واژگان محاسبـــاتی: ساختار مرکـــزی در سیستم‌های پردازش زبان طبیعی

مهرنوش شمس فرد
دانشجوی دکترا

احمد عبداله زاده
استادیار

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

چکیده

امروزه مرکزیت واژگان در سیستم‌های پردازش زبان طبیعی مورد تأکید است و لذا طراحی و ساخت واژگان محاسباتی یکی از فعالیت‌های زیربنایی ساخت چنین سیستم‌هایی به شمار می‌آید. در این مقاله ابتدا به تعریف واژگان و اجزاء آن، بررسی عوامل افتراق واژگان ها و انواع حاصله می پردازیم. سپس بدنبال تشریح اصول نظری طراحی و ساخت واژگان ها، یك نمونه عملی را در یك زیر سیستم پردازش زبان فارسی ارائه خواهیم نمود. در این نمونه عملی، پارامترها و عوامل تأکید شده در مباحث نظری را مورد توجه قرار داده و مقدار دهی می نماییم.

کلمات کلیدی

واژگان محاسباتی، پردازش زبان طبیعی، پایگاه دانش، هستان شناسی، یادگیری ماشینی

Computational Lexicon: The Central Structure in Natural Language Processing Systems

M. Shamsfard
Ph. D Student

A. Abdollahzadeh Barforoush
Assisstant Professor

Computer Engineering Department,
Amir Kabir University of Technology

Abstract

Lexicons play a central role in natural language processing (NLP) systems, so design and implementing computational lexicons are of foundamental activities in development of such systems.

In this paper first, we define lexicon, introduce its parts, study the difference factors in lexicons and describe their various types. Then after discussing the theoretical aspects design and implementation of lexicons, we will study a practical instance in a Persian natural language processing system. In this system, we will consider the mentioned theretical parameters and factors and initiate them practically.

Key words

Computational Lexicon, Natural Language Processing, Knowledge Base, Ontology, Machine Learning.

مقدمه

امروزه مرکزیت واژگان در سیستم های پردازش زبان مورد تأکید است. تقریباً بیش از ۱۰ سال است که محققین کار بر روی مسئله ورودی نامحدود را آغاز کرده اند، با این هدف که سیستم های پردازش زبان طبیعی بتوانند هر نوع جمله ای را دریافت کنند. به این منظور لازم است سیستم بتواند بطور وسیع واژگان و دستور زبان موردنظر را بپوشاند. لذا طراحی ساختار واژگان، اکتساب دانش لغوی، روش های جستجو و ذخیره و بازیابی اطلاعات در واژگان و نحوه ارتباط واژگان با سایر اجزاء سیستم از مباحث مطرح در تحقیقات جهانی هستند.

دردهه ۸۰، منابع لغوی را دیکشنری ها (فرهنگ های لغات) تشکیل می دادند. فرهنگ های ماشین خوان[1] (MRDs) فقط شکل تایپ شده فرهنگ های چاپی بودند که برخی اطلاعات فرهنگ نویسی اضافی مانند تغییرات قلم، دستورات تایپی و نمادهای خاص را در برداشتند. به مرور زمان نقش کامپیوترها در آماده سازی این فرهنگ ها بیشتر شد و جمع آوری داده ها، انتخاب و ساخت تعاریف، مرتب سازی مدخل ها و آزمون کامل و سازگار بودن اطلاعات وارده توسط خود ماشین صورت گرفت.

در دهه ۹۰، مهمترین منابع لغوی، پیکره های زبانی[2] بوده اند. یک پیکره یک منبع اولیه[3] است و حاوی مجموعه استراتژیکی از مواد زبان طبیعی (تمام متن یا نمونه هایی از متون با جملات بی ربط ازیک یا چند زبان) می باشد که در شکل ماشین خوان ذخیره شده است. داده های این پیکره ها معمولاً پوشاننده حالات گفتاری و نوشتاری زبان است و محدوده وسیعی از زبان غیر فنی رایج (در زمان حال) را که به صورت عادی (و نه در شعر یا داستان های درام یا تخیلی) توسط بزرگسالان با لهجه استاندارد (و نه محلی)، استفاده می شود، دربر می گیرد.

امروزه واژگان های محاسباتی در گستره وسیعی از کاربردها بکار می آیند. از آن جمله می توان به سیستم های ویرایشگر، تصحیح املاء، تجزیه نحوی، تولید گرامر، ساخت و شناسایی گفتار، تولید و درک متن، طبقه بندی معنایی، رفع ابهام و ... اشاره نمود. برای ساخت یک واژگان محاسباتی دو راه اصلی وجود دارد ([۱]): راه اول بازسازی یک یا چند MRD و تبدیل آن به شکل مورد نیاز کاربرد است و راه دوم ساخت مستقیم واژگان از روی پیکره های زبانی می باشد. در بسیاری موارد از ترکیب این دو روش استفاده شده، ابتدا با استفاده از MRD ها واژگان اولیه ساخته و سپس بر اساس پیکره ها اطلاعات آن تصحیح و سازگار شده و یا تکمیل می گردد.

دراین مقاله ابتدا پس از ارائه تعریفی از واژگان، اجزاء و همچنین وجوه افتراق انواع مختلف واژگان ها رابررسی خواهیم کرد. سپس مقایسه ای میان واژگان های محاسباتی و زبانی انجام داده و در نهایت به طرح ساختار و ویژگی های نمونه ای از واژگان های محاسباتی در یک سیستم پردازش زبان فارسی خواهیم پرداخت.

۲ـ تعریف واژگان

هر فرهنگ[4] نگاشتی از «مجموعه نمادها» به «مجموعه تفاسیر» است. مجموعه نمادها می تواند شامل ترکیبات حروف الفباء ارقام، اصوات، تصاویر، اشکال و ... و یا دنباله ای از ترکیب آنها باشد و مجموعه تفاسیر بسته به نوع فرهنگ مورد نظر، معانی نمادها یا نمادهای دیگری متناظر با نمادهای داده شده را در بر می گیرد. به عبارت دیگر اگر فرض کنیم رابطه D یک رابطه فرهنگ سازی، مجموعه S مجموعه n عضوی نمادها و مجموعه I مجموعه m عضوی تفاسیر باشد، آنگاه خواهیم داشت:

$$D = \{(x,y) \mid x \in S, y \in I\}$$

$$S = \{x \mid x \text{ is a string of symbols; symbol} \in A\}$$

$$I = \{y \mid y \text{ is a string of interpretations; interpretation} \in B\}$$

$$A = \text{Alphabet} \mid \text{Digits} \mid \text{Phonemes} \mid \text{Graphemes} \mid \ldots$$

$$B = \text{Sememes} \mid \text{Alphabet} \mid \text{Digits} \mid \text{Phonemes} \mid \text{Graphemes} \mid \ldots$$

متداول ترین نوع فرهنگ، فرهنگ لغوی یا واژگان می باشد. در این نوع فرهنگ، مجموعه A مجموعه حروف الفبا و مجموعه B بسته به نوع کاربرد می تواند اجتماع مجموعه های واحدهای معنایی، الفبایی و صوتی باشد. واژگان در حقیقت انباره ای از اطلاعات در مورد زبان و ویژگی های اجتماعی آن است. یک واژگان اطلاعات لغوی مناسب و مرتبط (مانند طرز تلفظ و معنا) را برای واحدهای مناسب و مربوط لغوی[5] (مانند لغات) ارائه می کند [۲]. معروف ترین شکل واحد لغوی کلمه است. اشکال دیگر آن می توانند قسمتی از کلمه یا ترکیبات کلمات باشند که در مورد آنها در بخش های بعد صحبت خواهیم کرد. هر واحد لغوی مجموعه ای از صور (تکواژها[6])ست که مجموعه ای از محتواها (واحد های معنا[7]) را بازنمایی می کند. این صور با مجموعه ای

گرافیکی مثل حروف و یا مجموعه ای صوتی مثل صداها محقق می شوند. ارتباط بین صورت و معنا برای یک واحد لغوی یک ارتباط n به m است که n می تواند بزرگتر، کوچکتر و یا مساوی m باشد. به بیان دیگر:

$$L=\{(x, y) \mid x \text{ is a set of morphemes, } y \text{ is a set of sememes}\}$$

انواع دیگری از فرهنگ ها نیز وجود دارند که مورد توجه ما در این مقاله نیستند. مانند فرهنگ انتساب اشکال گرافیکی به حروف (مثلاً جهت بازشناسی متن)، انتساب کلمات به تصاویر (مثلاً برای استفاده کودکان)، انتساب اصوات به حروف (مثلاً برای تشخیص گفتار) و ...

۳ـ اجزاء واژگان

هر واژگان دارای دو جزء اصلی است:

- ساختار کلی
- رویه جستجو

ساختار کلی واژگان حاوی همه اطلاعاتی است که در واژگان جای گرفته است و خود از کنار هم قرار گرفتن اجزاء مستقلی بنام «مدخل[8]» تشکیل شده است. یک مدخل، ساختاری است که اطلاعات مرتبط با یک واحد لغوی را در خود جای می دهد. این مدخل ها طوری مرتب شده اند که با یک رویه جستجو به سادگی قابل دستیابی اند. به بیان دیگر رویه جستجو امکان دستیابی سریع و آسان به مدخل های مرتب شده را فراهم می کند. نمونه ای از یک رویه جستجو می تواند ترتیب الفبایی باشد که بر اساس شکل نوشتاری واحد لغوی (و نه معنای آن) صورت می گیرد.

تصمیماتی نظیر تعیین اطلاعات مندرج در هر مدخل، نحوه کنار هم نشستن مدخل ها، چگونگی تخصیص واحدهای لغوی به مداخل و انتخاب رویه جستجوی مناسب بر عهده طراح واژگان است که بسته به نوع واژگان و کاربرد آن تعیین می نماید. اتخاذ تصمیمات متفاوت در هر زمینه می تواند منجر به ایجاد واژگان های مختلف گردد. نمونه هایی از این تفاوت ها در بخش بعد بررسی شده اند.

٤ـ انواع واژگان

واژگان ها بر اساس عوامل مختلفی طبقه بندی می شوند. از جمله این عوامل می توان به نوع و حدود اطلاعات مندرج در آنها و نحوه سازمان دهی این اطلاعات اشاره داشت. در دو بخش بعد عوامل طبقه بندی را بر اساس محتوا و ساختار واژگان خواهیم دید.

٤ـ۱ـ تنوع در محتوا

محدوده و تنوع اطلاعات موجود در یک واژگان به عوامل مختلفی وابسته خواهد بود که الزاماً مستقل از یکدیگر نیستند. برخی از این عوامل در زیر آورده شده است.

الف ـ درجه پوشا بودن واژگان: این عامل نشان دهنده وسعت دانشی است که واژگان تحت پوشش قرار می دهد. یک واژگان از این جهت می تواند در دو سطح عمل کند: [۳]

الف ـ ۱ـ لغت نامه: در چنین واژگانی هر لغت با کمک دانش زبانی موجود تعریف می شود. به عبارت دیگر لغات با کمک ارتباطاتشان با لغات دیگر مانند ترادف، تضاد و در بر گیری معرفی می شوند. چنین واژگانی دانش جهان را که برای عملیات استنتاج و استدلال مورد نیاز است، در بر نمی گیرد.

الف ـ ۲ـ دائرة المعارف: دایرة المعارف ها از جهت مجموعه تفاسیر با لغت نامه ها متفاوتند. آنها بخش عظیمی از دانش جهان را در بر دارند. این دانش انجام استنتاج و استخراج دانش جدید را ممکن می سازد.

ب ـ گروه مخاطبین واژگان: واژگان بسته به این که برای چه گروه از مخاطبین و به چه منظور طراحی می شود دارای اطلاعات متفاوتی خواهد بود. دو نمونه از عوامل دسته بندی مخاطبین در زیر آمده است:

ب ـ ۱ـ سن: رده سنی مخاطب در انتخاب نوع توضیحات، رسانه مورد استفاده و ابزارهای بکار گرفته شده در تفسیر کلمات مؤثر است. به عنوان نمونه واژگان طراحی شده برای کودکان به زبان ساده تر، با توضیحات بسیط تر و اکثراً همراه با تصاویر و اشکال متنوع ارائه می شود. در چنین واژگانی مجموعه تفاسیر حاوی زیر مجموعه ای از اشکال و تصاویر است.

ب ـ ۲ـ تخصص: برخی واژگان ها تخصصی هستند و با در برگیری لغات و اصطلاحات فنی یک رشته، برای متخصصین آن رشته طراحی می شوند، در حالیکه برخی دیگر عمومی هستند و کلمات و لغات معمول زبان که افراد عادی از آنها استفاده فرهنگ ها هم در مجموعه نمادهای انتخابی و هم در مجموعه تفاسیر ارائه شده قابل مشاهده است.

ج ـ تعداد زبان های تحت پوشش: برخی واژگان ها تک زبانه هستند و برخی دیگر چند زبانه.

ج ـ ۱ـ تک زبانه: واژگان های تک زبانه لغات و اصطلاحات یک زبان را به همان زبان معنی کرده و یا در مورد آنها توضیحاتی ارائه می کنند. فرهنگ های لغات و دائرة المعارف ها نمونه هایی از چنین واژگان هایی هستند.

ج ـ ۲ـ چند زبانه: واژگان های چند زبانه لغات و عبارات مترادف در دو یا چند زبان را معرفی می نمایند. این فرهنگ ها بیشتر برای ترجمه بکار می روند. در چنین حالتی مجموعه نمادها از لغات زبان مبدا و مجموعه تفاسیر از لغات زبان

مقصد انتخاب می شوند. متداولترین این فرهنگ ها دیکشنری های دوزبانه هستند.

د ـ کاربرد واژگان: یکی دیگر از عوامل تنوع در واژگان نوع کاربرد آن است. اطلاعات مندرج در واژگان بسته به کاربرد و استفاده مورد نظر، متفاوت خواهد بود. به عنوان نمونه چند کاربرد در زیر آورده شده است:

د ـ ۱ ـ درک: واژگان مورد استفاده در درک متون لازم است از لحاظ معنایی و کاربردشناسی بسیار غنی باشد. این نوع واژگان بیشتر در سطح دائرة المعارف عمل می کند و وظیفه آن نگاشت لغات و عبارات زبان به مفاهیم و معانی جهان واقعی است. در این حالت مجموعه تفاسیر، اجتماعی از مجموعه ویژگی های صرفی، نحوی و کاربرد شناسی است.

د ـ ۲ ـ تولید: واژگان مورد نیاز از جهت ویژگی شبیه به مورد قبل است با این تفاوت که در جهت معکوس وبه منظور نگاشت معانی و مفاهیم به لغات و عبارات زبان بکار می رود. به عبارت دیگر در این واژگان (نسبت به مورد د ـ ۱)، جای مجموعه نمادها و مجموعه تفاسیر عوض شده است.

د ـ ۳ ـ ترجمه: واژگان مورد نیاز چند زبانه است و میزان پوشا بودن آن بستگی به سطح ترجمه مورد نظر دارد. در این حالت مجموعه نمادها و تفاسیر از الفبای دو یا چند زبان متفاوت گرفته می شوند.

ه ـ کاربر واژگان: عامل مؤثر دیگر در تعیین محتوای واژگان، کاربر آن است. اطلاعات عمومی و قبلی کاربر در هنگام استفاده از واژگان، تعیین کننده سطح اطلاعاتی است که لازم و کافی است در واژگان جای گیرد.

ه ـ ۱ ـ انسان: در صورتی که کاربر انسان باشد، سطح وسیعی از دانش عرفی و عمومی را داراست. لذا می توان با استفاده از اصل کم کوشی، از درج مجدد اطلاعاتی که فرض می شود مخاطب دارد خودداری نمود. در این حالت انسان با دانش عمومی خود دانش مندرج در واژگان را تفسیر و خلاء های اطلاعاتی آن را بسهولت پر می کند. در چنین واژگانی بخشی از مجموعه تفاسیر و همچنین زیر مجموعه ای از مجموعه نمادها، دانسته فرض شده و از واژگان حذف می شوند.

ه ـ ۲ ـ ماشین: اگر کاربرد واژگان ماشین یا سیستم کامپیوتری باشد، لازم است دانش عرفی و عمومی مورد نیاز نیز در واژگان تعبیه شود. لذا حجم و محتوای دانش مندرج در واژگان محاسباتی که کامپیوتر از آن استفاده می کند با واژگان زبان شناسی مورد استفاده انسان ها متفاوت است. در واژگان محاسباتی، مجموعه تفاسیر و نمادها بزرگتر از واژگان زبان شناسی بوده، از جهت صورت بندی نیز بدلیل محدودیت های قابلیت خوانایی توسط ماشین، متفاوت خواهند بود.

۲ـ۴ـ تنوع در ساختار و نحوه انجام عملیات

سازمان دهی اجزاء یک واژگان بر اساس پارامترهای مختلف و به صور متفاوتی انجام می شود. این تفاوت در ساختار منجر به ایجاد تنوع در واژگان ها خواهد شد. برای سازمان دهی واژگان، طراحی ساختار و الگوریتم های انجام عملیات اصلی در واژگان، لازم است طراح به سؤالاتی پاسخ دهد. مجموعه پاسخ های طراح به این سؤالات، ساختار و اعمال لازم برای واژگان را تعیین خواهد کرد. نمونه هایی از موارد سؤال و برخی پاسخ های نوعی به آنها در زیر آورده شده است:

الف ـ نحوه برخورد با کلماتی با چند معنی

در هر زبانی تعدادی از کلمات دارای بیش از یک معنی هستند. نحوه سازمان دهی این معانی برای کلمه یکی از مسائلی است که طراح باید به آن پاسخ دهد. به این منظور دو امکان بررسی می شود:

الف ـ ۱ ـ آوردن همه معانی تحت یک مدخل: در این روش برای هر کلمه با هر تعداد معنی تنها یک مدخل در نظر گرفته می شود و تمامی معانی کلمه در همان مدخل جای می گیرند. در این حالت واژگان تابعی از مجموعه نمادها به مجموعه تفاسیر است. به عنوان مثال در چنین واژگانی برای کلمه «شیر» تنها یک مدخل وجود دارد که در مقابل آن معانی چون: حیوانی درنده، وسیله خروج آب یا مایعات از لوله و مایع خوراکی مترشحه از سینه پستانداران، درج می شود.

الف ـ ۲ ـ استفاده از مدخل های مختلف: در این روش معانی مختلف کلمه بر اساس معیارهای مختلف در مدخل های جداگانه جای داده می شوند. این معیارها می توانند داشتن ریشه کلمه، منشأ تاریخی، ویژگی های نحوی یا کاربرد شناسی متفاوت باشند. در این حالت واژگان دیگر یک تابع از مجموعه نمادها به مجموعه تفاسیر نیست بلکه رابطه ای است که عضو اول آن لزوماً یکتا نیست. در این واژگان برای همان کلمه «شیر» سه مدخل جداگانه در نظر گرفته می شود که هر یک حاوی یکی از معانی کلمه است.

الف ـ ۳ ـ حالت ترکیبی: این حالت که معمولاً متداول تر از حالات قبل است ترکیبی از حالات الف ـ ۱ و الف ـ ۲ است. در این واژگان معیارهای تفکیکی محدود شده و تنها به ازای برخی معیارها مثلاً منشاء تاریخی کلمه مدخل ها مجزا و سایر تفاوت ها تحت یک مدخل بیان می شوند.

ب ـ نحوه نگاشت مدخل ها به واحدهای لغوی

ب ـ ۱ ـ یک به یک: در این حالت هر واحد لغوی با هر تعداد معنی یک و فقط یک مدخل را به خود اختصاص می دهد (مشابه حالت الف ـ ۱) و هر مدخل نیز تنها به یک واحد لغوی تخصیص می یابد.

ب ـ ۲ـ یک به چند: در این حالت یک مدخل به چند واحد لغوی تخصیص می یابد. در چنین شرایطی یک واحد لغوی مدخل اصلی محسوب شده و چندین واحد دیگر تحت این مدخل به عنوان زیر مدخل یا وابسته های آن قرار می گیرند.

ب ـ ۳ـ چند به یک: در این حالت چند مدخل برای یک واحد لغوی صرف می شوند. این حالت وقتی اتفاق می افتد که یک واحد لغوی با چند معنی داریم که برای هر معنی یک مدخل در نظر گرفته ایم (مشابه حالات الف ـ۲ و الف ـ۳).

ج ـ نحوه تعیین واحدهای لغوی

ج ـ ۱ـ کلمات: ساده ترین و بدیهی ترین شکل واحدهای لغوی کلمات هستند. مانند خانه، آوردن و ...

ج ـ ۲ـ زیر کلمات: زیر کلمات تکواژهایی هستند که به ساخت تصریف ها یا مشتقات کمک می کنند مانند پیشوندها، پسوندها و میانوندها.

ج ـ ۳ـ زبر کلمات: ترکیباتی که از بیش از یک کلمه درست شده اند نیز می توانند واحد لغوی باشند. این ترکیبات دردو گروه اصلی جای دارند:

ـ واحدهای ترکیبی و آغازگرا: هر دو نوع این واحدها اجزائی از چند کلمه دارند با این تفاوت که در واحدهای ترکیبی این اجزاء از هر کجای کلمه می توانند انتخاب شوند ولی در واحدهای آغازگرا این اجزاء از حروف اول کلمات برگرفته می شوند. به عنوان مثال می توان به واحد «مشاصد» (ترکیبی از حروف کلمات مشاورین صنایع دفاع) به عنوان یک واحد ترکیبی و به واحد «نزاجا» (ترکیبی از حروف آغازین کلمات نیروی زمینی ارتش جمهوری اسلامی ایران) به عنوان یک واحد آغازگرا اشاره نمود.

ـ اصطلاحات و عبارات مرکب: این واحدها ترکیبی از چند کلمه کامل هستند. راه تشخیص این واحدها از دنباله ای از کلمات متوالی این است که در این حالت معنی کل عبارت با مجموع معانی کلمات تشکیل دهنده آن متفاوت است. مثل واحد «سر دواندن» یا اصطلاحات و ضرب المثل ها.

ج ـ ۴ـ موارد خاص شامل

اسامی خاص: یک گروه مهم از مدخل ها را اسامی خاص تشکیل می دهند. این مدخل ها ممکن است به اسامی اشخاص (مثل ارسطو یا فردوسی)، مکان ها (مانند حافظیه یا خزر) یا اصطلاحات عملی و فنی (مانند پنی سیلین یا رایانه) تخصیص یابند. برخی واژگان ها برای محدود نمودن مجموعه لغات خود این مدخل ها را حذف می نمایند.

صورت های روزمره: این عبارات واحدهای غیر لغوی هستند که کاربرد عملی دارند یا ترجمه واحد لغوی در زبان دیگر می باشند (مانند ترجمه استعارات).

د ـ یگانگی ساختار واژگان

د ـ ۱ـ تک ساختاری: در این روش همه واحدها در یک ساختار ریخته می شوند. شکل واژگان برای همه مدخل ها یکسان بوده، واژگان یکدست و بدون ضمائم ارائه می شود.

د ـ ۲ـ چند ساختاری: در روش چند ساختاری در کنار ساختار اصلی واژگان، از ضمائم برای معرفی واحدهای خاص مانند اسامی خاص و مخفف ها استفاده می گردد. ساختار این ضمائم ونوع اطلاعات مندرج در آنها می تواند به کلی با ساختار اصلی واژگان متفاوت باشد.

ه ـ تعیین مدخل های اصلی و زیر مدخل ها

متداول ترین ترتیب نزولی بر اساس احتمال مدخل بودن و ترتیب صعودی بر اساس احتمال زیر مدخل بودن در انگلیسی به شکل زیر است: تکواژها (مثل home)، واحدهای ترکیبی (مثل smog)، واحدهای آغازگرا (مثل VIP)، ترکیبات اسمی با نوشتار بسته (بدون فاصله) (مثلBlack-bird)، ترکیبات اسمی با نوشتار باز (با فاصله) (مثل night owl)، ترکیبات فعلی (افعال مرکب) (مثل give up)، اصطلاحات غیر فعلی (مثل at all)، اصطلاحات فعلی (مثل kick the bucket) و مشتقات با معنای شفاف (که معنای آنها از معنای اجزائشان قابل درک است) شامل: ترکیبات پسوند (زیر قسمت اصلی کلمه) (مثل lameness تحت مدخل lame)، و پیشوندی (به ترتیب الفبایی زیر پیشوند و یا به ترتیب غیر الفبایی زیر قسمت اصلی کلمه با ارجاع به محل اصلی الفبایی آنها) (مثل pre-war تحت مدخل ـpre).

و ـ ترتیب واحدهای با تشابهات صوری

در هر زبانی کلماتی وجود دارند که از لحاظ شکل نوشتاری بهم شبیه هستند ولی معانی متفاوتی دارند. این تشابهات صوری بر دو گونه اند: دسته اول کلمات با چند معنی هستند. در این حالت نه تنها شکل نوشتاری بلکه تلفظ کلمه نیز در معانی مختلف ثابت می ماند. (مانند کلمه شیر با سه معنی متفاوت) اما دسته دوم کلمات با شکل نوشتاری یکسان و تلفظ متفاوت هستند (مانند کلمه حکم با تلفظ های hokm, hakam). هر دو دسته در گروه تشابهات صوری قرار می گیرند. نکته قابل توجه طراح در مورد این تشابهات تعیین ترتیب وقوع این واحدها در واژگانی است که بر اساس شکل صوری کلمات مرتب شده است. به این منظور معیارهای مرتب سازی زیر پیشنهاد می گردد:

و ـ ۱ـ به ترتیب تاریخی: در این روش واحدهای مشابه بر اساس منشاء تاریخیشان (معمولاً از قدیم به جدید) مرتب می شوند.

و ـ ۲ـ به ترتیب فراوانی: در این روش واحدهای پرمصرف تر پیش از واحدهای مشابه کم مصرف آورده می شوند. به عبارت دیگر بر اساس مطالعات آماری، واحدها بر مبنای

بیشترین رخداد معانی مرتب می شوند.

و ـ ۳ـ به ترتیب الفبایی اجزاء کلامشان: راه دیگر مرتب سازی واحدهای مشابه در نظر گرفتن ترتیب الفبایی طبقه نحوی آنهاست. مثلاً اسم و صفت پیش از فعل و قید پس از آن ظاهر می شود.

و ـ ۴ـ استفاده از الگوریتم های خاص مثلاً

ـ قائل شدن ترتیب بین حروف کوچک و حروف بزرگ

ـ قائل شدن ترتیب بین واحدهای بسته (بی فاصله) و واحدهای باز (فاصله دار)

ـ قائل شدن ترتیب بین حروف خاص مثل خط فاصله و آپاستروف

ز ـ اهم اطلاعات ذخیره شده برای هر مدخل (واحد لغوی)

یکی از مهمترین مسائلی که طراح باید در مورد آن تصمیم بگیرد، انتخاب اطلاعاتی است که لازم است برای هر واحد لغوی ذخیره شود. در زیر برخی از این اطلاعات آورد ه شده است. واژگان های مختلف بسته به نوع و کاربردشان ممکن است برخی یا همه این اطلاعات را در خود جای دهند.

ز ـ ۱ـ شکل واحد (با گونه های نوشتاری مختلف): همان رسم الخط واحد لغوی است که معمولاً به عنوان اندیس مدخل هم استفاده می شود. مانند: برد، کتاب، ...

ز ـ ۲ـ اطلاعات آوایی (تلفظ و الگوهای تأکید): مانند: (برد: bord)، (کتاب: keta'b)

ز ـ ۳ـ طبقه بندی و زیر طبقه بندی های نحوی شامل:

ـ اجزاء کلام (اسم، فعل، صفت و غیره): مثال: (برد: فعل)، (کتاب: اسم)

ـ اطلاعات دستوری بیشتر (شمارش پذیری، تعدی و غیره): مثال: (برد: ماضی، متعددی)، (کتاب: مفرد، شمارش پذیر)

ز ـ ۴ـ گونه های تصریفی (جمع، ماضی، اشکال بی قاعده و ...): مثال: (برد: (مضارع: برد: barad)، (کتاب: (جمع: کتب: kotob)))

ز ـ ۵ـ گونه های اشتقاقی و ترکیبی: مثال: (برد: (برد و باخت: نتیجه مسابقه)، ...)، (کتاب: (مکتوب: نوشته شده)، (مکتب: ۱ـ محل نوشتن، ۲ـ نوع نگرش خاص)، (سر کتاب باز کردن: دعا نویسی)...)

ز ـ ۶ـ معرف معنی (ها) با ارائه تعاریف یا ارجاع به مترادف ها: مثال: (برد: ۱ـ انتقال داد، ۲ـ پیروز شد)، (کتاب: مجموعه چاپی از اوراق نوشته شده)

ز ـ ۷ـ مثال هایی از طرق مختلف استفاده از کلمه: (برد: ۱ـ علی کتاب را به مدرسه برد. ۲ـ علی مسابقه را برد.، علی شرط را برد.)، (کتاب: کتاب را خواندم.)

ز ـ ۸ـ موارد استفاده یا استفاده خاص: مثال: (برد: از میان برد: او آبروی مرا برد)

ز ـ ۹ـ اطلاعات در مورد **منشأ** واحد: در مورد ریشه کلمه و پیدایش و تغییر شکل آن در طول زمان تا رسیدن به شکل فعلی صحبت می کند.

ز ـ ۱۰ـ اطلاعات **جانشینی**[9] مانند مترادف، متضاد، وارون، فراگیر و ...: مثال: (برد: (مترادف: انتقال داد)، (فراگیر: انتقال فیزیکی)، (وارون: آورد))، (کتاب: (فراگیر: شیء فیزیکی ـ بی جان))

ز ـ ۱۱ـ اطلاعات **هم نشینی**[10] مانند عبارات و ترکیباتی که این کلمه معمولاً در آنها واقع می شود و یا محدودیت های گزینشی آن: مثال: (برد: (چیزی یا کسی را بردن: فاعل: متحرک، مفعول: قابل حمل))، (کتاب: (کتاب خواندن، کتاب نوشتن: فاعل: انسان، باسواد))

ز ـ ۱۲ـ اطلاعات **تشابهی**[11] (قیاس پذیر) حاوی اطلاعاتی در مورد حوزه لغوی که این واحد را شامل می شود (مانند استفاده از see also)

ز ـ ۱۳ـ اطلاعات **نظام مندی**[12] زبان مانند تعلق واحد به هسته زبان یا قرضی بودن آن، تعلق به زمان یا مکان و یا قلمروی تخصصی خاص، و یا رسمی یا غیر رسمی بودن واحد: مثال: (برد: متعلق به هسته زبان، عام)

ح ـ تکنیک های انجام توضیحات:

ح ـ ۱ـ نمایش: با جدول، شکل یا دیاگرام

ح ـ ۲ـ مثال آوری: ارائه مثال های مختلف از کاربرد واحد لغوی و نحوه ترکیب آن با واحدهای دیگر

ح ـ ۳ـ بسط: این نوع توضیح برای مخفف ها یا واحدهای ترکیبی و آغازگرا بیشتر بکار می رود و شکل کامل عبارتی را که واحد لغوی از آن گرفته شده بیان می کند. گاهی بیان شکل کامل عبارت اطلاعات کافی در اختیار کاربر می گذارد و نیازی به توضیح بیشتر نیست مانند توضیح واحد نزاجا به شکل نیروی زمینی ارتش جمهوری اسلامی ایران. اما بعضی مواقع بسط واحد به تنهایی اطلاع زیادی در مورد آن به ما نمی دهد مثلاً بسط واحد ناتو (NATO) به شکل "North Atlantic Treaty Organization" است که اطلاعی در مورد این سازمان، اعضاء عملکرد و اهداف آن نمی دهد.

ح ـ ۴ـ بحث: توضیح ساده استفاده روزانه از این واحد

ح ـ ۵ـ تعریف: شکستن واحد به اجزاء سازنده و ترکیب مجدد آنها طوری که محتوای واحد ساخته شده، معنی یا تعریف واحد مورد نظر باشد. تعریف ممکن است به صورت تعریف با مترادف، تعریف تحلیلی و یا تعریف فرمولی باشد.

۵ـ مقایسه واژگان محاسباتی و واژگان زبان

واژگان محاسباتی و واژگان زبان از جنبه های مختلف با هم متفاوتند. مهمترین و واضح ترین تفاوت در انتظاری است

که از آنها داریم. در واژگان زبان (اعم از شکل چاپی و یا تایپ شده آن) کاربر انسان است. لذا اطلاعات مندرج به زبان طبیعی و با صورت بندی قابل فهم برای انسان در واژگان ذخیره می شوند. درمقابل، کاربر اصلی واژگان محاسباتی، ماشین است. لذا اطلاعات با فرمت قابل خواندن توسط ماشین و عموماً به یک زبان بازنمایی میانی نگهداری می شود. در این سیستم ها درصورت نیاز، واسط های کاربری تعبیه می شود که اطلاعات مورد نیاز کاربر انسانی را تغییر شکل داده به صورت قابل فهم توسط انسان به او ارائه نماید.

به دلیل اختلاف فوق واژگان های محاسباتی و زبانی هم از جهت محتوایی و هم از جهت ساختاری با هم متفاوتند. اختلاف محتوایی واژگان محاسباتی و زبان شناسی ناشی از اختلاف در کاربر آنهاست. درجایی که کاربر واژگان انسان است، حجم زیادی از دانش عرفی و عمومی بشر، با فرض علم خواننده نسبت به آن صراحتاً آورده نمی شود. این بخش از دانش که برای انسان دانسته فرض شده، برای ماشین ناشناخته است. لذا در واژگان های محاسباتی لازم است صراحتاً بیان گردد. این تفاوت بیشتر در دانش های معناشناسی و کاربردشناسی نمود دارد.

اما اختلاف ساختاری واژگان های زبان شناسی و محاسباتی از چند جهت قابل بررسی است:

اولاً واژگان محاسباتی بدلیل دارا بودن حجم بیشتری از دانش، نیاز به ساختار بزرگ تر و پیچیده تری جهت ذخیره سازی این دانش دارد.

ثانیاً در واژگان محاسباتی برای دستیابی سریع و چند جهته به اطلاعات، اتصالات و ارتباطات بیشتری (نسبت به واژگان زبان شناسی) تعبیه شده اند. به عنوان نمونه در یک واژگان غیر محاسباتی که از سلسله مراتب ارث بری برای نمایش معنا استفاده می کند، معمولاً هر گره به گره پدرش اشاره دارد ولی به فرزندان و برادرانش اشاره نمی کند. مثلاً در تعریف کلمه «درخت» گفته می شود که «درخت» یک نوع «گیاه» است و در تعریف کلمه «گیاه» ارتباط فرزندی با گره «موجود زنده» قید می گردد. اما درتعریف گره «درخت» به فرزندان دیگر گره پدر (برادران) مانند «گل» اشاره نمی شود. همچنین ارتباط معکوس میان گره پدر و فرزند برقرار نیست. مثلاً از همان گره «درخت» راهی برای رسیدن به گره های «راش» و «افرا» وجود ندارد. در حقیقت این اطلاعات مفقوده اطلاعات ساختاری هستند. چرا که به هر حال از روی واژگان غیر محاسباتی نیز می شود آنها را استخراج کرد ولی کار بسیار مشکل و زمان بری خواهد بود. در حالیکه با افزایش اطلاعات ساختاری و ارتباطات میان اجزاء واژگان، سرعت دستیابی به این اطلاعات را افزایش داده ایم. نمونه دیگر این ارتباطات اضافه می تواند تعیین معنی ای از گره پدر باشد که گره جاری فرزند آن است. در بسیاری موارد گره پدر دارای بیش از یک معنی است. در واژگان های غیر محاسباتی، در چنین حالتی گره های فرزند تنها به گره پدر خود به شکل عام اشاره دارند و معنی مورد نظر خود از گره پدر را تعیین نمی کنند. در بسیاری از واژگان های محاسباتی (و از جمله در وردنت [۵]) برای افزایش سرعت استخراج اطلاعات چنین ارتباطاتی بوجود می آیند.

ثالثاً بدلیل استفاده ماشین ازواژگان محاسباتی، لازم است این واژگان به صورت کاملاً ساخت یافته و با ساختار از پیش تعیین شده ارائه گردد. به همین دلیل واژگان های مختلف، جهت تطابق ساختارهای لازم برای نمایش همه طبقات نحوی و معنایی زبان، تدابیر متفاوتی اندیشیده اند. مثلاً وردنت از واژگان چندساختاری بهره گرفته، ساختارهای متفاوتی برای بازنمایی اطلاعات در مورد اعضاء طبقات نحوی مختلف درنظر می گیرد. در برخی واژگان های محاسباتی دیگر، تک ساختاری ترجیح داده شده و ساختمان داده های مورد نیاز به شکل پویا و قابل انطباق تعریف می شوند. این در حالی است که اکثر واژگان های غیر محاسباتی تک ساختاری هستند و از رویه ای مسطح برای معرفی همه واحدهای لغوی استفاده می کنند (یا حداقل اگر چندساختاری باشند، معیار تقسیم بندی کاملاً متفاوت با واژگان های محاسباتی دارند). در چنین واژگانی اصراری برای انطباق ساختاری و یکسان سازی اجزاء تعریف همه واحدهای لغوی وجود ندارد.

تفاوت دیگری که هم از جهت ساختاری و هم از جهت محتوایی قابل بررسی است، تفاوت در نحوه ارائه معانی در دو نوع واژگان است. در واژگان زبان شناسی، معمولاً هر واحد لغوی با معرفی مترادف ها، بیان توضیحات و یا ارائه تصاویر معنی می شود. استفاده از کلمات ساده تر برای معرفی کلمات و عبارات دیگر بدلیل احتمال زیاد شناخت کاربر از کلمات ساده، معمول است. در چنین واژگانی گاهی بدلیل استفاده ازکلمات برای تعریف کلمات دیگر، با حلقه مواجه می شویم. اما در واژگان های محاسباتی نحوه معرفی معنا متفاوت است. در بعضی سیستم هایی که از واژگان بهره می گیرند (مانند مایکروکاسموس [۶])، برای نمایش معنا از ساختار دیگری به نام هستان شناسی استفاده می شود. هستان شناسی دانش مفاهیم جهان را درون خود سازمان دهی می نماید. معنی نمودن هر واحد لغوی در واژگان با معرفی جایگاه آن در سلسله مراتب هستان شناسی انجام می پذیرد. به این ترتیب کلمات مترادف آنهایی هستند که به مفاهیم و معانی یکسان در هستان شناسی ارجاع داشته

باشند. با این روش اگر چه مشکل معرفی مفاهیم در هستان شناسی اضافه می شود، ولی امکان بروز حلقه از میان رفته کار معرفی معنا در واژگان تسهیل می گردد. دربرخی واژگان های دیگر نیز که از هستان شناسی بهره نمی گیرند (مانند وردنت) معرفی معنی واحد لغوی با بیان ارتباطات مختلف آن با واحدهای دیگر انجام می گیرد. وسعت این ارتباطات منجر به حذف حلقه و رفع ابهام در واژگان می گردد.

۶ـ طرح یک نمونه عملی: واژگان در پروژهٔ هستی

۶ـ۱ـ معرفی هستی

هستی [۷و۸] سیستم ساخت خودکار هستان شناسی[13] بر اساس پردازش زبان فارسی است. این سیستم از مدل یک هسته یادگیر و تکاملی برای ساخت پایگاه دانشی برای مفاهیم استفاده می کند. هسته اولیه دارای مفاهیم و عملگرهای بنیادی است که برای اکتساب خودکار دانش مورد نیازند. به عبارت دیگر هسته حاوی فرادانش لازم جهت ساخت انواع هستان شناسی ها برروی آن، با توجه به دانش دریافتی از محیط از طریق جملات ساده زبان فارسی است. سیستم با شروع بکار با این هسته، با تراکنش هایی که به زبان طبیعی با محیط خود دارد، لایه های بعدی هستان شناسی را حول هسته اولیه می سازد. تنها وسیله ارتباطی هستی با محیط، متون نوشتاری زبان فارسی و هدف آن ساخت هستان شناسی براساس اطلاعات دریافتی از متون ورودی است. تأمین این اطلاعات وظیفه «مأمور پردازش زبان طبیعی» است که واژگان یکی از اجزاء اصلی در ساختار پایگاه دانش آن است. شکل (۱) ساختار کلی سیستم هستی را نشان می دهد.

۶ـ۲ـ ویژگی های واژگان در هستی

ساختار واژگان برای حفظ دانش لغات در سیستم طراحی شده است. این ساختار در ابتدا حاوی بخشی از دانش اولیه در مورد لغات است و سپس در طول حیات سیستم[14] دانش های جدید توسط سیستم، اخذ [۹] و در آن درج می شود. به عبارت دیگر واژگان، ساختاری با محتوای پویاست. برای طراحی واژگان لازم است نوع واژگان، اطلاعات مندرج در آن، ساختار مورد نیاز و نحوه دستیابی، ذخیره سازی، بازیابی و بهنگام سازی این اطلاعات مشخص گردد. در ادامه این بخش به بررسی مشخصات مذکور در طراحی واژگان در سیستم «هستی» می پردازیم.

الف ـ نوع واژگان

ـ درجه پوشا بودن: واژگان این سیستم خود به تنهایی به صورت یک لغت نامه عمل می کند. اما از آنجا که در ارتباط با هستان شناسی قرار می گیرد و از طریق مدخل های آن می توان به دانش گسترده مفاهیم و ارتباطات جهان دسترسی داشت، در سطح دائرة المعارف مورد استفاده قرار می گیرد.

ـ گروه مخاطبین: مخاطب این واژگان در حقیقت سیستمی است که قرار است از متون در حد کتاب فارسی کلاس اول دبستان، اطلاعاتی را استخراج کند. لذا در برگیرنده واژه هایی در حد همین کتاب است و واژه ها و عبارات تخصصی در آن جای ندارند.

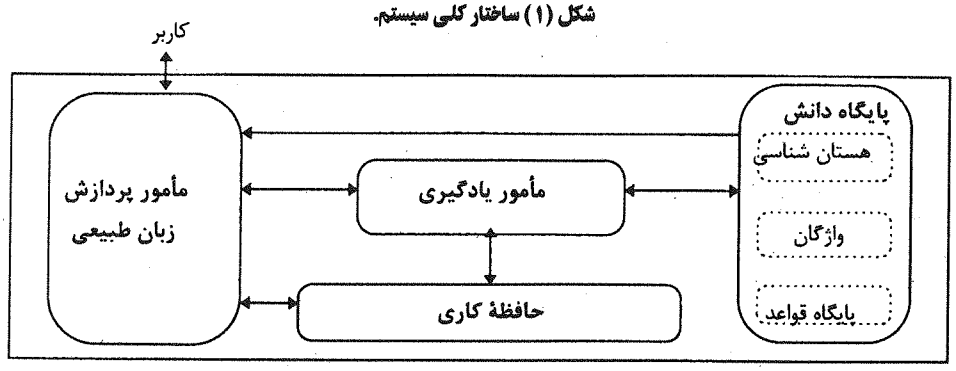
ـ تعداد زبان های تحت پوشش: واژگان مورد نظر در هستی تک زبانه است و فقط اطلاعات در مورد واژه های زبان فارسی را در بردارد.

ـ کاربرد واژگان: این واژگان فعلاً جهت درک متون به کار می رود اگر چه می توان در مراحل بعدی پروژه و یا در گسترش های آتی آن به مسئله تولید نیز توجه نمود.

ـ کاربر واژگان: این واژگان یک واژگان محاسباتی است و کاربر آن ماشین می باشد.

ـ نحوه برخورد با کلماتی با چند معنی: در این واژگان، برای حفظ نمود[15] ها (مفهوم ها)ی مختلف یک کلمه از یک مدخل با فلیدهای مختلف استفاده می شود. این تصمیم براساس ساختار انتخابی برای واژگان و امکانات زبان برنامه نویسی منتخب [۱۰] اتخاذ شده است. کلمات یکسان نه تنها ممکن است معانی مختلفی داشته باشند (مثل سه معنی کلمه شیر)، بلکه ممکن است دارای ویژگی های لغوی و نحوی متفاوتی نیز باشند (مثل شکست: فعل ماضی سوم شخص مفرد و

شکل (۱) ساختار کلی سیستم.



شکست: اسم). با توجه به اینکه در زبان منتخب (زبان لیسپ [۱۱]) امکان تخصیص پویای حافظه در صورت لزوم وجود دارد و اطلاعات به صورت نمادی پردازش می شوند، لذا می توانیم برای مدخل های مختلف تعداد و نوع ویژگی ها را متغیر در نظر بگیریم و هر گونه اطلاعات مرتبط با یک واژه را تنها در یک مدخل در کنار هم نمایش دهیم.

- نحوه نگاشت مدخل ها به واحدهای لغوی: با توجه به نحوه برخورد با کلمات هم شکل، نحوه نگاشت مدخل ها به واحدهای لغوی به صورت یک به یک خواهد بود. چرا که هر مدخل به یک واحد لغوی (یک کلمه با نمود ها یا مفهوم های متفاوت) تخصیص می یابد.

- نحوه تعیین واحدهای لغوی: واحدهای لغوی در این واژگان کلمات هستند. زیر کلمات (پیشوندها، پسوندها، میانوندها) در ساختار جداگانه ای در بخش ساختواژی نگهداری می شوند. زیر کلمات (کلمات مرکب یا چند کلمه ای ها) با توجه به انتخاب کلمات ساده در متون آزمون، در واژگان قرار نمی گیرند.

- یگانگی ساختار واژگان: واژگان مورد نظر از دو ساختار جداگانه بهره می برد: ساختار واژه ها و ساختار واژه سازها. ساختار واژه ها (که در این مقاله واژگان نامیده می شود) حاوی اطلاعات لازم در مورد واژه های زبان است و دارای ساختار یگانه است. به عبارت دیگر ساختار آن برای گروه ها و انواع مختلف واحدهای لغوی (اعم از اسم، فعل و ...) یکسان است. اما ساختار واژه سازها حاوی اطلاعات ساختواژی زبان در مورد وندهاست و اگرچه خود ساختار یگانه دارد ولی با ساختار واژگان متفاوت است. این ساختار از این پس واژگان ساختواژی نامیده می شود.

- تعیین مدخل های اصلی و زیر مدخل ها: در این واژگان مدخل ها تک سطحی بوده و زیر مدخلی وجود ندارد.

- ترتیب واحدهای با تشابهات صوری: واحدهای با تشابهات صوری به عکس ترتیب ورود (یا یادگیری) حفظ می شوند.

- تکنیک انجام توضیحات: از آنجا که کاربر، سیستم درک و یادگیری است، توضیحات به صورت متن نوشتاری تنها برای مواردی که لازم باشد عیناً در خروجی برای کاربر انسانی سیستم چاپ شود، ذخیره می شود. خود سیستم از این بخش استفاده ای نمی کند.

ب - اهم اطلاعات ذخیره شده برای هر مدخل در واژگان

در هستی با توجه به اینکه در حال حاضر واژگان جهت درک متون با کلمات و جملات ساده زبان بکار می رود، لذا از ذخیره سازی اطلاعات واجی و همچنین اطلاعاتی که جهت تولید زبان بکار می روند (مانند برخی اطلاعات جانشینی) و اطلاعات لازم جهت ساخت کلمات مرکب (برخی اطلاعات اشتقاقی) صرف نظر می کنیم. با توجه به این موارد اطلاعات ذخیره شده در واژگان هستی به شرح زیر خلاصه می شود:

- شکل نوشتاری کلمه

- دانش ساختواژی: در هستی بخش اعظم دانش ساختواژی زبان در قالب قواعد ساختواژی و یا به صورت جداگانه در ساختار واژگان ساختواژی (حاوی واژه سازها مانند پیشوندها، میانوندها و پسوندها) ذخیره می شود. بخش دیگری از این دانش که خاص کلمات و واژه هاست، در واژگان در مقابل هر واژه آورده می شود. مانند نحوه جمع بستن، جمع مکسر یا شکل مفرد یک اسم، ریشه کلمه، شکل ماضی یا مضارع فعل و ... به عبارت دیگر ما بیشتر به حفظ اطلاعات تصریفی توجه داریم و فعلاً اطلاعات اشتقاقی را نگهداری نمی کنیم.

- دانش نحوی: این بخش شامل طبقه یا نوع کلمه و ویژگی ها یا زیر طبقه کلمه می باشد. این زیر طبقات ویژگی های جزیی تری از کلمه را در طبقه خود بیان می کنند. به عنوان نمونه در طبقه فعل زیر طبقات می توانند نوع فعل (مثلاً ماضی یا مضارع) و یکسری ویژگی های دودویی دیگر مانند متعدی یا مجهول بودن و در طبقه اسم شامل ویژگی های دودویی مانند جمع یا مفرد، ذات یا معنی، شمارا یاناشمارا، عام یا خاص باشند. البته ما در مرحله اول پروژه با اطلاعات هم نشینی کار نمی کنیم ولی در مراحل بعد برخی اطلاعات هم نشینی کلمه مانند نحوه ترکیب آن با کلمات دیگر و یا با حروف اضافه و طبقات نحوی مجاز همراه شونده با فعل در این قسمت درج خواهد شد.

- دانش معنایی: در هستی، معنای اصلی واژه توسط اشاره به مفهوم متناظر آن در هستان شناسی تعیین می شود ولی ویژگی های خاص معنایی و یا محدودیت های گزینشی خاص واژه در واژگان قید می گردند. لذا بسیاری از اطلاعات معنایی که بعضاً در واژگان های محاسباتی دیده می شوند ([۵]) مانند اطلاعات جانشینی، در هستان شناسی هستی تعبیه می شود و واژگان از طریق بخش معنایی واژه به این اطلاعات دسترسی دارد.

- دانش کاربردی: در هستی بیشتر دانش کاربردی مورد نظر (مانند بخشی از اشارات ضمنی، پیش شرایط و نتایج و آثار اعمال و ...) در هستان شناسی جای دارد و بخش دیگری که متصل به واژه است (مانند ارزش واژه در میان واژه های هم گروهش) در واژگان و در ارتباط با هستان شناسی ذخیره می گردد. مثلاً برای ارزش گذاری واژه های داغ، گرم و ولرم در واژگان ارزش نسبی حرارتی به این واژه ها داده می شود که این ارزش مرتبط با مفهوم دما یا حرارت در هستان شناسی است و لذا بین همه واژه های مرتبط با این مفهوم مشترک بوده و قابل قیاس است.

۳-۶- ساختار واژگان در هستی

مفهوم های مختلف یک کلمه ممکن است دارای ویژگی های صرفی (ساختواژی)، نحوی، معنایی و کاربردی متفاوتی باشند. لذا لازم است در هر مدخل (که اطلاعات لازم در مورد یک واژه را حفظ می کند)، مفهوم های مختلف واژه را با ویژگی های متفاوت ثبت نماییم. در هستی مجموعه مفهوم های یک واژه که در صرف و نحو یکسان و در معنا متفاوتند را یک گروه مفهومی[16] می نامیم. در هر گروه مفهومی، اطلاعات ساختواژی، نحوی و لیستی از اطلاعات معنایی و کاربردی برای مفهوم های عضو این گروه ثبت می شود. ساختار یک مدخل شامل نماد شکل کلمه آن و لیستی از گروه های مفهومی منتسب به آن کلمه است. طرح واره ای از این ساختار در شکل (۲) آمده است. در این شکل مدخل i ام واژگان نشان داده شده است. در اینجا فرض شده که واژه مورد نظر دارای N گروه مفهومی است که در هر گروه، M دسته ویژگی های معنایی متفاوت وجود دارد و هر یک از این دسته ها خود دارای K گونه متفاوت کاربردی است.

برای هر فیلد از نمودهای یک واژه علاوه بر اطلاعات مذکور ضریب اطمینانی نیز ذخیره می شود. در حال حاضر مقدار ۲۵۵ یعنی مقدار اولیه وموجود در هنگام شروع بکار سیستم (تعیین شده توسط مهندس سیستم)، مقادیر میان ۱۲۹ و ۲۵۴ یعنی تأیید شده توسط کاربر شمارهٔ (ضریب منهای ۱۲۸)، مقدار ۱۲۸ یعنی تأیید شده با بیش از ۱۲۷ شاهد (جمله تأیید کننده در متن) و مقادیر کمتر از ۱۲۷ تأیید شده با تعداد ضریب جمله شاهد. جملات شاهد در ساختار نمایه سازی حفظ می شوند که شرح آن در بخش های بعد خواهد آمد. ضریب اطمینان کلی مفهوم مورد نظر توسط تابعی از ضرایب اجزاء بدست می آید. این تابع در مرحله اول به صورت ساده، تابع جمع است.

۴-۶- ساختار واژه سازها

ساختار واژه سازها ساختاری برای حفظ عناصر واژه سازی زبان (وندها) می باشد. این ساختار که در تحلیگر ساختواژی مورد استفاده قرار می گیرد، حاوی اطلاعاتی در مورد اجزاء قابل تفکیک واژه ها مانند پیشوندها، میانوندها و پسوندهاست. قواعد ساختواژی زبان که با ترکیب واژه سازها با واژه های موجود در واژگان، لغات و عبارات جدید می سازند ویا لغات مرکب را تجزیه می کنند، درپایگاه قواعد قرار دارند.

اهم اطلاعات ذخیره شده برای هر مدخل در ساختار واژه سازها عبارتند از:

- **شکل واژه ساز**
- **ویژگی های واژه ترکیب شونده:** این بخش حاوی اطلاعاتی در مورد محدودیت های گزینشی واژه ای است که با این واژه ساز قابل ترکیب می باشد. مانند طبقه نحوی و ویژگی معنایی واژه.
- **نوع ترکیب:** نشان می دهد که ترکیب به صورت پیشوندی، پسوندی و یا میانوندی صورت می گیرد.

شکل (۲) ساختار مدخل iام واژگان.

- ویژگی های واژه مرکب: طبقه نحوی و ویژگی های معنایی واژه حاصل از ترکیب را بیان می کند.

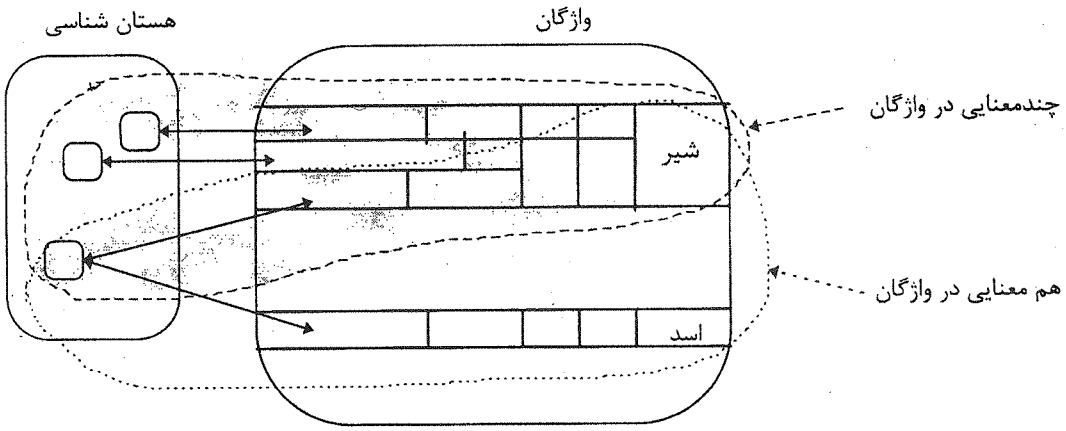
به عنوان نمونه، واژه ساز «بان» با واژه هایی با طبقه نحوی «اسم» می تواند ترکیب پسوندی بسازد و حاصل، واژه ای با طبقه نحوی «اسم» و ویژگی معنایی «محافظت» خواهد بود.

۶-۵- ارتباط واژگان با سایر اجزاء پایگاه دانش

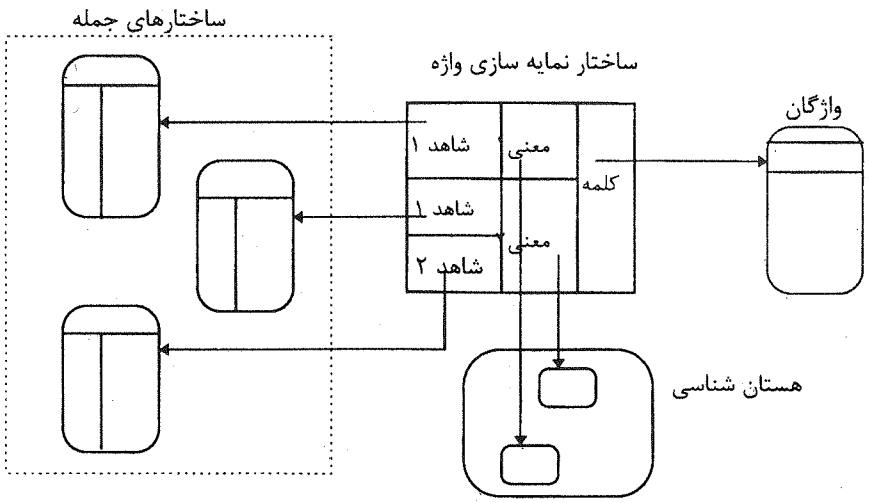
در واژگان هستی، بخش اعظم اطلاعات معنایی و کاربردی از طریق ارتباط با هستان شناسی تأمین می گردد. برای هر کلمه ممکن است بیش از یک ارتباط با هستان شناسی برقرار باشد (چند معنا بودن)[17] مثل کلمه «شیر» که بیش از یک معنا دارد و لذا با بیش از یک ارتباط معنایی به هستان شناسی مرتبط می شود. همچنین ممکن است به یک مفهوم در هستان شناسی بیش از یک کلمه مرتبط باشد (هم معنا بودن[18]). مثل مفهوم "speak" که معنایی برای کلمات مترادف گفتن، بیان کردن و صحبت کردن است. شکل (۳) طرح واره ای از ارتباط واژگان و هستان شناسی را در هستی نشان می دهد.

از آنجا که سیستم معانی لغات (مفاهیم و ارتباطات) را بتدریج و از روی متون دریافتی یاد می گیرد، به دلایل مختلف ممکن است لازم باشد در آموخته های خود تجدید نظر کرده ویا آنها را تغییر و تصحیح نماید. به این منظور لازم است سابقه ای از مسیر یادگیری خود را حفظ کند. ساختار نمایه سازی واژه ها، ساختاری است که با همین هدف طراحی شده است. این ساختار میان واژگان، هستان شناسی و ساختار متن قرار می گیرد و همه مفهوم های کلمه را در متن مشخص می کند. این ساختار برای

هر واژه لیستی از شاخص های ساختار جملات متن نگه می دارد که در آن جمله، آن کلمه با معنی مورد نظر بکار رفته است. این ساختار در حافظه کاری[19] سیستم نگهداری و با



شکل (۳) طرح واره ای از ارتباط واژگان و هستان شناسی.



شکل (۴) طرح واره ای از ارتباط واژگان، هستان شناسی و ساختارهای جمله در یك ساختار متن.

تعویض متن ورودی، آخرین نتایج مندرج در آن به ساختارهای واژگان و هستان شناسی منتقل و ساختار خالی خواهد شد. شکل (۴) طرح واره ای از ارتباط واژگان، هستان شناسی و ساختارهای جمله را در یک ساختار متن نشان می دهد.

۶-۶- یادگیری لغات و معانی تازه

واژگان هستی در ابتدای کار دارای تعداد محدودی واژه است که معانی بیشتر آنها را نمی داند. سیستم به مرور با دریافت جملات ساده، با استفاده از قواعد دستور زبان فارسی و با توجه به جملات دیگر متن، اطلاعات مربوط به کلمات موجود را تکمیل و یا کلمات جدید را به واژگان می افزاید. برای روشن تر شدن بحث، عمل یادگیری سیستم را برای مواجهه با کلمه جدید، و تکمیل اطلاعات ساختواژی، نحوی و معنایی به تفکیک بررسی می کنیم.

الف- کلمه جدید: در حالتی که یکی از کلمات متن در واژگان موجود نباشد ابتدا امکان تبدیل آن به یکی از کلمات موجود در واژگان با توجه به قواعد ساختواژی موجود وبر اساس واژگان ساختواژی بررسی می شود. مثلاً با برخورد به کلمه «کتابها» در شرایطی که واژه «کتاب» در وژاگان هست، گذر از یک مرحله تحلیل ساختواژی ما را با حذف علامت جمع به یک کلمه شناخته شده می رساند. اما اگر نتوانستیم کلمه را ساده تر کنیم (یعنی یا اجزاء کلمه در واژگان نبود یا کلمه قابل تفکیک نبود)، با یک کلمه جدید مواجهیم. در این حالت مدخلی برای آن در واژگان درنظر گرفته شده، بدنبال تکمیل اطلاعات این مدخل خواهیم بود.

ب) اطلاعات ساختواژی: در مرحله اول ساخت هستی ما از میان اطلاعات ساختواژی به نحوه جمع بستن، جمع مکسر یا شکل مفرد یک اسم جمع، ریشه کلمه، و شکل ماضی یا مضارع فعل توجه داریم. جمع های مکسر واسامی جمع در بسیاری موارد از طریق مطابقه فاعل و فعل قابل تشخیص می باشند. همچنین نحوه جمع بستن یک اسم و یا ریشه کلمه با گذر واژه از تحلیل گر ساختواژی استخراج می شود. اما در مورد بقیه موارد فوق در مرحله اول به یادگیری از طریق محاوره با کاربر بسنده کرده ایم.

ج- اطلاعات نحوی: برای دریافت و تکمیل اطلاعات نحوی یک واژه دو مسیر در سیستم وجود دارد. اول اینکه هنگام عبور واژه از مرحله تحلیل ساختواژی ممکن است نقش نحوی واژه قابل تشخیص باشد چرا که حاصل اعمال بسیاری قواعد ساختواژی برای سیستم شناخته شده است. مثلا اگر سیستم به کلمه ناشناس باغبان برخورد کند، در شرایطی که پسوند «بان» در ساختار واژه سازها آمده باشد، پس از عبور از تحلیگر ساختواژی نقش نحوی کلمه، اسم تشخیص داده می شود، چرا که سیستم می داند پسوند «بان» بر سر اسم آمده و اسم دیگر می سازد.

مسیر دوم دریافت اطلاعات نحوی، از طریق تجزیه نحوی جمله است. در این حالت ابتدا سیستم با کمک یک تجزیه گر چارت بالا به پائین[20] به دنبال قاعده ای در قواعد دستور زبان می گردد که با جمله جاری تطابق داشته باشد. این قاعده نقش دستوری کلمه یا کلمات ناشناس را تعیین می کند. در صورتی که بیش از یک قاعده برای جمله جاری یافت شده و یا بیش از یک نقش برای کلمه ناشناس بدست آمد، همه حالات ممکن برای تأیید یا رد بعدی در ساختار فرضیات حفظ خواهند شد. تأیید یا رد نتایج حاصله بر اساس جملات بعد و یا تراکنش با کاربر و توسط مأمور یادگیری انجام می پذیرد (برای بحث مفصل تر در مورد شرح وظایف و نحوه عمل مأمور یادگیری رجوع شود به [۱۳]). همچنین دریافت اطلاعاتی در مورد زیر طبقات افعال نیز در همین مرحله با توجه به کلمات یک جمله میسر است که در مراحل بعدی پروژه پیاده سازی خواهد شد.

د- اطلاعات معنایی و کاربردی: هنگامی که جمله با یکی از قواعد دستوری زبان تطابق یافت، با استفاده از الگوهای معنایی در نظر گرفته شده برای قواعد دستوری، نقش[21] کلمات جمله استخراج شده و در ساختاری به نام ساختار جمله حفظ می شود. ساختار جمله در حقیقت قاب حالتی[22] است که برای جملات کنشی شامل نقش های کنش، کنش گر، کنش پذیر، زمان، مکان، ابزار، توصیف گر کنش، مبدأ و مقصد است و برای جملات حالتی حاوی حالت پذیر، حالات، زمان و مکان می باشد.

پس از ساخته شدن ساختار جمله، به ازاء هر کلمه ای در جمله که معنای آن ناشناخته است مفهومی[23] درهستان شناسی ایجاد می گردد. این مفاهیم با توجه به طبقه کلمه و نقش آن در جمله در بالاترین سطح زیر هسته به ساختار هستان شناسی متصل می شوند[24] (به عنوان نمونه، اسامی تحت مفهوم Object، افعال تحت مفهوم Action و صفات تحت مفهوم Property). سپس این مفهوم از محل پدر اولیه خود به طرف برگ های سلسله مراتب ارث بری در هستان شناسی حرکت می نماید تا خاص ترین پدر خود را بیابد. منظور از یافتن خاصترین پدر، یافتن مفهومی است که اولاً از نظر ویژگی با مفهوم جدید تطابق داشته باشد، ثانیاً در میان مفاهیم مطابق، در پائین ترین سطح ممکن در سلسله مراتب قرار گرفته باشد. عمل یافتن تطابق خود رویه پیچیده ای دارد. ابتدایی ترین شکل تطابق در این مرحله، عدم وجود تضاد است.[25] در صورتی که همه یا هیچیک از فرزندان یک پدر با مفهوم جدید مطابقت داشته باشند، عمل انتقال متوقف و

مفهوم تحت پدر قرار می گیرد. در صورتی که بیش از یک فرزند مطابقت داشته باشد، با استفاده از یک تکنیک نگاه به جلو[26]، تا N سطح جلوتر (در مرحله اول N=۱) آزمایش می شود تا نزدیک ترین تطابق پیدا شود. در صورتی که همچنان بیش از یک انتخاب وجود داشت، در این مرحله مفهوم به گره پدر منتسب می شود[۲۷].

به مرور که مفاهیم مختلف به ازاء کلمات زبان تشکیل می گردند، این مفاهیم بر اساس ویژگی های مشترک در هم ادغام ویا بر اساس عوامل تفکیک به دسته های ریزتر شکسته می شوند. در شرایطی که اختلافی میان معنی موجود و معنی دریافتی از جمله جاری بیابد، نخست معنی جدید را با جملات قبل مقایسه می کند. در صورتی که معنی جدید با جملات قبل تطابق داشته باشد (مثلاً زیر مجموعه یا زبر مجموعه معانی سابق باشد) معنی واژه را به معنی جدید تصحیح می کند. در غیر این صورت معنی جدید را به عنوان مفهوم دیگری از واژه مورد نظر می آزماید. در هر دو صورت پیغامی مبنی بر تصمیم سیستم به کاربر انسانی ارسال می شود تا کاربر در صورت تمایل این تصمیمات را تصحیح یا تأیید نماید.

مثلاً فرض کنید اولین جمله ورودی جمله «حسن سیب را خورد» بوده، واژگان حاوی مدخل هایی با محتویات زیر باشد:

حسن: اسم خاص، مفرد.

سیب: اسم، مفرد،

را: حرف اضافه ـ نشانه

خورد: فعل ماضی، متعدی

ساختار جمله ساخته شده بر اساس قاعده دستوری:

«جمله ← گروه اسمی (کنش گر) + گروه اسمی (کنش پذیر) + را + فعل متعدی (کنش)»

نشان می دهد که حسن فاعل (کنش گر)، سیب مفعول (کنش پذیر) و خورد کنش جمله است. لذا در هستان شناسی مفاهیمی برای حسن، سیب و خورد تحت ابر مفاهیم چیز و کنش ایجاد می شود. در این مرحله چون این ابر مفاهیم خود در برگ ها قرار دارند، انتقالی صورت نمی گیرد. سپس روابط قوه کنش کاری، قوه کنش پذیری، قوه کنش کاری بر وقوه کنش پذیری از، میان این مفاهیم برقرار می گردد. پس از آن با رسیدن به جمله ای مثل «حسین سیب را خورد» با توجه به وجود مفاهیمی برای سیب و خورد، مفهومی برای حسین در هستان شناسی تشکیل می شود. در این لحظه حسن و حسین هر دو توان کنش گری یک کنش بر روی یک کنش پذیر را دارند و این یک اشتراک محسوب می شود. به مرور که جملات بیشتری توسط سیستم دریافت می شود، این اشتراکات افزایش می یابد. هنگامی که اشتراکات دو مفهوم از یک آستانه ای (آستانه ادغام[28]) فراتر رفت و در ضمن میزان اختلافاتشان از آستانه دیگری (آستانه تفکیک[29]) کمتر بوده، تضاد ویژگی نیز میانشان وجود نداشت، می توان این دو مفهوم را به یک مفهوم کلی تر ادغام و یا ابر مفهوم[30] مشترکی برای آنها در نظر گرفت ([۱۲]).

۶ـ۷ـ پیاده سازی

واژگان در این سیستم به زبان لیسپ و در محیط Allegro CL 6.0 تحت سیستم عامل Windows پیاده شده است. جهت ذخیره و بازیابی اطلاعات در واژگان از ساختار Hash-table استفاده شده و جستجو در آن مرحله اول با توابع hash موجود در محیط انجام می گیرد. واژگان در مرحله اول حاوی حروف اضافه، افعال ربطی، و تعدادی (کمتر از ۱۰ مورد) اسم و فعل است که به صورت تصادفی انتخاب شده اند.

۷ـ نتیجه گیری

درگذشته سیستم های پردازش زبان طبیعی (NLP)،به واژگان به عنوان یک جزء لازم که از توجهات نظری اندکی برخوردار بود وبه سادگی با کمک مجموعه ای از برنامه ها ساخته می شد، نگاه می کردند. در حالیکه امروزه مرکزیت واژگان در سیستم های پردازش زبان مورد تأکید است. انتخاب صحیح ویژگی های ساختاری و محتوایی و تعیین دقیق نوع و ساختار واژگان مورد نیاز بر اساس کاربرد سیستم مورد نظر از وظایف طراح است که نقش اساسی در عملکرد سیستم خواهد داشت. در نمونه عملی مورد مطالعه اگر چه هدف اصلی سیستم، پردازش زبان نبوده و عمل پردازش زبان به صورت محدود (بر روی ورودی محدود) و تنها به عنوان تغذیه کننده سیستم اصلی انجام می شود ولی بدلیل تعیین کننده بودن واژگان در کارآیی سیستم نهایی و اکتساب دانش مورد نیاز از متون به ساختار واژگان توجه خاص نموده ایم. این ساختار به سهولت قابل تعمیم برای دربرگیری ورودی های نامحدود زبان و یا تعمیم به کاربرد تولید متن می باشد.

زیرنویس ها

1- Machine Readable Dictionaries

2- Corpora

۳ ـ منابع زبانی به دو دسته اصلی تقسیم می شوند: الف ـ منابع اولیه منابع زبانی مورد استفاده هستند که به صورت عام وجود دارند و خاص استخراج اطلاعات واژگانی طراحی نشده اند. از جمله این منابع می توان به متون نوشتاری یا بخش های ضبط شده ای از گفتار اشاره

نمود.، منابع ثانویه: منابعی هستند که خود از منابع اولیه یا ثانویه دیگر استخراج شده اند مانند دیکشنری ها یا دستور زبان های موجود.

4- Dictionary

5- Lexical Unit

۶ ـ Morphemes ـ اجزاء ساخت یک کلمه، مثلاً «نا» (نفی) + «مادر» + «ی» تکواژهای ساخت کلمه «نامادری» هستند.

۷ ـ Sememes ـ اجزاء معنا که در کاربردهای مختلف متفاوت است مثلاً مفاهیم اولیه متناظر در هستان شناسی

8- Entry

9- Paradigmatic

10- Syntagmatic

11- Analogical

12- Diasystematic

13- Ontology

۱۴ ـ از لحظه آغاز بکار سیستم تا هر زمانی که سیستم مورد استفاده قرار گیرد.

15 - Sense

16- Sense Group

17- Polysemy

18- Synonymy

۱۹ ـ حافظه کاری محل نگهداری اطلاعات میان اکتسابی از محیط و حفظ ساختارهای کمکی سیستم برای انجام تحلیلات لازم است.

20- Top Down Chart Parser

21- Role

22- Case Frame

23- Concept

۲۴ ـ هسته هستان شناسی حاوی مجموعه کوچکی از مفاهیم و عملگرهای ابتدایی برای شروع به کار سیستم است. این مفاهیم در مرحله اول شامل: Thing در بالاترین سطح و Action, Object و Property در سطح دوم می باشند. همچنین عملگرهای اولیه در این مرحله عملگرهای Add جهت درج یک مفهوم، Delete جهت حذف یک مفهوم، Update جهت بهنگام سازی ویژگی های یک مفهوم، Merge جهت ادغام دو یا چند مفهوم، Split جهت تفکیک یک مفهوم و Transfer جهت انتقال یک مفهوم در ساختار هستان شناسی، می باشند.

۲۵ ـ به منظور یافتن تطابق، هر ویژگی دارای نشانگری است که نشان می دهد مقدار آن قابل ادغام، اجتماع، بازنویسی و یا تغییر توسط مقدار جدید هست یا نه. به منظور یافتن تطابق لازم است ویژگی های غیر همسان، اولاً متصاد نباشند، ثانیاً مقادیر آنها متضاد نباشند، ثالثاً این مقادیر براساس نشانگرشان با هم قابل تطابق باشند. تشخیص تضاد براساس ویژگی های منصوب به مفهوم ویژگی مورد نظر در هستان شناسی میسر می گردد.

26- Look Forward

۲۷ ـ توضیح در مورد محتوای اولیه هستان شناسی، رویه های ادغام، تفکیک و انتقال مفاهیم و نحوه تشخیص تضاد، همسانی و تطابق، بحث مفصلی را می طلبد که باتوجه به تأکید این مقاله بر بخش واژگان، در حوصله این بخش نیست. شرح این موارد در مرجع [۱۲] آورده شده است.

28- Marge Treshold

29- Split Threshold

30- Super Concept

مراجع

[1] **Atwell E. S.,** "Machine Readable Dictionaries and Lexical Datatbases", School of computer Studies, Leeds university, 1998.

[2] **Malmkjaer K.,** "The Linguistics Encyclopedia", Routledge, 1996.

[۳] شمس فرد م.، «واژگان: مباحث زبانی و محاسباتی»، گزارش فنی شماره ۱۰۱ـ۴۰۰۱ـ۷۸، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، مرداد ماه ۱۳۷۸.

[4] **Cavazza M., Zweigenbaum P.,** "Lexical Semantics: Dictionary or Encyclopedia?", in [14], pp 336-347, 1995.

[5] **Miller G. A.,** "WordNet: A Lexical Database for English", Communication of the ACM, Vol 38, No. 11, p 39-41, NOV. 1995.

[6] **Nirenburg S., et al,** "Lexicons in the Mikrokosmos project", AISB Workshop on Multilinguality in the Lexicon, Brighton, UK, 1996.

[۷] عبداله زاده ا.، شمس فرد م.، «هستی مدل پیشنهادی هستان شناسی در سیستم های درک زبان طبیعی»، هفتمین کنفرانس مهندسی برق ایران، اردیبهشت ۱۳۷۸

[8] **Shamsfard M., Abdollahzadeh A.,** "A Basis for Evolutinary Ontology Construction", 18th Iasted International Conference on Applied Informatics (AI 2000), Feb. 2000, Austria.

[۹] شمس فرد م.، «یادگیری ماشینی»، گزارش فنی شماره ۱۰۲ـ۴۰۰۱ـ۷۸، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، شهریور ماه ۱۳۷۸.

[۱۰] شمس فرد م.، «طراحی محیط پیاده سازی و تست هستی: سیستم یادگیری هستان شناسی مبتنی بر درک متون فارسی، گزارش فنی شماره ۱۰۴ـ۴۰۰۱ـ۷۹، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، خرداد ماه ۱۳۷۹.

[11] Winston **P. H., Horn B. K. P.** "Lisp", 3rd edition, Addison Wesley, 1993.

[۱۲] شمس فرد م.، «هسته هستان شناسی، مفاهیم و عمگرهای اولیه»، گزارش فنی شماره ۱۰۶ـ۴۰۰۱ـ۸۰، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، بهار ۱۳۸۰.

[۱۳] شمس فرد م.، «طراحی ساختار اولیه هستی: سیستم یادگیری هستان شناسی مبتنی بر درک متون فارسی»، گزارش فنی شماره ۱۰۵ـ۴۰۰۱ـ۷۹»، آزمایشگاه سیستم های هوشمند، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، شهریور ماه ۱۳۷۹

[14] **Saint-Diezer P., Viegas E.,** "Computational Lexical semantics", Cambridge university press, 1995.