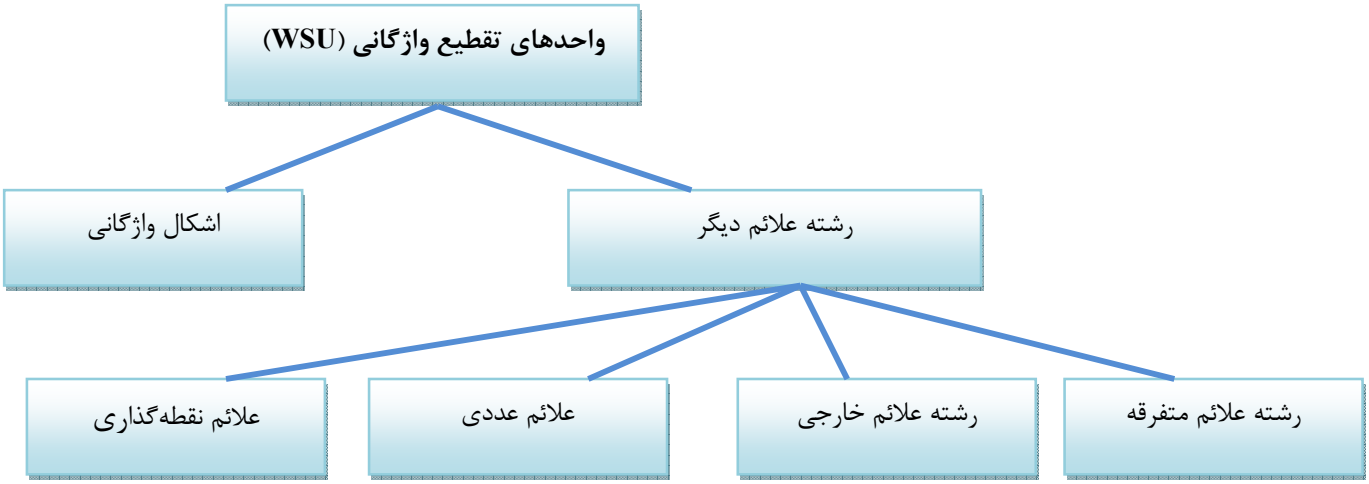
1- Multi Word Expression

واژه‌ای را تشکیل می‌دهد، چرا که ترکیب واژگان«بادمجان بم»متداول است وحتی درعبارات اصطلاحی استفاده می‌شود، « بادمجان بم آفت ندارد»، در حالی که « بادمجان اصفهان» چنین خواصی را ندارد.



شکل ۳- انواع عبارات چند واژه‌ای

واحدهای تقطیع واژگانی ازاشکال واژه‌ها و رشته علائم دیگری تشکیل می‌شود. رشته علائم شامل علائم عددی یا خارجی، علائم نقطه گذاری ویا برخی ازرشته علائم دیگر مانندعلائم صامت و مصوت در متن عربی هستند یا با آنها ترکیب می‌شوند. به عنوان مثال،«هان!»حاوی علامت تعجب در واژه است.



شکل ۴- انواع واحدهای تقطیع واژگانی

۴-۲ منابعی که می‌تواند تقطیع واژگان را تسهیل کند

روند تقطیع واژگان درحوزه زبانی خاص می‌تواند از اجزا و منابع زیر بهره‌گیرند:

الف- قاموس مربوطه؛

ب-لیستی از وندها، شامل پیشوندها، پسوندها، ومیان وندها، در صورت وجود هر کدام؛

پ-لیستی از تک واژهای مقید، علاوه بر وندها؛

ت-مشخصات برای واژه شناسی زبان- جهت تعیین خروجی تقطیع واژگانی براساس پدیده های وابسته به زبان،تحت اصول شرح داده شده در بند۴؛

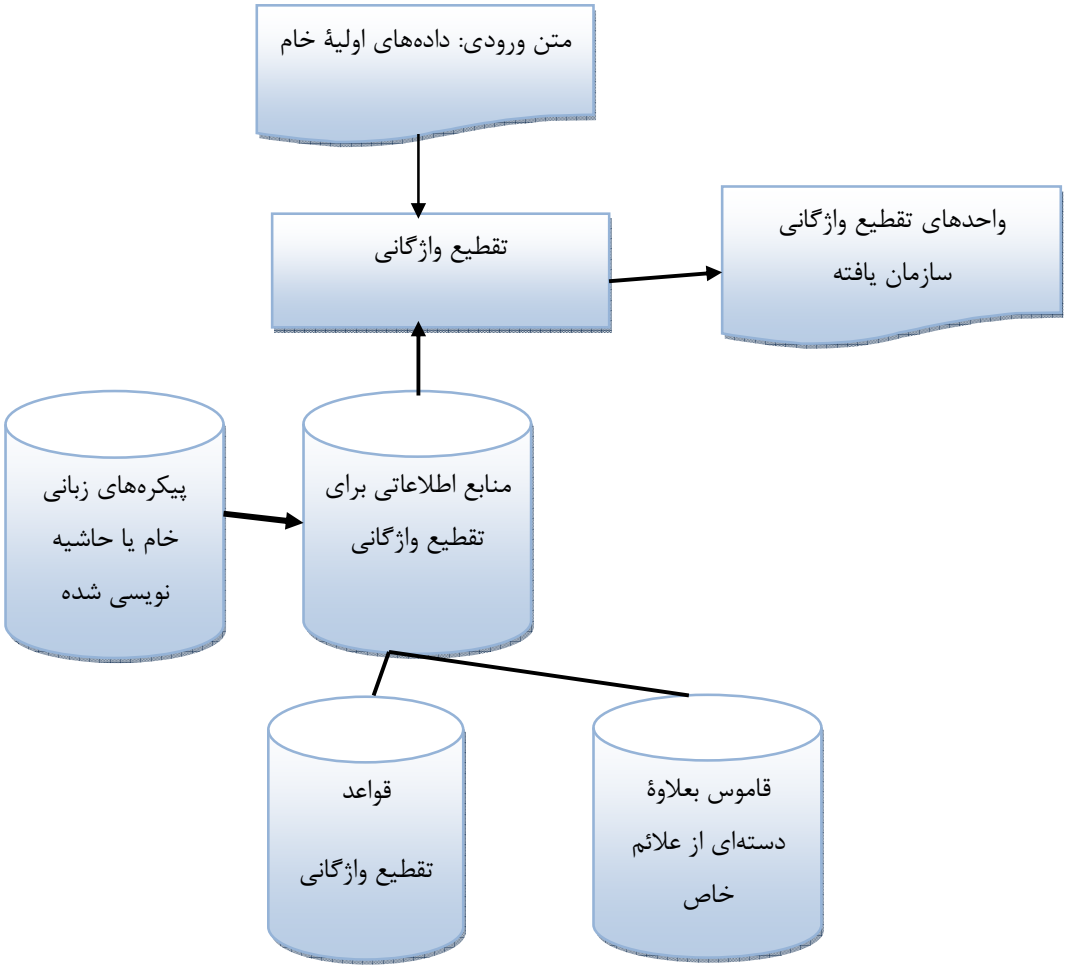
ث- پیکره‌ای زبانی به نمایندگی از یک زبان.

به منظور مطمئن شدن از سازگاریدرتقطیع واژگانی متن‌های مختلف (یایکمتن باابزارهای مختلف)وبرای اطمینان ازاینکه وقتی که برای شمارش نشانه‌هایِ (رجوع کنید به ۴-۳) متن یک سند تقطیع اعمال می‌شود، اعدادقیاس پذیری را فراهم می‌کند، منابع ذکر شده از الف) تا هـ) در بالا باید باتوجه به محتویاتشان با جزئیات شرح داده شوند.

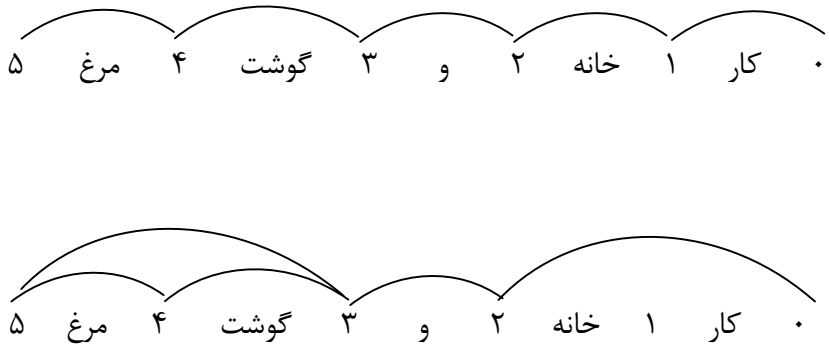
۴-۳ روندتقطیع واژگان

شکل ۵ روند تقطیع واژگانی راتشریح می‌کند.

با داده‌های خام اولیهٔ معین، متن به علائم تقطیع و با شاخص‌های مکانی علامت‌گذاری می‌شود، و سپس با توجه به استاندارد ISO 24612 به واحدهای اصلی مناسب تقطیع می‌شود. پیکرهٔ زبانیِ خام و حاشیه‌نویسی شده، پایه‌ای برای ایجاد یک قاموس فراهم می آروند که شامل اشکال واژگانی و احتمالاً لیستی از تکواژهای مقید و علائم باشد. دسته‌ای از قواعد تقطیع واژگانی نیز فراهم می‌شود. پیکره زبانی، قواعد تقطیع واژگانی و قاموس با هم منابعی را تشکیل می‌دهند که برای تبدیل تقطیع اساسی به یک تقطیع دربردارندهٔ واحدهای تقطیع واژگانی ضروری هستند.



شکل ۵- روند تقطیع واژگانی



شکل ۶- نمایش تقطیع اساسی و تقطیع واژگانی

در سطح تقطیع اولیه، هر نویسه به عنوان یک محدوده بین دو شاخص مکانی (به عنوان مثال اولین علامت «کار»در شکل ۶ با یک محدوده <۱، ۰>مشخص شده است) در سطح حاشیه نویسی زبانی تقطیع واژگانی اولین خروجی «کارخانه» به عنوان یک واژه شناخته می‌شود، با محدودهٔ <۲، ۰> مشخص شده است، زیرا دو واژه را نمی‌توان به تنهایی به حساب آورد. واحد دوم یک واژهٔ تک علامتی «و» است. واحد سوم «گوشت مرغ» یک ترکیب عبارتی است، با محدودهٔ <۵، ۳> با یک ساختار داخلی که متشکل از دو واحد تقطیع واژگانی «گوشت» <۴، ۳> و «مرغ» <۵، ۴>مشخص شده اند.در مورد دوم دو واحد تقطیع واژگانی وجود دارند چون این دو علامت می توانند به طور مستقل وجود داشته باشند و هر یک می‌تواند به معنای واژه کمک کنند.

تقطیع واژگانی بر روی متن خام به‌کار می‌رود.تقطیع واژگانی باعث تقسیم متن داده شده به رشته‌ای از واحدهای تقطیع واژگانی می‌شود، وقتی که تقطیع‌های جایگزین مجاز باشند، یک واحد تقطیع واژگانی می تواند یک ساختار تقطیع درونی داشته باشد. در جملهٔ«رضا ایالات متحده آمریکا را ترک کرد» در این بخش از متن بر اساس برخی از قواعد تقطیع، ابتدا می‌توان آن‌را به بخش‌هایی به نام«نشانه»[1]تقطیع کرد، که در اینجا به سادگی بر اساس فاصله می‌باشد(در زبان‌هایی مانند: چینی که از فاصله‌ها استفاده نمی‌کنند، قواعد مختلفی برای اجرای نشانه‌گذاری[2] باید مورد استفاده قرار گیرد).سپس با مراجعه به قاموس، رشته‌ای از برخی از این بخش‌ها مانند «ایالات متحده آمریکا» می‌تواند به عنوان یک واحد زبانی به نام «واژه» و یا به عنوان عبارت چند واژه‌ای[3]، که به عنوان یک نوع واژه در نظر گرفته شده، تلقی شوند. نتایج مرحله دوم به محتویات قاموس بستگی دارد، برخی قاموس‌ها ممکن است شامل تمام رشته «ایالات متحده آمریکا» به عنوان یک مدخل واژه‌نامه این باشند، اما «ایالات متحده » یا حتی «آمریکا»را شامل شوند.

۵اصول عمومی تقطیع واژگان

۵-۱ اصول جهانی واژه شناسی

اصل جهانی و بنیاد اساسی استاندارد ISO 24614این است که هر زبانی دارای واژه ها و واحدهای کوچکتری نام«تک‌واژ»می‌باشد.

1- Token

2- Tokenization

3- Multi word expression

۵-۲ اصول معتبر ساختن یک واحد تقطیع واژگانی

۵-۲-۱ کلیات

دو دسته از اصول مستقل از زبان برای معتبر ساختن واحد تقطیع واژگانی معین هستند: یکی از دیدگاه زبانی و دیگری از دیدگاه کاربردی.استثنائات مختصزبان در دیگر بخش‌های استاندارد ISO 24614 که با زبان‌های خاص سرو کار دارد، توضیح داده شده است. اصول مختلف ممکن است در موقعیت‌های مختلف اعمال شوند، حتی برای رشته‌های یکسان متن.

۵-۲-۲ اصول از دیدگاه زبانی

الف- اصل تک‌واژ مقید[1]

اگر تک‌واژ مقید به یک واژه وصل شود، آنگاه نتیجه یک واحد تقطیع واژگانی است (به عنوان مثال « بی» به عنوان تکواژ مقید در « بی‌نتیجه»).

ب- اصل تمامیت لغوی[2]

استفاده از قواعد نحوی ساختار داخلی یک واژه را مورد بررسی قرار نمی‌دهد .اگر کاندید یک واژه این اصل را برآورده سازد، آنگاهاحتمالاً یک واحد تقطیع واژگانی است .به عنوان مثال، بین علائم «کاخ سفید»، وقتی که به محل اقامت ریاست جمهوری ایالات متحده اشاره دارد، هیچ چیز نمی تواند درج شود ، به عنوان مثال«کاخ پاک سفید » در حالی‌که می توان گفت «کاخ سفید تمیز»وقتی‌که به هر کاخ سفیدیاشاره دارد.

پ- اصل غیر قابل پیش بینی بودن معنای یک واژه اززیربخش‌های آن

اگر کاندید یک واژه دارای یک خاصیت غیرقابل‌پیش‌بینی‌بودن معنایی را باشد، پس آنگاه یک واحد تقطیع واژگانی می‌باشد.به عنوان مثال، یک «تخته سیاه» لزوما نباید سیاه باشد، ممکن است سبز باشند، بنابراین، این واژه تشکیل دهندهٔ یک واحد تقطیع واژگانی است.

---

[1]- Principle of bound morpheme

2- Principle of lexical integrity

ت- اصل استفادهٔ اصطلاحی[1]

اگر رشته‌ای از شکل‌های واژگانی به صورت اصطلاحی استفاده شوند، پس آنگاه به عنوان یک تک واحد تقطیع واژگانی به حساب می‌آیند(به عنوان مثال «دار فانی را وداع گفتن»به عنوان یک عبارت اصطلاحی استفاده می‌شود.)

ث- اصل نازایایی[2]

اگر یک کاندید واژه در ساختار نازایا باشد، پس یک واحد تقطیع واژگانی است. برای مثال، «پشتکار»، یک واژه نازایای فارسی است، چرا که علامتی با معنای «پشت» را نمی توان با هیچ علامتی با معنای محل جایگزین کرد که ترکیب به‌دست آمده در فارسی وجود داشته باشد.

۵-۲-۳ اصول از دیدگاه کاربردی

الف- اصل تواتر[3]

تواتر یک معیار اساسی برای تعیین درجهٔ واژه‌بندی[4]کاندیدواژه است. یک واژه یا رشته‌ای از واژگان بسیار مکرر یک واحد تقطیع واژگانی است.

ب- اصل گشتالت[5] (از علوم شناختی[6])

چیزها بیشتر به عنوان یک کل دیده می‌شوند. این اصل شواهدی خواهد بود برای برخی ترکیبات عبارتی به عنوان مدخل واژه‌نامه‌ای در قاموس حتی اگر آنها در ظاهراقلامی جداگانه به نظر برسند.

1- Principle of idiomatic use

2- Principle of non-productivity

3 - Principle of frequency

4- Lexicalization

5- Gestalt principle

6- Cognitive science

پ- اصل اعضای اصلی در دسته بندی (از زبان شناسی شناختی)[1]

با توجه به نظریه الگوی اصلی[2]در مورد قاموس روانی، اعضای اصلی در دسته بندی های از اعضای غیر اصلی برجسته تر هستند .انسان آنها را با دقت بیشتری در حافظه کوتاه مدت به یادآورده می‌آورد و به راحتی در حافظه بلند مدت حفظ شده و در دسترس می‌باشند .این اصل منطقی را فراهم می‌آورد برای شامل شدن برخی از ترکیبات عبارتیکه می توانند به عنوان نمونه‌های اصلی در یک الگوی تشکیل واژگانی بکار روند، به عنوان مثال «بادمجان بم» و «پشتکار» در زبان فارسی به ترتیب با الگوهای « بادمجان + شهر» و « محل + کار» در قاموس.

ت- اصل اقتصاد زبان[3]

اگر شمول واژهٔ کاندیدی در قاموس می‌تواند مشکل تجزیه و تحلیل زبانی آن‌را کاهش دهد، پس مناسب است یک واژه باشد .به عنوان مثال، «فتا»در فارسی مخفف («پلیس فضای تولید و تبادل اطلاعات نیروی انتظامی جمهوری اسلامی ایران») است، اگر رشته «فتا» در قاموس گنجانده شود، تشخیص آن به عنوان یک واحد تقطیع واژگانی آسان‌تر است.

۵-۳ اصل مدخل کاملِ واژگان

در اصل، همهواحدهای تقطیع واژگانی در استفاده مکرر در قاموس گنجانده می‌شوند. قاموس نیز باید با ایجاد واحدهای تقطیع واژگانی جدید پویا و سازگار باشد.

۵-۴ اصولی برای نتایج تقطیع واژگانی

الف- اصل دانه دانه بودن[4]

روند تقطیع واژگانی ممکن است ساختارهای داخلیدر واحدهای تقطیع واژگانی به‌دست آمده تولید کند تا تقطیع‌های جایگزین احتمالی برای دیگر کاربردهای مختلف را نشان دهد.

---

1 - Cognitive linguistic

2- Prototype theory

3- Principle of language economy

4- Principle of granularity

ب- اصل بیشینه کردن دامنه وندها

یک ستاک تمام وندهایی که به آن متصل می‌شودرا تابع خود می‌کند، و وندها بخشی از همان واحد تقطیع واژگانی را تشکیل می‌دهند، به عنوان مثال، « بابی‌تفاوتی» مفروض در یک متن: « با- بی- تفاوت-ی» به عنوان یک کل خود یک واحد تقطیع واژگانی است، با این‌حال اگر در یک متن با « بی‌تفاوت» یا «بی‌تفاوتی» مواجه شویم، خود به عنوان واحدهای تقطیع واژگانی به حساب می‌آیند.

پ- اصل بیشینه کردن دامنه ترکیب‌ها

اگر یک ترکیب در متنی یافت شود که شامل یک ترکیب دیگر باشد، با توجه به قاموس مرجع، آنگاه ترکیب بزرگتر خود به عنوان یک واحد تقطیع واژگانی به حساب می‌آید به همراه ترکیب کوتاه‌تر که احتمالاً به عنوان یک واحد تقطیع واژگانی داخلی دیگر نشان‌گذاری می‌شود. به عنوان مثال، رشتهٔ مفروض«سامانه عامل شبکه» در یک متن، تمام رشته به صورت یک کل به عنوان یک واحد تقطیع واژگانی به حساب می‌آید، که در آن ترکیب کوتاه «سامانه عامل» نامزد جایگزین واحد تقطیع واژگانی است.

ت- اصل تقطیع رشته های دیگر

هر رشته علائم، از جمله رشته های عددی و یا علائم خارجی و یا علائم نگارشی و هر ترکیبی از آن‌ها، می‌تواند یک واحد تقطیع واژگانی باشد در صورتی که احساس شود که در یک متن حامل برخی از توابع نحوی هستند.به عنوان مثال، در مثال«جنگ جهانی دوم در سال ۱۹۴۵ به پایان رسید»، رشتهٔ عددی «۱۹۴۵» یک واحد تقطیع واژگانی است .جملهٔ فرضی «¬ اولین علامت صامت در زبان کره‌ای است»، «¬»با توجه به این متن خاص که فاعل این جمله است،یک واحد تقطیع واژگانی می‌باشد.

۵-۵اصل پوشش کامل و ثبات در استفاده از این استاندارد

این استاندارد قصد دارد به شیوه‌ای سازگار مورد استفاده قرار گیرد تا هر متنیرا در هر زبان پوشش دهد.با این حال، همانگونه که در بخش دیگری از این استانداردشرح داده شد، برای زبان هایی خاص برخی از تغییرات مورد نیاز است .اهمیت متن در دانه دانه بودن و محدوده در تقطیع واژگانی نیز در این بخش‌ها بیشتر مورد بحث قرار خواهد گرفت و نشان داده می‌شود.

پیوست الف

(اطلاعاتی)
نمایش تقطیع واژگانی در XML[1]

این مثال از نمایش واحدهای تقطیع واژگانی در XML بر اساس ISO 24611 می‌باشد.

```
<?Xml version="1.0" encoding="UTF-8"?>
<maf addressing="xpointer">
  <seg xml:id="seg0">کارخانه و گوشت مرغ</seg>
  <token xml:id="tok1" target="#string-range(seg0,0,1)" />
  <token xml:id="tok2" target="#string-range(seg0,1,1)" />
  <token xml:id="tok3" target="#string-range(seg0,2,1)" />
  <token xml:id="tok4" target="#string-range(seg0,3,1)" />
  <token xml:id="tok5" target="#string-range(seg0,4,1)" />
  <wordForm lemma="کارخانه" tokens="#tok1 #tok2"
      entry="urn:lexicon:cn:کار- خانه" />
  <wordForm lemma="و" tokens="#tok3" entry="urn:lexicon:cn:و" />
  <wordForm lemma="گوشت مرغ" tokens="#tok4 #tok5" entry="urn:lexicon:cn:گوشت مرغ">
    <wordForm lemma="مرغ" tokens="#tok4" entry="urn:lexicon:cn:مرغ" />
    <wordForm lemma="گوشت" tokens="#tok5" entry="urn:lexicon:cn:گوشت" />
  </wordForm>
</maf>
```

1- Extensible Markup Language

کتابنامه

1. ISO 639-1:2002, Codes for the representation of names of languages — Part 1: Alpha-2 code
2. ISO 639-2:1998, Code for the representation of names of languages — Part 2: Alpha-3 code
3. ISO 639-3:2007, Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages
4. ISO 639-5:2008, Codes for the representation of names of languages — Part 6: Alpha-3 code for language families and groups
5. ISO 704, Terminology work — Principles and methods
6. ISO 860, Terminology work — Harmonization of concepts and terms
7. ISO 1087-1:2000, Terminology work — Vocabulary — Part 1: Theory and application
8. ISO 1087-2:2000, Terminology work — Vocabulary — Part 2: Computer applications
9. ISO 24611, Language resource management — Morpho-syntactic annotation framework)
10. ISO 24612, Language resource management — Linguistic annotation framework (LAF))
11. ISO 24613:2008, Language resource management — Lexical markup framework (LMF)
12. ISO 12620, Computer applications in terminology — Data categories
13. ISO 16642:2003, Computer applications in terminology — Terminological markup framework
14. ISO 30042:2008, Systems to manage terminology, knowledge and content — TermBase eXchange (TBX)
15. Britannica Online Encyclopedia, http://www.britannica.com
16. ALLEN, J., Natural Language Understanding, (1994) Addison Wesley
17. ARONOFF, M. and REES-MILLER, J., The Handbook of Linguistics. 2001, Blackwell
18. BIBER, D. et al., Corpus Linguistics. 1998, Cambridge University Press
19. BUSSMANN, H., Routledge Dictionary of Language and Linguistics. 1996, Routledge
20. CRYSTAL, D., The Cambridge Encyclopedia of Language. 1997, Cambridge University Press
21. JOHNSON, K. and JOHNSON, H., Encyclopedia Dictionary of Applied Linguistics: A Handbook for Language Teaching. 1999, Blackwell
22. KENNEDY, G., An Introduction to Corpus Linguistics. 1998, Addison Wesley Longman

23. MATTHEWS, P.H., Morphology. 1991, Cambridge University Press

24. PACKARD, J.L., The Morphology of Chinese: A Linguistic and Cognitive Approach. 2000, Cambridge University Press

25. POOLE, S.C., An Introduction to Linguistics, 1999, Macmillan

26. RICHARDS, J. et al., Longman Dictionary of Applied Linguistics. 1985, Longman

27. UNGERER, F. and SCHMIDT, H-J., An Introduction to Cognitive Linguistics. 1996, Addison Wesley Longman

28. Zhu, Dexi, Lecture on Grammar, 2003, Commercial Press (written in Chinese)