




	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

عنوان زیرپروژه:

تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت
<h2>فهرست مطالب</h2>			
شماره صفحه	عنوان		
7	1- مقدمه		
8	1-1- مفاهیم پایه		
9	2-1- ویژگی‌های هستان‌شناسی		
11	2 - انواع هستان‌شناسی		
11	1-2- هستان‌شناسی عمومی		
11	2-2- هستان‌شناسی‌های تخصصی		
13	3 - مزایا و کاربردهای هستان‌شناسی		
17	4 - چالش‌ها و پیش‌نیازهای تولید هستان‌شناسی برای زبان فارسی		
19	1-4- چالش‌های ذاتی زبان فارسی		
22	5 - مراحل تولید هستان‌شناسی		
22	1-5- تعیین حوزه هستان‌شناسی		
23	1-1-5- پرسش‌های توانش		
24	2-5- استفاده مجدد از هستان‌شناسی‌های موجود		
25	3-5- شمارش واژه‌های مهم در هستان‌شناسی		
25	4-5- تعریف کلاس‌ها و سلسله مراتب کلاس‌ها		
26	5-5- تعریف ویژگی‌های کلاس‌ها		
27	6-5- تعریف جنبه‌های ویژگی‌ها		
29	7-5- ایجاد نمونه‌ها:		

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی			

6 - رویکردهای تولید هستان‌شناسی 30

1-6- ساخت دستی 30

عنوان شماره صفحه

2-6- استفاده از ابزارهای مهندسی هستان‌شناسی 30

3-6- ساخت نیمه خودکار 31

4-6- یادگیری و ساخت خودکار هستان‌شناسی‌ها 31

1-4-6- پیش پردازش 32

2-4-6- استخراج دانش 33

3-4-6- چک درستی و پالایش هستان‌شناسی 35

7 - مروری بر هستان‌شناسی‌های موجود 37

1-7- هستی 37

1-1-7- پیش پردازش 37

2-1-7- استخراج معنای متن 38

3-1-7- جایابی عناصر هستان‌شناسی - خوشه بندی برخط 44

4-1-7- اکتشاف دانش‌های جدید در سطح متن 44

5-1-7- پالایش هستان‌شناسی 46

2-7- پروژه سایک 47

3-7- WordNet 48

4-7- مدل عمومیت یافته سطح بالا 49

5-7- هستان‌شناسی سوا 50

6-7- هستان‌شناسی دالگرن 52



7-7- سیستم زبانی واحد پزشکی 53

8-7- Ontolingua 55

9-7- MESA 55

10-7- سازمان مجازی تورنتو 56

11-7- GENSIM 58

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

59..... SENSUS -12-7

59..... پلی‌نیوس -13-7

60..... قالب تبادل دانش -14-7



شماره صفحه عنوان

61..... OntoLearn -15-7

62..... نحوه ارزیابی هستان‌شناسی - 8



64..... نتیجه‌گیری - 9

66..... مراجع

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



فهرست شکل‌ها

شماره صفحه	عنوان
39	شکل 1- نمایش صوری ساختار جملات حالتی
39	شکل 2- ساختار جمله حالتی
40	شکل 3- نمایش صوری ساختار جملات فعلی
41	شکل 4- ساختار جمله فعلی
42	شکل 5- نمایش صوری ساختار جملات فعلی
47	شکل 6- سلسله مراتب حاصل از پردازش متن
49	شکل 7- سطوح اولیه در WordNet
52	شکل 8- سلسله مراتب مفاهیم سطح بالا در سوا
53	شکل 9- ساختار سلسله مراتب سطح بالای دالگرن
54	شکل 10- سلسله مراتب در ULMS
57	شکل 11- سلسله مراتب موجودیت - سازمان در TOVE
58	شکل 12- هستان‌شناسی سطح بالای GENSIM

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

فهرست جدول‌ها

شماره صفحه	عنوان
56.....	جدول 1- چک کردن ارتباط بین مفاهیم

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

1 - مقدمه



امروزه استفاده گسترده و همه جانبه از هستان‌شناسی و گوناگونی کاربردهای آن، مهندسی هستان‌شناسی (Ontology)، را به یکی از مهم‌ترین موضوعات مورد تحقیق و بررسی در زمینه بازنمایی و کاربرد دانش تبدیل کرده است و پیشرفت سریع علوم و تکنولوژی، و لزوم دستیابی به آخرین اطلاعات روز، بحث به اشتراک گذاشتن دانش را بیش از پیش حیاتی و مهم جلوه می‌دهد.

هستان‌شناسی، پایگاه دانش مفهومی بزرگی است که مفاهیم، روابط بین مفاهیم و اصول صریح عمومی و یا خاص یک قلمرو را نمایش می‌دهد. هستان‌شناسی به عنوان فناوری کلیدی برای شکل دهی و بکارگیری دانش، بخصوص در محیط‌های گسترده مانند وب معنایی بکار می‌رود و یکی از اهداف ساخت و به کاربردن آن، افزایش سازگاری و قابلیت استفاده مجدد اطلاعات، افزایش درون عملگری سیستم‌ها و به اشتراک گذاشتن دانش است. امروزه در دریای وسیع اطلاعات، نه تنها یک هستان‌شناسی تک برای برآورده نمودن نیازهای یک محیط توزیع شده بزرگ کفایت نمی‌کند، بلکه حتی در قلمروهای کوچک نیز بیش از یک هستان‌شناسی معتبر و مورد استفاده وجود دارد و کاربردهای مختلف در محیط‌های گسترده معمولاً لازم است به هستان‌شناسی‌های متعدد در قلمروهای متفاوت دسترسی داشته باشند.

برای هستان‌شناسی در متون هوش مصنوعی تعاریف گوناگونی می‌توان یافت. یک هستان‌شناسی توصیفی صریح و فرمال از مفاهیم یک حوزه از بحث است. کلاس‌ها که گاهی اوقات مفاهیم نیز نامیده می‌شود، ویژگی‌های هر یک از این مفاهیم و صفات خاص و خصایص مختلف این مفهوم را توصیف می‌کنند. تعریف دیگری که برای هستان‌شناسی وجود دارد نمایش اطلاعات جهان است به صورتی که قابل خواندن توسط کامپیوتر باشد و بتوان بر اساس آن استنتاج انجام داد.

یک هستان‌شناسی به همراه یک مجموعه از نمونه‌های منفرد از کلاس‌ها تشکیل یک پایگاه دانش را می‌دهند. در عمل، مرز باریکی وجود دارد که در آنجا هستان‌شناسی پایان می‌یابد و پایگاه دانش آغاز می‌گردد. به عبارت دیگر، هستان‌شناسی مدلی انتزاعی از جهان واقع است که مفاهیم و روابط میان آن را در قلمروی مورد بحث نمایش می‌دهد.

تاکنون، تعداد زیادی از هستان‌شناسی‌ها به وسیله گروه‌های مختلفی، با رویکردهای متفاوت، و با روش‌ها و تکنیک‌های متفاوت، ایجاد شده‌اند. مهندسی هستان‌شناسی به مجموعه‌ای از فعالیت‌ها می‌پردازد که به فرایند توسعه هستان‌شناسی، چرخه عمر هستان‌شناسی، روش شناسی‌ها، ابزارها، و زبان‌های مورد استفاده برای ساخت هستان‌شناسی‌ها مربوط می‌شوند. پیشرفت در ساخت هستان‌شناسی‌ها، به منظور

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی			

مدل سازی دیدگاه‌های مورد توافق از واقعیت، گستره قابل ملاحظه‌ای یافته است. اما با این حال، اجماع در این مسیر، همچنان با محدودیت‌هایی رو به رو است.



1-1- مفاهیم پایه

در قلمروی کامپیوتر، هستان‌شناسی را می‌توان با یک چهارتایی (C, R, F, A) تعریف کرد [1] که در آن:

- C مجموعه مفاهیم موجود در جهان مدل شده است.
- R مجموعه روابط میان مفاهیم است و خود به دو زیر مجموعه مجزای R_N و R_T افراز می‌شود.
 - R_T مجموعه روابط طبقه‌ای (taxonomic) میان مفاهیم است که سلسله مراتب مشمول را ایجاد می‌کند و دودویی می‌باشد.
 - R_N مجموعه روابط غیر طبقه‌ای است که ممکن است n تایی نیز باشد. ($1 < n$)
- F مجموعه تصریحات هستان‌شناسی در مورد مفاهیم و روابط آن‌هاست و خود به دو زیر مجموعه F_T و F_N افراز می‌شود
 - F_T مجموعه تصریحات هستان‌شناسی درباره روابط طبقه‌ای مفاهیم است. به عبارت دیگر، سلسله مراتب شمول را نشان می‌دهد.
 - F_N مجموعه تصریحات هستان‌شناسی درباره روابط غیر طبقه‌ای مفاهیم است.
- A مجموعه اصول بدیهی (axioms) هستان‌شناسی است که به زبان صوری، مثل منطق بیان می‌شود.

به بیان دیگر معانی در هستان‌شناسی با استفاده از پنج نوع اجزا تقسیم‌بندی می‌شوند [2].

- مفاهیم
- ساختار سلسله مراتبی مفاهیم
- روابط بین مفاهیم
- بدیهیات یا قواعد

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			



- نمونه‌ها یا موجودیت‌های واقعی

عنصر اساسی یک هستان‌شناسی همان مفاهیم یا کلاس‌ها می‌باشند. مفاهیم در هستان‌شناسی‌ها معمولاً به صورت لغوی یا ساختاری سازمان دهی می‌شوند و در واقع بیانگر معنی یا مفهوم خاص می‌باشند. کلاس نشانگر مجموعه‌ای از مفاهیم است که دارای خصوصیات یکسان می‌باشند. روابط بین مفاهیم به صورت توصیفی یا ارتباطی می‌باشند. مثلاً رابطه بین دو مفهوم کابل و هسته شامل بودن است. توابع نیز جهت بیان روابط بین دو مفهوم که رابطه بین آن‌ها یک رابطه ریاضی می‌باشد استفاده می‌گردند. کلاس‌ها مورد توجه اکثر هستان‌شناسی هستند و مفاهیم حوزه را توصیف می‌کنند. به عنوان مثال یک کلاس از نوشیدنی‌ها، نمایشگر تمام نوشیدنی‌هاست و یک نوشیدنی خاص نمونه‌ای از کلاس نوشیدنی‌ها به شمار می‌آید. یک کلاس می‌تواند زیر کلاس‌هایی داشته باشد که مفاهیم خاص تر ابرکلاس‌شان را توصیف می‌کنند. به عنوان مثال کلاس نوشابه‌ها را می‌توان به نوشابه‌های زرد، سیاه و سفید تقسیم کرد. (بر اساس رنگ نوشیدنی) هم چنین می‌توان نوشابه‌ها را به دو گروه نوشابه‌های الکلی و غیرالکلی نیز تقسیم کرد. بدیهیات، جملات همیشه صحیح یا قوانین بدیهی حاکم بر مفاهیم می‌باشند.



اسلات‌ها (ویژگی)، ویژگی‌های کلاس‌ها و نمونه‌های آن را توصیف می‌کنند. نمونه‌ها یا مثال‌ها برای نمایش اعضای هستان‌شناسی استفاده می‌شود. در کنار این اجزا، هستان‌شناسی را می‌توان با زبان‌های مختلف بیان کرد. این زبان‌ها به چهار گونه معنایی، نیمه معنایی، غیر معنایی و نیمه غیر معنایی می‌باشند. تعدادی از زبان‌های هستان‌شناسی مطرح موجود روی دستور خاص زبان xml شکل گرفته‌اند. دلیل این امر گسترش چشمگیر فن‌آوری xml در وب و انتخاب یک دستور زبان آسان و قابل واریسی و اعتبارسنجی با گسترشی از ابزارهای موجود بوده‌است.

1-2- ویژگی‌های هستان‌شناسی

هر هستان‌شناسی باید به صورت موثری معنای مورد نظر برای اصطلاحات تعریف شده را انتقال دهد [3]. تعاریف (Objective) باید عینی باشند. تعاریف می‌توانند براساس اصول فرمال بیان شوند و تعریف کامل به تعریف جزئی ارجحیت دارد. تعریف کامل با شرایط لازم و کافی بیان می‌شود و تعریف جزئی یا فقط شرایط لازم یا فقط شرایط کافی را دارد. تمام تعریف‌ها باید با زبان طبیعی مستند شوند. اصول تولید هستان‌شناسی، شفافیت، انسجام، توسعه‌پذیری و تولید هستان‌شناسی کمینه است که می‌بایست در نظر گرفته شود. خصوصیات که یک هستان‌شناسی می‌بایست داشته باشد به قرار ذیل است [4]:

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- هستان‌شناسی می‌بایست به نمایش دانش وابسته نباشد
- هر هستان‌شناسی مفاهیم پایه و قواعد ساختاری دارد که دیگر مفاهیم بر اساس این مفاهیم پایه تعریف می‌شوند
- طراحی هستان‌شناسی می‌بایست بر اساس تصمیم‌گیری مهندسی انجام شود. به عنوان مثال، تصمیم‌گیری‌هایی درباره فرار دادن مفاهیم خاص در هستان‌شناسی بر اساس تحلیل‌های خاص.

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18

2 - انواع هستان‌شناسی

هستان‌شناسی‌ها را از ابعاد مختلف می‌توان تفکیک نمود. در این تحقیق، با توجه به موضوع مورد بحث، تنها به دو بعد عمومی و تخصصی آن اشاره می‌شود.



2-1- هستان‌شناسی عمومی

این نوع هستان‌شناسی‌ها برای بازنمایی دانش عرفی قابل استفاده در حوزه‌های تخصصی مختلف به کار می‌روند. این هستان‌شناسی‌ها شامل واژگان مرتبط با چیزها، وقایع، زمان، فضا، علیت، رفتار، تابع (Mereology) و مانند آن هستند.



نمونه‌ای از هستان‌شناسی‌های عمومی، "هستان‌شناسی واحدهای استاندارد" است که شامل تعاریفی درباره واحدهای اندازه‌گیری (به طور مثال، دقیقه و ثانیه) است. هستان‌شناسی‌های سطح بالا مفاهیمی بسیار کلی را توصیف می‌کنند و در حقیقت مفاهیمی را ارائه می‌دهند که در ذیل آنها تمام اصطلاحات ریشه‌ای در هستان‌شناسی‌های موجود باید مرتبط شوند. مسئله اصلی در اینجا آن است که هستان‌شناسی‌های سطح بالای متعددی وجود دارند که در معیارهای مورد استفاده برای طبقه بندی کلی ترین مفاهیم رده بندی‌هایشان با یکدیگر متفاوتند.

2-2- هستان‌شناسی‌های تخصصی

هستان‌شناسی‌های تخصصی [5] در حوزه تخصصی مفروضی قابل استفاده‌اند. حوزه‌های تخصصی هستان‌شناسی‌ها شامل مواردی نظیر حوزه‌های پزشکی، داروسازی، مهندسی، حقوق، تجارت، و خودرو هستند. این هستان‌شناسی‌ها واژگان‌هایی درباره مفاهیم مطرح در یک حوزه تخصصی و روابط بین آنها فراهم می‌آورند، که به فعالیت‌های مربوط به آن حوزه‌ها، و نظریه‌ها و اصول اساسی حاکم بر آن حوزه‌ها، می‌پردازند. بین هستان‌شناسی‌های تخصصی و هستان‌شناسی‌های سطح بالا، مرزی روشن وجود دارد. مفاهیم مطرح در هستان‌شناسی‌های تخصصی معمولاً حالت‌های خاص مفاهیمی هستند که در هستان‌شناسی‌های سطح بالا تعریف شده‌اند، و تعبیری مشابه نیز درباره روابط بین مفاهیم، معتبر است. به طور مثال، اصطلاح "شهر" در یک هستان‌شناسی تخصصی حالت خاصی از مفهوم کلی تر "مکان" است؛

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

مکان نیز خود حالت خاصی از اصطلاح عمومی‌تر ”نقطه در فضا“ (Spatial Point) است که در هستان‌شناسی سطح بالا تعریف می‌شود به صورت مشابه، رابطه ”متصل می‌سازد“ (Connects) که در یک هستان‌شناسی سطح بالا تعریف شده است، می‌تواند به صورتی اختصاصی شود تا نقش راه‌ها در اتصال دو شهر به یکدیگر را در یک هستان‌شناسی تخصصی، بیان کند. این رابطه در این صورت به شکل ”راه‌ها شهرها را به یکدیگر متصل می‌کنند“ (Road Connects Cities) خواهد شد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			



3 - مزایا و کاربردهای هستان‌شناسی

در سال‌های اخیر، توسعه هستان‌شناسی‌ها به منزله یک توصیف فرمال و صریح از واژه‌های یک حوزه خاص و ارتباطات میان آن‌ها از یک کار آزمایشگاهی در آزمایشگاه‌های هوش مصنوعی به یک کار در کاربردهای واقعی تبدیل شده است [3]. هستان‌شناسی‌ها در دنیای وب مقبولیت پیدا می‌کنند. همچنین هستان‌شناسی‌ها در وب یک حوزه وسیع را پوشش می‌دهند. این حوزه وسیع در یک سو، شامل دسته بندی سایت‌های وب (نظیر آنچه که در دایرکتوری‌های گوگل و یاهو صورت می‌گیرد) و در سوی دیگر، دسته بندی محصولات قابل فروش و ویژگی‌های آن‌ها در یک سایت (نظیر سایت Amazon.com) را شامل می‌گردد. کنسرسیوم WWW (W3C)، استاندارد RDF (چارچوب توصیف منابع) را توسعه داد. استاندارد RDF، یک زبان برای کدینگ دانش موجود در صفحات وب، به منظور قابل فهم کردن این دانش برای عامل‌های الکترونیکی جستجوگر اطلاعات بود. سازمان DARPA نیز با همکاری کنسرسیوم WWW، زبان DAML¹ را ایجاد کرد. زبان DAML یک توسعه از RDF بود که در آن ساختارهای توصیفی بیشتری استفاده شده بود. هدف این ساختارها تسهیل تعامل عامل‌ها در وب بود [6].

هستان‌شناسی یک واژه مشترک برای پژوهشگرانی که نیاز به "اشتراکی کردن اطلاعات" در یک حوزه و حوزه خاص دارند، تعریف می‌کند. این واژه نامه شامل تعاریف قابل فهم توسط ماشین، از مفاهیم پایه یک حوزه و ارتباطات میان آن‌ها است. یکی از کاربردهای هستان‌شناسی ایجاد سیستم‌هایی برای ذخیره و بازیابی بخش‌هایی از دانش می‌باشد. به عنوان مثال در صورتی که بتوان مفهوم هستان‌شناسی یک موضوع را داشت، می‌توان ارتباطات بین مفاهیم را کشف نمود و وابستگی بین یک مفهوم با مفهوم دیگر را استخراج نمود [7].

هستان‌شناسی‌ها که پایگاه دانش مفهومی (conceptual knowledge) هستند، در محدوده وسیعی از قلمروها کاربرد دارند که برای نمونه می‌توان به شبکه‌های جهان‌گستر معنایی (semantic web)، موتورهای جست و جو، تجارب الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، سیستم‌های چند عاملی، مدل‌سازی کیفی از سیستم‌های فیزیکی، طراحی پایگاه‌های داده، سیستم‌های اطلاعات جغرافیایی، تحقیقات فیزیکی، شبیه‌سازی، مدل‌سازی و کتابخانه‌های رقمی اشاره نمود.

¹ DARPA Agent Markup lang

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18



همان‌شور که بیان شد، یکی از کاربردهای مشهود هستان‌شناسی در بازیابی اطلاعات و موتورهای جستجو است. از آنجایی که هستان‌شناسی مجموعه‌ای از مفاهیم و ارتباط بین آن‌ها می‌باشد، به بازیابی هوشمند و دقیق اطلاعات کمک می‌نماید و از بازیابی اطلاعات غیر مرتبط به مورد سوال کاربر جلوگیری می‌نماید [8].

دلایل مختلفی برای توسعه یک هستان‌شناسی متصور است. برخی از این دلایل که مرتبط به کاربرد هستان‌شناسی می‌باشد به شرح ذیل می‌باشد:

- به اشتراک گذاری فهم همگانی از ساختار اطلاعات میان عامل‌های انسانی یا ماشینی: این مساله یکی از اهداف معمول در توسعه هستان‌شناسی‌ها است [9]. به عنوان مثال چندین وب سایت مختلف شامل اطلاعات پزشکی یا سرویس‌های تجارت الکترونیک حوزه پزشکی را در نظر بگیرند، در این صورت عامل‌های کامپیوتری قادر به استخراج و جمع آوری و تجمع اطلاعات از این سایت‌های مختلف هستند. این عامل‌ها از اطلاعات بدست آمده برای پاسخ گویی به سوالات کاربران و یا به عنوان داده‌های ورودی به سایر کاربرها بهره می‌گیرند.

- قابلیت استفاده مجدد از دانش حوزه: یکی از مسائل مهمی که در ورای تحقیقات اخیر در حوزه هستان‌شناسی‌ها، دیده می‌شود، می‌باشد. به عنوان مثال، مدل‌ها برای بسیاری از حوزه‌های متفاوت، نیاز به نمایش مفهوم زمان دارند (به عبارت ساده تر، مفهوم زمان در بسیاری از حوزه‌ها وجود دارد و لذا باید به نحوی در مدل سازی حوزه‌ها آن را نمایش داد). این نمایش شامل مفاهیم بازه‌های زمانی، نقاط زمانی، سنجش نسبی زمان و موارد دیگر می‌باشد. اگر یک گروه از محققان، یک هستان‌شناسی را با ذکر جزئیات برای این مفهوم توسعه دهند، دیگران نیز به سادگی می‌توانند از این هستان‌شناسی در حوزه مورد نظر خود استفاده کنند و علاوه بر این، در صورت نیاز به ساخت یک هستان‌شناسی بزرگ، می‌توان از هستان‌شناسی‌های موجود که هر یک، بخشی از حوزه بزرگ بیان شده را پوشش می‌دهند استفاده کرده و با تجمیع آن‌ها در کنار یکدیگر، هستان‌شناسی حوزه بزرگ تر را به دست آورند. هم چنین می‌توان یک هستان‌شناسی عمومی را مورد استفاده قرار داد و با توسعه آن یک هستان‌شناسی در زمینه‌ای خاص تر به دست آورد.

- جداسازی دانش حوزه از دانش عملیاتی (واضح سازی فرضیات حوزه تحت پیاده سازی): چنانچه دانش ما درباره حوزه تغییر کند، فرضیات ما نیز به سادگی ممکن است تغییر کند.



	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

- تحلیل دانش حوزه



کاربردهای مختلف نیاز به انواع مختلف هستان‌شناسی دارند. هستان‌شناسی‌ها از نظر محتوا و اعمالی که پشتیبانی می‌نمایند با یکدیگر متفاوت می‌باشند. برای مثال کاربردهای حل مسائل علمی، تخیلی یا تجاری نیازمند هستان‌شناسی‌های کوچک، باریک و عمیق با جزئیات خاص کد شده در قالب اصول بدیهی هستند. در حالی که سیستم‌های بازیابی اطلاعات (موتورهای جستجو) ممکن است به هستان‌شناسی‌های بزرگ و سطحی که سلسله مراتبی از مفاهیم با تعداد اندکی روابط میان آن‌ها بدون اصول بدیهی را در بر می‌گیرند، نیاز داشته باشد تا کارایی جستجوی مبتنی بر کلید واژه‌ها را افزایش دهد. کاربردهایی که نیاز به درک عمیق زبان و انجام استنتاجات مختلف دارند همانند برنامه‌سازی خودکار، برخی برنامه‌های مشاوره و تصمیم‌گیری با واسط زبان طبیعی و برخی سیستم‌های پاسخ‌گویی به سوالات نیازمند هستان‌شناسی عمیق‌تر با پیش زمینه غنی و قضایای بدیهی و دارای قابلیت خودکار هستند، در حالی که برای برخی کاربردهای دیگر پردازش زبان مثل سیستم‌های ترجمه ماشین تجاری که برای تولید پیش نویس سریعی از متون در زبان مقصد بکار می‌روند واژگان معنایی همانند Wordnet [10] کفایت می‌نماید.

از دیگر کاربردهای هستان‌شناسی در پردازش زبان طبیعی می‌توان کمک به خوشه‌سازی کارای اسناد [11] و همچنین استفاده از هستان‌شناسی در امر خلاصه‌سازی اسناد نام برد. از آن جایی که هستان‌شناسی به مفاهیم و ارتباط بین آن‌ها می‌پردازد، کمک شایانی به بالابردن کیفیت این دو کار خواهد نمود. همچنین هستان‌شناسی در بازیابی و تجمیع اطلاعات از منابع مختلف توزیع شده نیز قابل استفاده است [12].

همان‌طور که بیان شد، بسیاری از هستان‌شناسی‌ها را برای انواع کاربردهای مربوط به زبان‌های طبیعی می‌سازند که شامل کاربری‌هایی نظیر دریافت دانش از متن تا بازیابی معنایی اطلاعات می‌شود. سایک، مدل عمومیت یافته سطح بالا، WordNet، هستان‌شناسی UMLS [13] در این مقوله می‌گنجند. با این که سایک [14] برای کاربری خاصی در حوزه زبان‌های طبیعی طراحی نشده است، اما یکی از انگیزه‌های عمده در این پروژه، بازنمایی دانش عمومی بوده است. سیستم‌های Wordnet و UMLS سیستم‌های مرجع بزرگی هستند که برای کاربری در سیستم‌های پردازش زبان‌های طبیعی طراحی شده‌اند، اما حوزه کاربردی خاصی را هدف گیری نکرده‌اند. در پروژه TOVE [15] بازیابی هوشمند اطلاعات، مقصود و هدف اصلی است. اما این بازیابی، بازیابی متن نیست. پروژه TOVE مدل‌هایی از سازمان‌ها را می‌سازد و سپس

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



به پرسش‌های فرمال مطرح شده در این مدل‌ها نه تنها براساس آنچه به صورت صریح در حالت ایده‌ال در خود دارد، پاسخ می‌دهد، بلکه می‌تواند برای پاسخ دادن، استنتاج نیز داشته باشد. طبقه دیگری از هستان‌شناسی‌ها جزو تحقیقات نظری هستند که مستقیماً به دنبال ایجاد یک سیستم عملیاتی نیستند. هستان‌شناسی [16] سوا نمونه‌ای از این چنین سیستمی است. سوا مبانی فلسفی لازم برای ساخت مدل‌های دانش را بررسی می‌کند و همچنین نگاهی به تاریخ هستان‌شناسی‌ها از ارسطو تاکنون دارد و سپس مدل خود را براساس این مطالعات اولیه ارائه می‌دهد.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

4 - چالش‌ها و پیش‌نیازهای تولید هستان‌شناسی برای زبان فارسی



روش‌هایی که بر اساس متون و با استفاده از روش خودکار اقدام به ایجاد و به روزآوری هستان‌شناسی می‌پردازند، نیازمند آن هستند که پیش‌پردازش کامل روی کم‌توان انجام شود تا کلمات مهم موجود در متن استخراج شود و مرز جمله‌ها برای مشخص کردن نقش معنایی کلمات مشخص شود. بنابراین تمامی چالش‌های مطرح شده در زمینه ریشه‌یابی متون، مشخص کردن مرز جملات و چالش‌های موجود در نگارش متن فارسی جزو چالش‌های اولیه تولد هستان‌شناسی فارسی نیز می‌باشد. در ریشه‌یابی افعال زبان فارسی که یکی از مراحل پردازش زبان طبیعی برای تولید هستان‌شناسی می‌باشد، چالش‌های ذیل موجود می‌باشد:

- ابهام در قواعد تولید بن مضارع و بی‌قاعده بودن برخی افعال. به عنوان مثال ("کردن"، "کرد"، "کن").
- نیاز به وجود اطلاعات آوایی از بن‌ها برای فرایند تبدیل همانند (کندن، کند، کن).
- عدم وجود اعراب در زبان فارسی و از آن جایی که یک کلمه بدون اعراب می‌تواند معانی مختلفی داشته باشد، تشخیص فعل از اسم را مشکل می‌نماید.
- شناسه‌های چسبان به فعل: ابهام میان فعل ماضی ساده و فعل مضارع دوم شخص یا سوم شخص جمع همانند "گردید"، "رنجید"، "گشتید". شناسه‌های چسبان به سادگی قابل تمایز از بخش بن در یک فعل نیستند. زیرا بعضی از آن‌ها در ترکیب با یک بن مضارع منجر به تولید یا بن ماضی می‌شوند.
- شناسه‌های چسبان به فعل، ابهام میان فعل مضارع با بن منتهی به "ی" در مورد شناسه‌های اول شخص مفرد و جمع همانند گویم، آیم. همچنین افعال مضارع با بن منتهی به "ن" در مورد شناسه‌های سوم شخص مفرد و جمع. و نیز بن‌های ماضی منتهی به (ید، د، ند...). برخی از شناسه‌های چسبان، حروف انتهایی مشترک با هم دارند. در بن مضارع منتهی به "ی" نیز تمایز میان شناسه‌های اول شخص مفرد و جمع نیاز به بررسی دقیق‌تر دارد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

- حروف التزامی در ابتدای فعل (این حروف در برخی از بن‌ها با حرف آغازین بن یکسان است. در نتیجه بعد از حذف این حروف باید مطمئن شد که باقی مانده یک بن است همانند "بگرد" از "گردیدن" و "ببر" از "بردن".
- فعل‌های مرکب با پیشوند حرفی همانند "برگرفتن" و "برداشتن"
- اتصال ضمیر مفعولی به فعل همانند "خورندش" که "آن را خوردند" است. تشخیص اینکه یک فعل به ضمیر مفعولی متصل شده است یا خیر برای یک نرم‌افزار واقعا دشوار خواهد بود.
- مختصر شدن بعضی افعال و اتصال به لغات غیر فعلی همانند "بیدارند"، "خوشحالند". صیغه‌های برخی از افعال مانند "استن" به صورت مختصر نوشته می‌شوند و به لغت قبل از خود می‌چسبند. در نتیجه کشف این افعال بسیار دشوار می‌شود.
- تعداد زیاد افعال فارسی که کار ریشه‌یابی را دچار مشکل می‌نماید.
- افعالی با اجزای پخش شده در جمله همانند "من به خانه رفته و غذا خورده‌ام"
- نبود قواعد ریشه‌یابی مشخص ابهام در امر پردازش متن را افزایش داده و آن را امری پیچیده نموده است.

استناد به زبان طبیعی جزو مسائلی است که بخصوص در زبان فارسی حساسیت بالایی می‌یابد، چرا که منابع واژگانی محدودی برای تعریف به زبان طبیعی داریم و اغلب منابع موجود فارسی در حوزه‌های دانشی مختلف، از نظر توصیف و تعریف واژگان و اصطلاحات ضعف دارند یا اساساً موجود نیست. حتی برای هستان‌شناسی‌های عمومی نیز، دسترسی به منابع توصیفی اصطلاحات به زبان طبیعی به خوبی میسر نیست. به این ترتیب یکی از پیش نیازهای تولید هستان‌شناسی‌ها تدارک همین مرجع استنادی برای تعریف و توصیف واژه‌ها به زبان فارسی است. همین ضعف باعث شده است که در کارهای انجام شده در زبان فارسی، روی طیف دادگان تا هستان‌شناسی، بیشتر در حوزه دادگان بمانیم و چندان موفقیتی در تولید هستان‌شناسی فارسی به دست نیامده باشد، چرا که اجماع نظر اهل فن در آن حوزه که جزو شرایط لازم تحقق هستان‌شناسی است، بدون استفاده از تعریف و توصیف محکم در (زبان طبیعی) فارسی، دشوار است و معمولاً تیم‌های کارشناس و مهندسی دست اندرکار تهیه تعریف‌های فرمال، یا نمی‌توانند توصیف و تعریف‌های زبان طبیعی فارسی را تهیه کنند و اگر در مواردی هم این کار را انجام داده باشند، به مجموعه‌ای از تعاریف قابل استناد و قابل اجماع نرسیده‌اند.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت

در ساخت دستی هستان‌شناسی برای زبان فارسی می‌بایست دانش اولیه آن وجود داشته‌باشد که در زبان فارسی، زیر ساخت جامعی در این زمینه هنوز ایجاد نشده‌است.

از آن جایی که ساخت دستی هستان‌شناسی مقرون هزینه بسیار، انعطاف پذیری کم و وقت‌گیر می‌باشد، ساخت خودکار هستان‌شناسی از اهمیت بالایی برخوردار است. بنابراین نیاز به استخراج دانش مفهومی از منابع ورودی و ساخت یک هستان‌شناسی بر اساس آن‌ها می‌باشد. برای این کار زیر ساخت‌های ذیل مورد نیاز می‌باشد:



- تحلیل‌گر لغوی برای استخراج لغات
- تحلیل‌گر نحوی (تعیین نقش کلمات و نوع رابطه آن‌ها با فعل)
- تعیین نقش موضوعی کلمات در جمله
- استخراج الگوهای انواع جملات (ربطی، غیر ربطی، ...)
- تبدیل ساختار جملات به عناصر هستان‌شناسی (بدست آوردن الگوهای معنایی)

چالش دیگری که در زبان فارسی مطرح می‌باشد استخراج الگوهای زبانی و الگوهای جملات می‌باشد.

4-1- چالش‌های ذاتی زبان فارسی

زبان فارسی در ساختار و قاعده (semantic) با زبان انگلیسی متفاوت می‌باشد. برخی از مشکلات ذاتی مربوط به متون فارسی در ذیل طبقه بندی شده است. لازم به ذکر است که این چالش‌ها در هنگام تولید خودکار هستان‌شناسی از روی متون فارسی در بخش پردازش زبان طبیعی نمود پیدا می‌نماید.

- نبود نکات گرامری تعریف شده همانند آن‌چه که در زبان انگلیسی وجود دارد. این مساله در ریشه‌یابی و پیش پردازش متن تاثیر گذار خواهد بود
- وجود لغات ترکیبی چند جزئی همانند "آب‌سردکن"
- در زبان فارسی، ضمائر مفعولی و نشانه‌های جمع به لغات متصل می‌شوند. این مسئله نیز منجر به پیچیدگی تفکیک لغات برای رایانه می‌شود زیرا حروف ضمائر مفعولی و نشانه‌های جمع با

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

تعدادی از وندها و نیز واژگان فارسی مشابهت دارند در نتیجه رایانه نمی‌تواند به سادگی در مورد بخش اصلی لغت قضاوت کند. به عنوان نمونه، "ان" در "درختان" نشانه‌ی جمع است در حالی که "ان" در "خوران" نشانه‌ی صفت فاعلی می‌باشد.

- در زبان فارسی قواعد تبدیل بن مضارع به بن ماضی و بالعکس به صورت قانون‌مند وجود ندارد، در نتیجه رایانه برای آنکه بتواند لغات حاصل از بن مضارع یا بن ماضی را تشخیص دهد (مثلاً برای آنکه بداند "ان" در لغت "خوران" نشانه‌ی جمع است یا نشانه‌ی صفت فاعلی)، هیچ راه مشخصی ندارد.

- ابهام ساختاری به نحوی که یک کلمه می‌تواند معانی مختلف داشت. به عنوان مثال، شیر 3 معنی متفاوت دارد. شیر حیوان، شیر آب، شیر خوراکی.

- عدم وجود قاعده خاص برای تشخیص اسامی و مکان‌های خاص همانند آنچه در زبان انگلیسی موجود می‌باشد.

- ابهام در معنی کلمه به علت نبود اطلاعات آوایی همانند "مرد" و "مُرد"

- عدم وجود دستورالعمل قطعی برای استفاده از نیم‌فاصله.

- عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات



- ناآگاهی رایانه از نقش لغت در جمله همانند "آیین نامه نوشتن" و "آیین نامه رانندگی". از آنجا که در این ترکیبات، تمام کلمات به تنهایی معنی دارند، رایانه قادر به تعیین مرز لغات نمی‌باشد.

- وجود کلمه‌های ترکیبی و امکان در نظر گرفته شدن دو کلمه مجزا. به عنوان مثال سیب زمینی که کلمه مرکبی است که از 2 کلمه سیب و زمینی تشکیل شده است.



- وجود اشتباهات نگارشی در مستندات فارسی همانند نگارش با کاراکترهای متفاوت همانند آئین‌نامه و آیین‌نامه، به کار بردن همزه به صورت‌های مختلف، و نیاز به یکسان سازی این موارد برای پیش پردازش متن به صورت کارا.

- ابهام در دستور خط زبان فارسی: پیوسته نویسی کلمات مرکب ترکیب شده با پسوند

- فقدان هر نوع دادگان بزرگ مانند یک گرامر کامل، یک مجموعه از جملات تجزیه شده نمونه و یا آمارهای ارزشمند از کاربرد لغات در جملات مختلف و جایگاه‌های مختلف در جمله.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- همانند سایر زبان‌های هند و اروپایی وجود پیشوند و پسوند و استفاده غلط از واژه‌هایی که طبق عرف تغییر معنا داده‌اند.
- فقدان یک واژگان کامل زبان فارسی با توجه به شرایط محیطی استفاده از کلمات.
- فقدان یک ارزیاب اتوماتیک برای هر قسمت از سیستم‌های پردازش زبان طبیعی.
- فقدان تجزیه‌کننده اتوماتیک ساختار کلامی متن که قسمت‌های پایه و پیرو متن را تشخیص دهد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیک متن فارسی - 3 - ت
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی			

5 - مراحل تولید هستان‌شناسی



در عمل، توسعه یک هستان‌شناسی به طور کلی شامل مراحل زیر است.

- تعریف کلاس‌ها در هستان‌شناسی
- مرتب کردن کلاس‌ها در یک سلسله مراتب "زیر کلاس - ابر کلاس"
- تعریف ویژگی‌ها و توصیف مقادیری که این ویژگی‌ها مجاز به اختیار آن‌ها هستند.
- تعیین مقادیر ویژگی‌ها برای نمونه‌های کلاس
- ایجاد پایگاه دانش با تعریف نمونه‌های منفرد از این کلاس‌ها، تعیین مقادیر خاص ویژگی‌ها و تعیین محدودیت‌های اضافی موجود بر ویژگی‌ها
- اشکال زدایی و ارزیابی نسخه اولیه هستان‌شناسی. این کار با استفاده از کاربرد هستان‌شناسی به دست آمده در یک کاربرد و یا یک PSM انجام می‌شود. هم چنین می‌توان با بحث‌های کارشناسی با صاحب نظران حوزه‌ای که هستان‌شناسی برای آن تهیه شده است، این کار را انجام داد. به هر حال باید به طریقی، فرآیند تجدید نظر (revise) در هستان‌شناسی اولیه به دست آمده صورت گیرد. این فرآیند تکراری در تمام مراحل چرخه حیات هستان‌شناسی ادامه می‌یابد.

5-1- تعیین حوزه هستان‌شناسی

مراحل مختلف توسعه یک هستان‌شناسی را با تعریف حوزه هستان‌شناسی آغاز می‌کنیم. این کار شامل پاسخ گویی به سوالات پایه زیر است:

- هستان‌شناسی در حال توسعه، چه حوزه‌ای را پوشش می‌دهد؟
- برای چه منظوری از این هستان‌شناسی استفاده می‌کنیم؟
- برای چه نوع سوالاتی، اطلاعات هستان‌شناسی باید قادر به پاسخ گویی باشد؟
- چه کسی از هستان‌شناسی استفاده می‌کند و مساله نگهداشت آن را بر عهده دارد؟

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

باید به این نکته توجه داشت که پاسخ به این سوالات در فرآیند طراحی هستان‌شناسی ممکن است تغییر کنند. اما در هر حال در هر زمان، به محدود سازی حوزه مدل کمک خواهند کرد. به عنوان مثال هستان‌شناسی "اطعمه و اشربه" را در نظر بگیرید. ما قصد داریم که از این هستان‌شناسی در کاربردی استفاده کنیم که ترکیب مناسب غذا و نوشابه را پیشنهاد می‌کند. نمایش غذا و نوشابه، حوزه هستان‌شناسی را تشکیل می‌دهند. به صورت طبیعی، مفاهیم در این هستان‌شناسی توصیف از انواع مختلف نوشیدنی‌ها، انواع مختلف غذاهای اصلی و ترکیب‌های مناسب غذا و نوشیدنی ارائه می‌دهند.

اگر هستان‌شناسی‌ای که قصد طراحی آن را داریم، به عنوان دستیار در پردازش زبان‌های طبیعی مقالات چاپ شده در یک مجله تخصصی نوشیدنی‌ها مورد استفاده قرار می‌گیرد، در این صورت لحاظ کردن مفاهیم‌ای چون مترادف‌ها و اطلاعات مربوط به بخش بندی متون و سخنرانی در این هستان‌شناسی، ضروری است. اگر هستان‌شناسی مورد طراحی، برای کمک به مشتریان یک رستوران جهت سفارش نوشیدنی مورد استفاده قرار می‌گیرد، در این صورت اطلاعات مربوط به قیمت خرده فروشی نیز باید در آن لحاظ گردد. اگر هستان‌شناسی برای افرادی با زبان دیگری به غیر از زبانی که هستان‌شناسی بر مبنای آن طراحی شده است مورد استفاده قرار گیرد، باید نگاهی میان این زبان‌ها ایجاد کرد.



5-1-1- پرسش‌های توانش

یکی از روش‌های تعیین حوزه یک هستان‌شناسی تهیه لیستی از پرسش‌هایی است که سیستم بر پایه دانش مبتنی بر آن هستان‌شناسی باید قادر به پاسخ گویی به آن‌ها باشد. این پرسش‌ها را پرسش‌های توانش گویند [17]. برخی از پرسش‌های توانش می‌تواند به شرح زیر باشد:

- آیا هستان‌شناسی اطلاعات کافی جهت پاسخگویی به این نوع سوالات را دارا می‌باشد؟
- آیا پاسخ‌ها نیاز به سطح معینی از جزئیات و یا نمایش از یک حوزه خاص را دارا می‌باشند؟

این سوالات گزینشی هستند و نه فراگیر. به عنوان مثال در همان حوزه "اشربه و اطعمه" که قبلاً نیز در مورد آن سخن گفتیم، برخی از پرسش‌های توانشی می‌توانند به شرح زیر باشند:

- هنگامی که نوشیدنی خاص را انتخاب می‌کنیم چه ویژگی‌هایی را باید مدنظر قرار دهیم؟

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- بهترین نوشیدنی موقعی که سفارش چلوکباب می‌دهیم کدام است؟
- آیا ترکیب دوغ با غذاهای دریایی مناسب است؟

بر اساس این لیست مشخص است که هستان‌شناسی باید شامل اطلاعاتی در خصوص ویژگی‌های انواع مختلف مواد غذایی و نوشیدنی، طبقه بندی غذاها و پیشنهادهای جهت ترکیب غذا و نوشیدنی باشد.



5-2- استفاده مجدد از هستان‌شناسی‌های موجود

در نظر گرفتن کاری که قبلاً توسط دیگران انجام گرفته است و اعمال تغییرات، اصلاحات و یا توسعه دادن منابع موجود جهت انطباق با حوزه و کاربرد خاص مورد نظر ما، یک فرآیند با ارزش تلقی می‌گردد. استفاده مجدد از هستان‌شناسی‌های موجود، هنگامی که سیستم مورد نظر، نیازمند تعامل با سیستم‌های کاربردی دیگری باشد که از هستان‌شناسی خاصی استفاده کرده‌اند (یا از یک واژه نامه خاص)، امری ضروری است.

بسیاری از هستان‌شناسی‌ها، به فرم الکترونیکی موجود هستند و در محیط‌هایی که جهت توسعه هستان‌شناسی مورد استفاده قرار می‌گیرند، قابلیت بازخوانی، اصلاح و استفاده از آن‌ها جهت وسعت دادن و یا ساخت هستان‌شناسی‌های دیگر وجود دارد.

فرمت ظاهری هستان‌شناسی‌ها موضوع مهمی تلقی نمی‌شود و بسیاری از سیستم‌های نمایش دانش، قابلیت پشتیبانی از اشکال مختلف نمایش هستان‌شناسی را دارا هستند. گرچه امکان تبدیل از یک شکل به اشکال دیگر نیز وجود دارد. در وب و نیز در متون مرتبط با مباحث هستان‌شناسی، کتابخانه‌ای از هستان‌شناسی‌هایی با قابلیت استفاده مجدد وجود دارد. برای مثال می‌توانیم از کتابخانه هستان‌شناسی [9] استفاده کنیم.

به عنوان مثال ممکن است یک پایگاه دانش از نوشیدنی‌های فرانسوی، از قبل موجود باشد. اگر از هستان‌شناسی‌ای که سیستم مبتنی بر پایگاه دانش مثال مورد بحث، بر اساس آن بنا نهاده شده است استفاده کنیم، نه تنها یک طبقه بندی از نوشیدنی‌های فرانسوی در اختیار داریم، بلکه نخستین گام در طبقه بندی تشخیص، توصیف و تمایز میان نوشیدنی‌ها بر اساس اطلاعات موجود در این هستان‌شناسی نیز برداشته شده است.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

5-3 - شمارش واژه‌های مهم در هستان‌شناسی

تهیه لیستی از تمامی واژه‌هایی که تمایل داریم برای آن‌ها جمله بسازیم یا در یک کاربرد درباره آن‌ها توضیح دهیم نیز کاری مفید محسوب می‌گردد. تمایل داریم در مورد چه واژگانی صحبت کنیم؟ این واژه‌ها چه ویژگی‌هایی دارند؟ درباره این واژه‌ها چه می‌خواهیم بگوییم؟ و بسیاری موارد دیگر.



دو گام بعدی شامل توسعه سلسله مراتب کلاس‌ها و تعریف ویژگی‌های مفاهیم، بسیار نزدیک به هم و در هم تنیده هستند. به گونه‌ای که تمایز میان آن‌ها و قائل به تقدم و تاخر شدن میان آن‌ها بسیار سخت است. معمولاً با ایجاد تعریف برای تعدادی محدود از مفاهیم در سلسله مراتب آغاز می‌کنیم و سپس با توصیف ویژگی‌های آن‌ها (مفاهیم ایجاد شده) کار را ادامه می‌دهیم. این دوگام، مهمترین گام‌ها در فرآیند طراحی هستان‌شناسی هستند. ما در اینجا به صورت مختصر درباره هر یک از این دو صحبت می‌کنیم.

5-4 - تعریف کلاس‌ها و سلسله مراتب کلاس‌ها

رهیافت‌های متفاوتی در توسعه سلسله مراتب کلاس‌ها وجود دارد [18]

- رهیافت بالا به پایین: فرآیند توسعه بالا به پایین با تعاریف مفاهیم عام در حوزه آغاز می‌کند. سپس با ایجاد زیر کلاس‌هایی خاص تر از این مفاهیم، فرآیند توسعه را ادامه می‌دهد.
- رهیافت پایین به بالا: فرآیند توسعه پایین به بالا با تعریف کلاس‌های بسیار خاص آغاز می‌کند - برگ‌های سلسله مراتب کلاس‌ها سپس با گروه بندی این کلاس‌ها در قالب مفاهیم عام تر، فرآیند توسعه را ادامه می‌دهد.
- رهیافت ترکیبی: فرآیند توسعه ترکیبی، ترکیبی از روش‌های بالا به پایین و پایین به بالا محسوب می‌گردد. در این روش، ابتدا مفاهیم برجسته و حائز اهمیت تعریف می‌شوند. سپس با عمومی‌سازی (ترکیب) و یا خصوصی‌سازی (تجزیه) این مفاهیم، فرآیند توسعه ادامه می‌آید.

هیچیک از این سه روش، به صورت ذاتی از مابقی بهتر نیست. رهیافتی که جهت توسعه هستان‌شناسی اتخاذ می‌گردد، کاملاً به دیدگاه شخصی طراح، از حوزه وابسته است اگر یک توسعه دهنده، دیدگاه بالا به پایین و سیستماتیک به حوزه داشته باشد، در این صورت بهتر است از روش بالا به پایین بهره ببرد. اما، به

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



نظر می‌رسد که رهیافت ترکیبی، برای بسیاری از توسعه دهندگان هستان‌شناسی‌ها راه حل ساده تر محسوب می‌گردد و تمایل بیشتری به استفاده از این روش وجود دارد. جدای از این که کدام یک از این رهیافت‌ها را مورد استفاده قرار می‌دهیم، معمولا با تعریف کلاس‌ها آغاز می‌کنیم. از لیستی که در گام سوم ایجاد کرده ایم واژه‌هایی را انتخاب می‌کنیم که اشیاء مستقل را توصیف می‌کنند. این واژه‌ها و هستان‌شناسی " کلاس " را مشخص کرده و نقاط لنگرگاه در سلسله مراتب کلاس‌ها را تشکیل می‌دهند. ما، کلاس‌ها را در یک الگوی سلسله مراتبی جای خواهیم داد. این الگوی سلسله مراتبی بر این اساس شکل می‌گیرد که اگر یک شی نمونه‌ای از یک کلاس باشد، آیا نمونه‌ای از کلاس دیگری نیز خواهد بود یا خیر؟ اگر یک کلاس A، ابر کلاس، کلاس دیگری چون B باشد، در این صورت هر نمونه از B نیز یک نمونه از A خواهد بود. به بیان دیگر، کلاس B، نشانگر مفهوم " نوعی از A " می‌باشد.

5-5- تعریف ویژگی‌های کلاس‌ها

کلاس‌ها به تنهایی، حاوی اطلاعات کافی جهت پاسخگویی به پرسش‌های توانش (در گام 1) نیستند. یک بار که برخی از کلاس‌ها را از لیست واژه‌هایی که در 5-4 ایجاد کردیم، استخراج و انتخاب می‌کنیم. اکثر واژه‌های باقی مانده ویژگی‌های کلاس‌هایی هستند که آن‌ها را ایجاد کرده ایم. برای هر ویژگی موجود در لیست (واژه‌های باقی مانده) باید مشخص کنیم که کدام کلاس آن را توصیف می‌کند. این ویژگی‌ها به کلاس ضمیمه می‌شوند.

در حالت کلی، انواع مختلفی از ویژگی‌های اشیاء وجود دارند که می‌توانند ویژگی‌های یک هستان‌شناسی را تشکیل دهند. این ویژگی‌ها را از یک دیدگاه کلی می‌توان به گروه‌های زیر تقسیم کرد

- ویژگی‌های درونی (intrinsic)
- ویژگی‌های بیرونی (extrinsic)
- اجزاء: اگر شیئی داری ساختار باشد. " بخش " می‌تواند فیزیکی یا منطقی باشد. (به عنوان مثال، بخش‌های یک وعده غذایی)

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

- ارتباط با سایر اعضا: عبارت است از ارتباط منفرد میان یک عضو کلاس و سایر اعضا (به عنوان مثال سازنده یک نوشیدنی، نمایشگر ارتباطی میان نوشیدنی و کارخانه سازنده نوشابه و ماده اولیه نوشابه را نشان می‌دهد)



باید توجه داشت که تمام زیر کلاس‌های یک کلاس، ویژگی‌های کلاس اصلی را به ارث می‌برند. مثلاً تمام ویژگی‌های کلاس نوشیدنی به زیر کلاس‌های کلاس‌های نوشیدنی شامل زیر کلاس‌های نوشیدنی گازدار، نوشیدنی طبیعی نیز می‌رسد. هر یک از این زیر کلاس‌ها علاوه بر ویژگی‌های کلاس اصلی، شامل ویژگی‌های منحصر به فرد دیگری نیز هستند. (ویژگی‌های خاص خود را دارند).

5-6- تعریف جنبه‌های ویژگی‌ها

ویژگی‌ها می‌توانند جنبه‌های مختلفی داشته باشند که این جنبه‌ها توصیفی از نوع مقادیر، مقادیر مجاز، کاردینالیتی مقادیر و سایر ویژگی‌هایی که ویژگی‌ها می‌توانند داشته باشند را توصیف می‌کنند. به عنوان مثال، مقدار name یک ویژگی (مثلاً نام یک نوشیدنی) یک رشته یا string است. لذا "name یک ویژگی با مقادیری از نوع string است". ویژگی تولید کردن می‌تواند مقادیر چندگانه داشته باشد و این مقادیر نمونه‌هایی از کلاس نوشابه هستند. به عبارت دیگر produces یک ویژگی با مقادیری از نوع Instance است که کلاس نوشابه (wine) را به عنوان مقادیر مجاز می‌گیرد. در ادامه به توصیف مختصری از برخی جنبه‌های عمومی ویژگی‌ها می‌پردازیم.

5-6-1- کاردینالیتی

این ویژگی تعداد مقادیری که یک ویژگی می‌تواند داشته باشد را تعریف می‌کند. برخی از سیستم‌ها تنها مقادیر منفرد که در آن اجازه داده می‌شود هر ویژگی حداکثر یک مقدار بگیرد و کاردینالیتی چندگانه که در آن اجازه داده می‌شود هر ویژگی به هر تعداد مقدار بگیرد، تمایز قائل می‌شوند و کاردینالیتی را به صرف تعیین یکی یا چندگانه بودن مقادیر تعریف می‌کنند. در برخی دیگر از سیستم‌ها برای تعیین حدود مقادیر مجاز یک ویژگی، کمترین و بیشترین تعداد مقادیر یک ویژگی تعیین می‌گردند. کمترین مقدار کاردینالیتی N، به معنای آن است که ویژگی باید حداقل N مقدار مختلف داشته باشد و به همین ترتیب



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

بیشترین مقدار کاردینالیتی M به معنای آن است که ویژگی حداکثر M مقدار مختلف می‌تواند داشته باشد.

5-6-2- نوع مقادیر

این ویژگی نوع مقادیری که توسط ویژگی‌ها اخذ می‌شوند را تعیین می‌کند. در این جا لیستی از برخی مقادیر معمول را ذکر می‌کنیم:



- String ساده ترین نوع مقادیری که برای ویژگی‌ها استفاده می‌گردد. این مقدار معمولاً برای ویژگی‌هایی نظیر name استفاده می‌گردد
- Number (که در برخی موارد انواع خاص تر آن نظیر اعشاری یا عدد صحیح نیز مورد استفاده قرار می‌گیرد) مقادیر عددی ویژگی‌ها را توصیف می‌کنند. مثلاً Price. قیمت یک نوشیدنی، مقداری عددی از نوع اعشاری است.
- Boolean ویژگی‌های ساده از نوع "بله-خیر" یا منطقی را تبیین می‌کند. برای مثال اگر کلاس نوشابه‌های الکلی را به عنوان یک کلاس جداگانه در نظر نگرفته باشیم، تعیین این که یک نوشابه الکلی یا غیر الکلی است از طریق یک ویژگی با مقادیر Boolean می‌تواند صورت بگیرد. در این ثورت مقدار True به منزله Yes و به معنای الکلی بودن نوشابه و مقدار false به منزله no و به معنای غیر الکلی بودن نوشابه تلقی می‌گردد.
- Enumerated (شمارشی) برای ویژگی‌هایی که یک لیست از مقادیر مجاز مشخص شده برای ویژگی وجود دارد، مورد استفاده قرار می‌گیرند. در protégé-2000 این نوع از ویژگیها به صورت مقادیری از نوع symbol تعریف می‌شوند.
- Instance این نوع مقادیر، اجازه تعریف ارتباط میان کلاس‌ها را می‌دهند.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

5-7- ایجاد نمونه‌ها

آخرین گام ایجاد نمونه‌های منفرد در سلسله مراتب است. تعریف نمونه‌های منفرد یک کلاس مستلزم انجام این سه مورد می‌باشد.

- انتخاب یک کلاس
- ایجاد یک نمونه منفرد از آن کلاس
- پرکردن ویژگی‌ها با مقادیر تعیین شده و مجاز

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

6 - رویکردهای تولید هستان‌شناسی

امروزه با توجه به روند رو به رشد استفاده از هستان‌شناسی‌ها در سیستم‌های اطلاعاتی، ساخت هستان‌شناسی، روش‌شناسی ساخت، ابزارهای ساخت، ساخت خودکار و یادگیری هستان‌شناسی‌ها از مباحث مطرح در میان محققان است.



هستان‌شناسی‌ها را ممکن است به صورت دستی، با استفاده از ابزارهای مهندسی هستان‌شناسی، با روش‌های اکتساب دانش و ساخت نیمه خودکار و یا ساخت خودکار تولید کرد که در مورد هر یک، در زیر، نمونه‌هایی ارائه شده است.

6-1- ساخت دستی

در این روش، حجم عظیمی از دانش مفهومی به وسیله افراد در ماشین کد می‌شود و پایگاه‌های دانش بزرگ عموماً تخصصی ایجاد می‌گردند. Cyc [14] و Mikrokosmos [19] نمونه‌هایی از هستان‌شناسی‌های ساخته شده با این روش هستند. ساخت دستی انواع هستان‌شناسی، برای قلمروها و کاربردهای مختلف، پرهزینه، وقتگیر و مستعد خطاست و هستان‌شناسی‌هایی که به صورت دستی ساخته می‌شوند، معمولاً گران، متمایل به نظرات شخصی طراح، در مقابل تغییرات غیر منعطف و دقیقاً خاص منظوری که برای آن تهیه شده‌اند، می‌باشند.

6-2- استفاده از ابزارهای مهندسی هستان‌شناسی

در سال‌های اخیر ابزارهایی برای پشتیبانی ساخت هستان‌شناسی ساخته شده‌اند. [20] Protégé، [21] Ontosaurus، [22] Ontolingua، [23] WebOnto با فراهم کردن واسط کاربر مناسب، محیط را برای اکتساب دانش مفهومی از کاربرد آماده و دسته دیگر مانند [24] DODDLE II و [25] SVETLAN داده‌ها و ساختارهای لازم برای ساخت هستان‌شناسی را از منابع ورودی استخراج می‌کنند و در اختیار سازنده هستان‌شناسی (انسان یا ماشین) قرار می‌دهند. وجود ابزارهای مهندسی هستان‌شناسی که فقط به صورت واسط کاربرد عمل می‌کنند، نیاز به سازنده انسانی را منتفی نمی‌کنند، بلکه فقط محیط را برای وی مهیا می‌سازند.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

6-3- ساخت نیمه خودکار



برای کاهش مشکلات ساخت هستان‌شناسی‌ها و تولید نیمه خودکار آن‌ها دو راهکار پیشنهاد شده است:

- یکپارچه سازی
- استفاده مجدد از هستان‌شناسی‌های موجود [27] [26] [28]. در سیستم‌های نیمه خودکار معمولاً کاربرد در پیشنهاد هستان‌شناسی اولیه، اعتبارسنجی و تغییر تصمیمات سیستم، انتخاب الگوهای روابط، کنترل سطحی تجرید و مواجهه با نویز، بر چسب زنی به مفاهیم جدید و تعیین وزن‌های عناصر هستان‌شناسی دارد.

6-4- یادگیری و ساخت خودکار هستان‌شناسی‌ها

مثال‌هایی از این روش سیستم‌های [31] Asium, [30] Syndikate, [29] Text-to-onto و هستی [32] می‌باشد. خودکارسازی عمل ساخت هستان‌شناسی نه فقط هزینه ساخت را کاهش می‌دهد، بلکه به تولید هستان‌شناسی با انطباق بیشتر با کاربرد آن منجر می‌شود. حرکت به سمت خودکارسازی اکتساب دانش از روی منابع، متون، پایگاه‌های داده و هستان‌شناسی‌های دیگر، مشکلات مهندسی هستان‌شناسی را محدود و هزینه ساخت و استفاده اشتراکی از هستان‌شناسی‌ها را کاهش می‌دهد. مسائل و مراحل مطرح در تولید خودکار هستان‌شناسی به قرار ذیل می‌باشد.



- عنصر آموختنی هستان‌شناسی می‌تواند دانش مفهومی و هستان‌شناسانه به تنهایی و یا ترکیب آن با دانش لغوی باشد. اصلی‌ترین عناصر لغوی که سیستم‌های موجود یاد می‌گیرند، کلمات و اصیل‌ترین عناصر آموختنی هستان‌شناسی، مفاهیم، روابط مفهومی اعم از روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی هستند.
- سیستم‌های یادگیر هستان‌شناسی از دانش پیش زمینه خود استفاده می‌کنند و از منابع ورودی، دانش‌های جدید را می‌آموزند. دانش پیش زمینه ممکن است از نوع دانش زبانی (دستور زبان، دانش لغوی، الگوها) و یا از نوع دانش مفهومی (هستان‌شناسی مبنا) باشد. در اکثر سیستم‌های موجود، برای پردازش متن از واژگان معنایی از پیش تعریف شده‌ای مانند SYNDIKATE, DODDLE II, TEXT-TO-ONTO, WordNet [10] استفاده می‌شود که معمولاً حاوی دانش

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

مفهومی نیز هست. در سیستم‌های مختلف، هستان‌شناسی اولیه ممکن است تنها هسته کوچکی از دانش اولیه [32] [30] یا یک هستان‌شناسی تخصصی در یک قلمروی خاص (SYNDIKATE, TEXT-TO-ONTO) و یا یک هستان‌شناسی کلی بزرگ حاوی دانش عرفی مانند Cyc باشد. منبع ورودی سیستم‌های یادگیر ممکن است داده‌های ساخت یافته مانند ساختار جداول پایگاه داده [33] هستان‌شناسی‌های موجود دیگر [34-36]، پایگاه‌های دانش [35] و داده‌های نیمه ساخت یافته [36] مانند فرهنگ‌های لغات و مستندات HTML, XML و یا داده‌های غیر ساخت یافته مانند متون زبان طبیعی موجود در پیکره‌های زبانی (SVETLAN', SYNDIKATE) و متون موجود در مستندات وب (TEXT-TO-ONTO) باشد.

6-4-1- پیش پردازش

متداولترین پیش پردازش استفاده شده در سیستم‌های یادگیر هستان‌شناسی از متن، پیش پردازش زبانی است. این پیش پردازش می‌تواند به صورت درک عمیق متن و یا پردازش سطحی آن باشد. درک عمیق، به کشف روابط خاص و پردازش سطحی به استخراج روابط کلی میان مفاهیم منجر می‌شود. از آنجا که درک عمیق معمولاً مشکل‌تر است و سرعت ساخت هستان‌شناسی را کاهش می‌دهد، اکثر سیستم‌ها مانند ASIUM و TEXT-TO-ONTO برای استخراج ساختارهای مورد نیازشان از متن، از تکنیک‌های پردازش سطحی متن مانند نمونه‌برداری، تعیین اجزای کلام، تحلیل نحوی و ... استفاده می‌کنند. برخی سیستم‌ها مانند SYNDIKATE نیز از روش‌های درک عمیق برای استخراج دانش مفهومی بهره می‌برند. همچنین پیمانه‌هایی وجود دارند که برای استخراج ساختارهای خاص از ورودی، به کار می‌روند. این ساختارها به وسیله سیستم یادگیر برای ساخت هستان‌شناسی استفاده می‌شوند. که از آن جمله می‌توان به SVETLAN برای استخراج طبقه‌بندی اسامی اشاره نمود.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت
		تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی	

6-4-2- استخراج دانش

روش‌های استخراج دانش در حوزه‌ای از روش‌های با دانش ضعیف مانند تکنیک‌های آماری تا روش‌های غنی از دانش (مانند استدلال منطقی) گسترده‌اند. این روش‌ها ممکن است به صورت با نظارت و یا بدون نظارت و همچنین به صورت بر خط یا برون خط مورد استفاده قرار گیرند.

6-4-2-1- روش‌های آماری



سیستم‌های بهره‌گیر از روش‌های آماری مانند [37] DODDLE II عمدتاً بر اساس فرکانس تکرار و یا فرکانس هم‌وقوعی و هم‌مکانی کلمات و عبارات عمل می‌کنند و از تحلیل آماری داده‌های هم‌مکانی برای یادگیری طبقات و روابط مفهومی از متون استفاده می‌کنند. به عنوان مثال، در [38] پس از استخراج کلمه‌های متن، با استفاده از کلمات نشانه‌ای که در [39] پیشنهاد شده‌است به کشف ارتباط بین کلمات همانند جز به کل و شامل بودن و غیره می‌پردازد.

6-4-2-2- روش‌های منطقی

سیستم‌های بهره‌مند از روش‌های منطقی، دانش جدید را با استفاده از قیاس و یا استقرا به دست می‌آورند و با گزاره‌ها، منطق درجه اول یا درجات بالاتر نمایش می‌دهند. روش‌های یادگیری منطبق بر منطق ممکن است دانش جدید را با قیاس و یا استقرا بدست آورند و با گزاره‌ها، منطق درجه اول یا درجات بالاتر نمایش دهند.

6-4-2-3- روش‌های زبان پایه

روش‌های زبان پایه مانند تحلیل نحوی، تحلیل ساختوازی-نحوی [40]، تجزیه الگوهای لغوی-نحوی، پردازش معنایی و درک متن (SYNDIKATE) عموماً وابسته به زبان هستند و برای استخراج دانش از منابع غیر ساخت یافته زبان طبیعی به کار می‌روند.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت

6-4-2-4-6 روش‌های مکاشفه‌ای

روش‌های مکاشفه‌ای ممکن است در کنار هر یک از روش‌های دیگر به کار گرفته شوند. به عبارت دیگر، روش‌های مبتنی بر مکاشفه روش‌های مستقلی نمی‌باشند بلکه برای پشتیبانی و تسهیل روش‌های دیگر بکار روند. به عنوان نمونه‌هایی از کاربرد این روش‌ها، می‌توان به سیستم‌های TEXT-TO-ONTO (برای افزایش معیار توجه روابط وابستگی زبانی اشاره نمود).



6-5-2-4-6 روش‌های ترکیبی

برخی سیستم‌ها نیز مانند TEXT-TO-ONTO از ترکیب روش‌های فوق برای یادگیری هستان‌شناسی بهره می‌برند.

6-6-2-4-6 روش‌های بر پایه الگو

در روش‌های مبتنی بر الگو، ورودی (معمولاً متن) به دنبال الگو یا کلمات کلیدی خاص که نشانگر رابطه مفهومی خاصی است، جست و جو و اطلاعات مورد نظر از متن استخراج می‌شود. الگوهای ذیل متصور می‌باشد:

- تعیین الگوهای نحوی: توابع متصل به دستور زبان هستند که به سیستم کمک می‌کنند تا ساختار جمله از روی جمله ورودی ساخته شود. این الگوها نقش‌های موضوعی اجزای کلام را تعیین می‌کنند و تعیین نقش‌های موضوعی، منجر به استخراج برخی روابط غیر طبقه‌ای جهت ساخت هستان‌شناسی می‌شود. نقش‌های موضوعی که آن‌ها را نقش‌های معنایی، حالت و یا نقش‌های تتا نیز نامیده‌اند، روابط مفهومی میان مفهوم متناظر با فعل و مفاهیم متناظر با سایر اجزای جمله هستند. محققان مختلف، نقش‌های موضوعی متفاوتی برای اجزای یک جمله معرفی کرده‌اند. این تفاوت‌ها ممکن است به دلیل تفاوت در ساختار زبان‌های مختلف، تفاوت در نظریه زبان‌شناسی مورد قبول محقق و یا تفاوت در دیدگاه مفهومی زبان باشد. از این رو، تعداد و نوع نقش‌های لازم برای توصیف روابط معنایی اجزای کلام مورد اتفاق نظر نیست. تعداد نقش‌های

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18



پیشنهادی در طیف وسیعی از تعداد بسیار کم تا صدها نوع گسترده است. در دو سر این طیف می‌توان از یک سو به نظریه اندرسون [41] مبنی بر وجود تنها سه نقش مبدأ، موقعیت و هدف و از سوی دیگر، به نظریه پولار [42] مبنی بر وجود نقش‌های موضوعی خاص برای هر فعل که با نقش‌های فعل دیگر متفاوت‌اند اشاره نمود. در حالات متعادل میانی، معمولاً تعداد محدودی نقش که اشتراکات بسیاری نیز با هم دارند، پیشنهاد می‌شوند. مثلاً پالمر [43] نقش‌های اصلی را عامل، پذیرا، ذینفع، ابزار و مکان یا موقعیت می‌داند؛ آرتس [44]، این نقش‌ها را عامل، پذیرا، موضوع، تجربه‌گر، هدف، ذینفع، مبدأ و موقعیت بر می‌شمرد و رادفورد [45] آن‌ها را موضوع / پذیرا، عامل / علت، تجربه‌گر، دریافت کننده / مالک و هدف در نظر می‌گیرد.

- تعیین نقش موضوعی کلمات: در مورد تعیین نقش موضوعی کلمات در جمله، نظرات مختلفی ارائه و الگوریتم‌هایی پیشنهاد شده است [46]. تعیین نقش‌های موضوعی جملات در اعمالی چون حاشیه‌نویسی پیکره‌های زبانی، جهت استخراج اطلاعات مختلف از آن‌ها نیز کاربرد دارد. در سال‌های اخیر، برای خودکارسازی این عمل و انجام آن به وسیله سیستم‌های پردازش متون، روش‌هایی معرفی و آزموده شده‌اند که معمولاً بر اساس ساختار نحوی و ویژگی‌های معنایی کلمات در جمله عمل می‌کنند. به عنوان نمونه‌ای از این تحقیقات می‌توان به اسمیت [47] اشاره کرد که از روی طبقات نحوی کلی در جملات زبان آلمانی، از توابع نحوی برای تعیین نقش‌های عامل، دریافت کننده و پذیرا استفاده می‌کند.



- تعیین الگوهای معنایی: از الگوهای معنایی برای تبدیل ساختارهای جمله به عناصر هستان‌شناسی استفاده می‌شود. آن‌ها، روابط معنایی و مفهومی میان مفاهیم مختلف موجود در جمله را تعیین می‌کنند و برای استخراج روابط طبقه‌ای و غیرطبقه‌ای میان مفاهیم استفاده می‌شوند. این الگوها را می‌توان بر اساس حوزه عملکرد آن‌ها و یا بر اساس نوع عنصری که استخراج می‌کنند، به دو نوع طبقه‌بندی بر اساس حوزه عملکرد و بر اساس نوع عنصر استخراجی تفکیک کرد.

6-4-3- چک درستی و پالایش هستان‌شناسی

مفاهیم ایجاد شده لازم است که مکرراً در مرحله تست و اصلاح قرار گیرند. این مساله برای چک کردن درستی ساختار و ارتباط بین مفاهیم لازم است. به بیان دیگر، هستان‌شناسی لازم است به صورت ادواری

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

مورد سازمان دهی مجدد و پالایش قرار گیرد. دوره پالایش ممکن است به کوتاهی پردازش یک جمله و یا به بلندی پردازش کل متن و یا در سیکل‌هایی مابین این دو مقدار باشد. پالایش هستان‌شناسی شامل انتقال ویژگی‌های مشترک به سطوح بالایی سلسله مراتب، هرس کردن و حذف مفاهیم اضافی و کاهش افزونگی‌ها و زوائد می‌باشد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت

7 - مروری بر هستان‌شناسی‌های موجود



در این فصل مروری بر بعضی از هستان‌شناسی‌های مطرح‌تر می‌پردازیم. به هستان‌شناسی فارسی با جزئیات بیشتر می‌پردازیم و توضیح کلی بر روی بقیه هستان‌شناسی‌ها انجام می‌دهیم.

7-1- هستی

تنها هستان‌شناسی مطرح در زبان فارسی می‌باشد. هستی [32] سیستمی برای استخراج دانش مفهومی از متون ساده زبان فارسی و ساخت هستان‌شناسی از روی آن‌هاست. هستی، از پایه، به ساخت خودکار هستان‌شناسی از یک روش ترکیبی می‌پردازد. منظور از پایه، عدم وجود هستان‌شناسی مبنا، عمومی‌ا تخصصی و همچنین نبود واژگان معنایی جهت کمک به فرآیند یادگیری می‌باشد. در ابتدای کار سیستم، واژگان تقریباً تهی و هستان‌شناسی فقط حاوی هسته اولیه یادگیری است که به صورت دستی ساخته خواهد شد. سپس واژگان معنایی و هستان‌شناسی‌های مختلف می‌توانند بر مبنای آن با توجه به ورودی‌هایشان از محیط، که متون زبان فارسی هستند، ساخته می‌شوند. هسته اولیه که تنها بخش هستان‌شناسی است که به وسیله خود سیستم ساخته نمی‌شود، حاوی فرادانش لازم جهت اضافه، حذف و تغییر عناصر هستان‌شناسی می‌باشد. در این روش، عناصر لغوی اعم از مفاهیم، کلمات و همه عناصر هستان‌شناسی روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی از متن استخراج و آموخته خواهند شد. هستان‌شناسی حاصل دارای ساختاری مرکب از گراف جهت‌دار شامل روابط سلسله‌مراتبی و غیر سلسله‌مراتبی و اصول بازنمایی شده با منطق خواهد بود. هستی، از یک رهیافت نمادین ترکیبی، مرکب از روش‌های منطقی، زبان پایه و مبتنی بر الگو و از طریق استفاده‌های متعدد از مکاشفه برای استخراج دانش بهره می‌برد و ادعا شده‌است که قادر به ساخت هستان‌شناسی‌های عمومی و تخصصی است.

7-1-1- پیش پردازش

ابتدا از تجزیه‌گر نحوی استفاده می‌شود. این تجزیه‌گر، یک تجزیه‌گر چارت بالا به پائین است و خروجی آن لیستی از درخت‌های تجزیه‌گر می‌باشد. در طول کار این قسمت هرگاه سیستم با کلمه ناشناسی برخورد کند، تحلیلگر ساختوازی سعی در تقلیل آن به یکی از کلمات و ریشه‌های شناخته شده در سیستم



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

می‌کند. اگر در این کار موفق نشود، کل کلمه به عنوان یک واحد لغوی در نظر گرفته می‌شود. برای استخراج دانش لغوی کلمه ابتدا لیستی از قواعد دستور زبان که جمله جاری با آنها تطابق دارد تهیه می‌شود. با توجه به این فرض که تمام جملات ورودی از نظر دستوری صحیح هستند، لیست فوق به پیشگو این امکان را می‌دهد که نقش‌های دستوری یا نحوی کلمه ناشناس را تعیین کند. در این مرحله کلمه جدید با ویژگی‌های استخراج شده آن به مدیر واژگان ارسال می‌شود تا در صورت لزوم به واژگان افزوده شود [48].

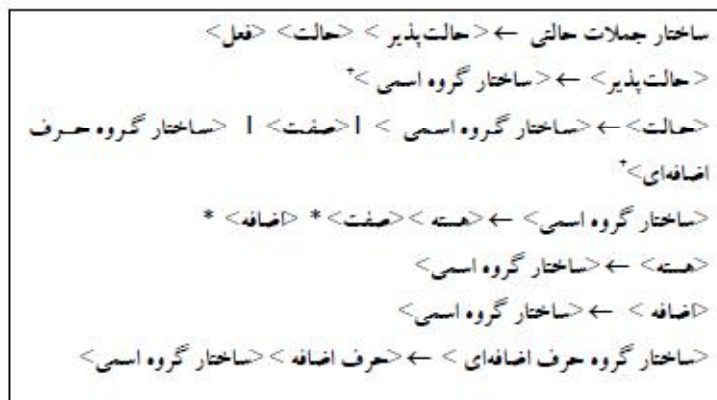
7-1-2- استخراج معنای متن

در این سیستم، ابتدا، درخت تجزیه حاصل از هر جمله که در یک تجزیه گر چارت بالا و پایین ساخته می‌شود و با استفاده از الگوهای نحوی، به ساختار جمله تبدیل می‌گردد. سپس، ساختارهای جمله ساخته شده با استفاده از الگوهای معنایی به مجموعه‌ای از عناصر هستان‌شناسی که شامل مفاهیم، روابط طبقه‌ای و غیر طبقه‌ای و اصول بدیهی می‌باشند، تبدیل می‌شوند. ساختارهای جمله حاوی اطلاعاتی در مورد نقش یک جمله می‌باشند. در هستی، جملات فارسی به دو دسته حالتی، فعلی متناظر با جملات ربطی و غیر ربطی تقسیم شده‌اند. جملات حالتی، جملاتی هستند که دارای فعل ربطی مانند "است، بود و شد" می‌باشند و وجود یا ایجاد حالت یا وضعیت را می‌رسانند. سایر جملات، فعلی هستند. ساختار جمله در حقیقت، قالب حالتی است که برای جملات فعلی شامل نقش‌های کنش، کنشگر یا عامل، کنشپذیر یا پذیرا، ذینفع، زمان، مکان، ابزار، توصیفگر کنش، مبدأ و مقصد است و برای جملات حالتی حاوی حالت‌پذیر، حالات، زمان و مکان می‌باشد. مشخصه اصلی جملات حالتی، داشتن یک مسندالیه یا نهاد و یک گزاره یا مسند است. مسندالیه در این جملات یک گروه اسمی است و مسند می‌تواند هر یک از سه حالت زیر را داشته باشد. در هر حالت، زیر مسند خط کشیده شده است [49].

- گروه اسمی در جایگاه مسند؛ مانند "حافظ، شاعر بزرگی است" و یا "احمد، دوست علی است".
- گروه صفتی در جایگاه مسند مانند "رنگ لباس علی سبز روشن بود" و یا "اتاق مینا، تمیز و مرتب شد".
- گروه حرف اضافه‌ای در جایگاه مسند مانند "این سینی از مس است" و یا "گره مریم زیر میز اتاق بود".

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			



لذا ساختار جمله حالتی دارای سه بخش اصلی نهاد، گزاره و اطلاعات فعل است. دو بخش اول، فهرستی از ساختارهای گروه اسمی‌اند و بخش سوم، زمان و نوع فعل ربطی را مشخص می‌کند. هر ساختار گروه اسمی، اجزای مختلف یک گروه اسمی شامل هسته و وابسته‌ها را به تفکیک در بخش‌های هسته، صفات و اضافات نمایش می‌دهد. برای یکسان‌سازی نمایش اطلاعات، جهت نمایش یک گروه صفتی، از یک ساختار گروه اسمی با هسته تهی و جهت نمایش یک گروه حرف اضافه‌ای، از یک ساختار گروه اسمی با هسته و صفات تهی استفاده می‌شود. شکل 1 نمایش صوری ساختار جملات حالتی و شکل 2، دو مثال از ساختارهای جمله حالتی ساخته شده برای جملات لباس خواهر مینا، پیراهن زیبایی بود و کتاب تمیز مریم روی میز بزرگ است را نشان می‌دهند



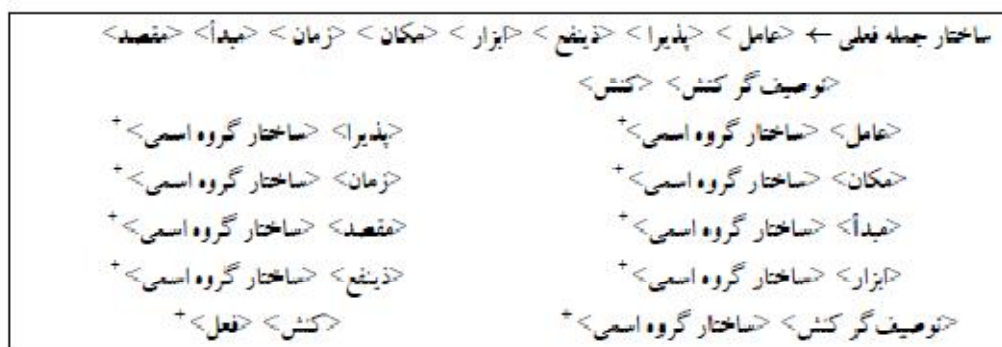
شکل 1- نمایش صوری ساختار جملات حالتی



شکل 2- ساختار جمله حالتی



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

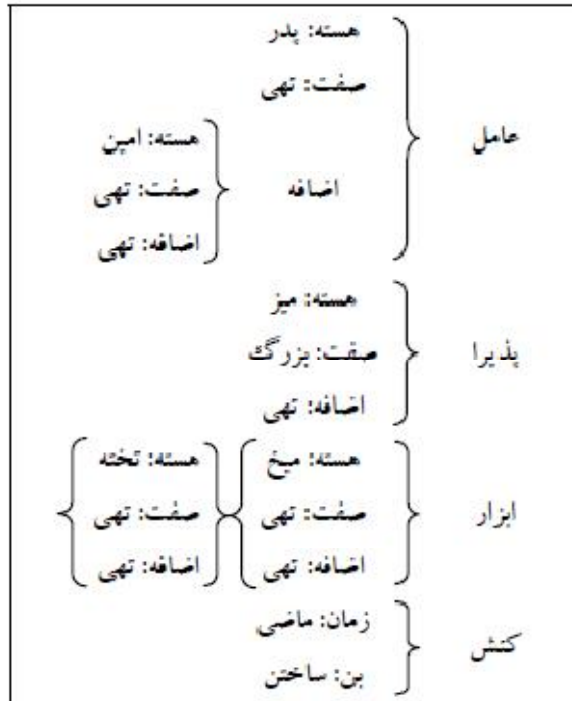
ساختار جمله فعلی حاوی اطلاعاتی در مورد نقش‌های موضوعی جمله است. نقش‌های مورد توجه عبارتند از: عامل، پذیرا، ذینفع، ابزار، مکان، زمان، مبدأ، مقصد، کنش و توصیف‌گر کنش. جملات غیر ربطی، به ساختار جمله‌ای با دوازده بخش تبدیل می‌شوند. ده بخش اول مربوط به 10 نقش موضوعی فوق، بخش یازدهم حاوی فهرست عباراتی که نقش آن‌ها در جمله شناخته نشده و بخش دوازدهم مربوط به مشخصات فعل است. یازده بخش اول هر یک، فهرستی از ساختارهای گروه اسمی‌اند و بخش دوازدهم، شامل اطلاعاتی در مورد زمان، شخص و ریشه فعل می‌باشد. شکل 3 نمایش صوری ساختار جملات فعلی را نمایش می‌دهد.



شکل 3- نمایش صوری ساختار جملات فعلی

به عنوان مثال، در جمله "پدر امین با میخ و تخته، میز بزرگ را ساخت" ساختار جمله ساخته شده حاوی بخش‌های عامل، پذیرا، ابزار و کنش خواهد بود که متناظر ساخت، تخته، میخ، میز بزرگ و پدر امین به عبارات نسبت داده می‌شوند. شکل 4 نمایی از ساختار جمله فعلی ساخته شده برای این جمله را نشان می‌دهد.



	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت



شکل 4- ساختار جمله فعلی

7-1-2-1- استفاده از الگوهای نحوی

نقش‌های مورد توجه در هستی، عامل، پذیرا، ذینفع، ابزار، موقعیت مکان، زمان، مبدأ، مقصد و توصیف‌گر کنش می‌باشند. این انتخاب محدودیت‌هایی است که فقدان دانش اولیه در مورد دلیل استخراج نقش‌ها، به وجود می‌آورند. تشخیص نقش کلمات و نوع رابطه آن‌ها با فعل، علاوه بر ساختار نحوی جمله، به ساختار معنایی آن بستگی دارد و نیازمند داشتن دانش معنایی هم در مورد کلمه مورد نظر و هم درباره فعل جمله است. برای مثال، در جمله "علی سیب می‌خورد" علی، عامل یا کنش‌کار است در حالی که در جمله "علی سیب دوست دارد" علی نقش تجربه‌گر را دارد و یا در جمله "او کتاب را دید"، کتاب موضوع محسوب می‌شود در حالی که در جمله "او کتاب را پاره کرد" کتاب پذیرا است.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

7-1-2-2- تعیین نقش کلمات

در هستی، با توجه به محدودیت دانش اولیه درباره ویژگی‌های کلمات، از رویه ساده و همسانی برای تشخیص نقش‌ها استفاده می‌شود. در این رویه، عامل، متناظر با فاعل و جملات معلوم و پذیرا و متناظر با مفعول جملات دارای فعل متعدی در نظر گرفته شده‌اند. در زیر، الگوی نحوی تعیین عامل (فاعل) و پذیرا (مفعول مستقیم) در جملات دارای فعل متعدی آمده است.

<p>جمله < ← < گروه اسمی > «عامل» < گروه فعلی ></p> <p>< گروه فعلی > ← < گروه اسمی > «پذیرا» < فعل متعدی > «کنش»</p>
--



شکل 5- نمایش صوری ساختار جملات فعلی

برای تعیین سایر نقش‌ها، بر اساس نوع حرف اضافه‌ای که پیش از آن‌ها ظاهر می‌شود و ویژگی‌های زیر، افعال در صورت وجود طبقه‌بندی می‌شوند. به این منظور، نقش‌های محتمل و همچنین نقش پیش فرض (محتمل‌ترین نقش) برای متمم، بعد از هر حرف اضافه در پایگاه دانش نوشته شده است. هر فعل می‌تواند این مقادیر پیش فرض را در مدخل خود در واژگان بازنویسی کند. مثلاً، گرچه نقش پیش فرض برای متمم پس از حرف "به" نقش مقصد است ولی برای فعل "دادن" می‌توان آن را به نقش ذینفع بازنویسی نمود. در صورتی که این اطلاعات زیر طبقه بندی، در واژگان موجود نباشند و یا بیش از یک نقش را برای یک متمم معرفی کنند، کلیه نقش‌های محتمل با تعیین حالت پیش فرض برای پردازش‌های بعدی، در نظر گرفته خواهند شد.

7-1-2-3- تعیین الگوی معنایی

برای تعیین الگوی معنایی با رویکرد حوزه عملکرد از سه الگوی ذیل استفاده شده است:

- الگوهای جمله ربطی: در جملات ربطی الگوهای معنایی برای تعیین روابط شمول، جزء-کل، ویژگی داشتن و ویژگی بودن بکار می‌روند.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



- الگوهای جمله غیر ربط: در این جملات روابط مفهومی اصلی را نقش‌های موضوعی و ارتباطات میان فعل و سایر اجزاء جمله می‌سازند. این الگوها که برای استخراج روابط مفهومی از جملات غیر ربطی به کار می‌روند، ارتباطات میان فعل و سایر اجزای جمله را بر اساس نقش‌های موضوعی استخراج می‌کنند. میزان جزئی بودن روابط استخراج شده به وسیله پارامترهای سیستم قابل تنظیم است. برای این میزان دو حالت متصور است که در یکی فقط روابط فعل با سایر اجزا و در دیگری روابط همه اجزا با هم فعل و غیر فعل استخراج می‌شوند که در هر دو حالت، رابطه میان نمونه فعل با نمونه هر جزء، همان نقش موضوعی جزء مورد نظر است و در حالت دوم، رابطه سایر اجزا با هم به فعل بستگی دارد.

- الگوهای عبارات اسمی: این الگوها روابط مفهومی موجود میان مفاهیم متناظر با صفت و موصوف مانند ویژگی داشتن و یا میان مضاف و مضاف الیه مانند مالکیت را استخراج می‌کنند. این الگوها روابط مفهومی موجود در عبارات اسمی را استخراج می‌کنند. هر عبارت اسمی دارای یک هسته و تعدادی وابسته است. هر وابسته که به هسته اضافه می‌شود، ممکن است یک گروه صفتی یا یک گروه اسمی باشد که متناظراً نوع ترکیب را وصفی یا اضافی نامیده شده. در ترکیبات وصفی، رابطه میان هسته و صفات آن یک رابطه ویژگی داشتن / ویژگی بودن است که در صورتی که نوع ترکیب وصفی درست تشخیص داده شده باشد، استخراج این رابطه بر اساس الگوی وصف به سهولت انجام می‌شود.

الگوهای معنایی به طریق دیگری نیز بر اساس آن‌چه استخراج می‌کنند دسته‌بندی و لحاظ شده است که به صورت ذیل دسته بندی خواهد شد.

- الگوهای استخراج روابط طبقه‌ای: روابط طبقه‌ای، روابطی هستند که عناصر مرتبط را در یک طبقه‌بندی یا سلسله مراتب جا می‌دهند؛ مانند رابطه شمول معنایی. از معروف‌ترین الگوهای استخراج این روابط، شش الگوی هیرست [50] استفاده شده است و آن را برای فارسی تطبیق داده‌اند.

- الگوهای استخراج روابط غیر طبقه‌ای: برخی از روابط غیر طبقه‌ای جزو هسته هستان‌شناسی و از پیش تعریف شده می‌باشند. الگوهای استخراج این روابط، مفهوم جدیدی برای این روابط ایجاد نمی‌کنند و فقط به استخراج موارد رخداد آنها و کشف مفاهیم مرتبط به وسیله آنها می‌پردازند.

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18

مهم‌ترین روابط از پیش تعریف شده، روابط میان مفهوم کنش متناظر با فعل و مفاهیم متناظر با نقش‌های موضوعی موجود در جملات غیر ربطی هستند که به وسیله الگوهای جملات غیر ربطی استخراج می‌شوند. همچنین روابط دیگری چون تعلق، هم مرجعی و ویژگی داشتن نیز از روابط هسته می‌باشند. سایر روابط، از پیش تعریف شده نیستند و الگوهای معنایی نه تنها موارد رخداد آنها، بلکه مفهوم متناظر با آنها را نیز ایجاد می‌کنند. به عنوان مثال‌هایی از این گروه می‌توان به روابط برادر، کارمند، هم مکان و رنگ اشاره نمود.



بدین ترتیب پس از ساخته شدن مفاهیم ابتدایی طی یک عمل خوشه بندی برخط یا افزایشی در هستان‌شناسی جا داده می‌شوند.

7-1-3- جایابی عناصر هستان‌شناسی - خوشه بندی برخط

این نوع خوشه بندی در سطح جمله عمل می‌کند. در این حالت با رسیدن دانش جدید استخراج شده از جمله جاری در مورد یک مفهوم، پیمانه جایابی جهت یافتن پدر این مفهوم در سطح مناسبی از تجرید فعال می‌شود. پیمانه جایابی مفهوم را از محل اولیه خود به سمت پایین به طرف برگ‌های سلسله مراتب جهت یافتن خاص‌ترین پدر آن حرکت می‌دهد. خاص‌ترین پدر یک مفهوم عمیق‌ترین نواده پدر اولیه آن است که مجموعه ویژگی‌های آن، زیرمجموعه‌ای از مجموعه ویژگی‌های آن مفهوم باشد.

7-1-4- اکتشاف دانش‌های جدید در سطح متن

پس از تبدیل جمله به عناصر هستان‌شناسی، سیستم با اعمال قواعد استنتاجی مانند استلزام منطقی، قواعد ارث‌بری ویژگی‌ها و قوانین تعدی روابط به اکتشاف دانش‌های جدید از پایگاه دانش می‌پردازد. این اکتشاف، معمولاً به درخواست کاربر و یا بر اساس نیاز جهت سازماندهی هستان‌شناسی صورت می‌گیرد. نخستین گام در سازماندهی هستان‌شناسی خوشه بندی مفاهیم و روابط موجود است که منجر به کشف مفاهیم و یا روابط جدید می‌شود. در مدل پیشنهادی، عمل خوشه بندی به دو صورت برخط و برون خط صورت می‌پذیرد.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



7-1-4-1- خوشه بندی برون خط یا دفعی

پس از ایجاد مفاهیم و جایابی آنها در هستان‌شناسی، این مفاهیم ممکن است بر اساس ویژگی‌های مشترک در هم ادغام و یا بر مبنای وجوه افتراقشان از هم بگسلند. این کار در طی عمل خوشه بندی دفعی انجام می‌شود. شرط اصلی جمع شدن در یک خوشه و قرار گرفتن تحت ابر مفهوم پدر مشترک، داشتن روابط مشابه با آنتل‌های مشابه است. الگوریتم خوشه بندی در سه سطح مختلف از شرایط ادغام اعمال می‌شود:

- سطح یک شباهت: داشتن روابط مستقیم (صریح) یکسان با مفاهیم یکسان
 - سطح دوی شباهت: داشتن روابط غیر مستقیم (غیر صریح و قابل نتیجه گیری) یکسان با مفاهیم یکسان،
 - سطح سه شباهت: داشتن روابط یکسان با مفاهیم مشابه (که در دو سطح یک و دو شبیه تشخیص داده شده‌اند)
- در الگوریتم‌های خوشه بندی معیار شباهت تعیین کننده میزان شباهت میان دو کلاس یا مفهوم است. در این سیستم معیار شباهت میان دو مفهوم b و a به صورت زیر تعریف می‌شود:

$Sim(c)$

در این فرمول L_j ثابت سطح و W_x نمایشگر وزن x می‌باشد که x می‌تواند یک رابطه یا یک عنصر مرتبط باشد. $card(x)$ تعداد اعضای x را بر می‌گرداند. $N_f(c)$ مجموعه ویژگی‌های غیر طبقه‌ای مفهوم c می‌باشد. ثابت سطح مقدار ثابتی است که به هر سطح شباهت منتسب می‌شود و با افزایش سطح کاهش می‌یابد. جهت کنترل فرایند یافتن شباهت در خوشه بندی، به روابط و مفاهیم مرتبط وزن‌هایی منتسب می‌شوند. این وزن‌ها نمایانگر میزان اهمیت و برجستگی عناصر وزن داده شده از دید کاربر جاری و یا در قلمرو یا کاربرد جاری هستند و موجب می‌شوند که هستان‌شناسی ساخته شده پویا و انعطاف پذیر در مقابل تغییرات کاربر و یا قلمرو باشد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیکرمتن فارس - 3 - ت

7-1-5- پالایش هستان شناسی

سه عمل به منظور پالایش دوره‌ای هستان شناسی تعریف شده‌اند:

7-1-5-1- فاکتورگیری



وقتی همه فرزندان یک گره دارای ویژگی‌های مشترک هستند، ویژگی‌های مشترک همه فرزندان به پدر منتقل و از تک تک فرزندان حذف می‌شوند. به این عمل، فاکتورگیری از ویژگی‌ها گوییم که حالتی از یادگیری استقرائی است.

7-1-5-2- هرس کردن

در عمل هرس کردن، گره‌های اضافی و غیر ضروری گراف هستان شناسی حذف و فرزندان آنها به تحت پدر بلافصل گره اضافی منتقل می‌شوند. یک گره غیر ضروری گره‌ای است که هیچ واحد لغوی به آن ارجاع نمی‌کند و مجموعه ویژگی غیر طبقه‌ای آن تهی است.

7-1-5-3- حذف مسیرهای تکراری

برای حذف مسیرهای تکراری اگر گره‌ای رابطه طبقه‌ای مستقیمی بی واسطه (با یکی از نیاکان) پدران غیر بلافصل خود داشته باشد آن رابطه حذف خواهد شد. در این سیستم افزونگی‌هایی که صراحتاً در متن قید شده‌اند حفظ می‌شوند. شکل 6 نمایی از درخت سلسله مراتب حاصل از پردازش متن نمونه توسط سیستم هستی را نشان می‌دهد.

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18





شکل 6- سلسله مراتب حاصل از پردازش متن

7-2- پروژه سایک

پروژه سایک [14] به منظور ایجاد یک هستان‌شناسی عمومی برای دانش عرفی در سال 1985 آغاز شد. سایک بیش از ده هزار گونه مفهومی در خود دارد که در مجموعه قواعد و اطلاعات گنجانده در پایگاه دانش پروژه به کار رفته‌اند. سطح بالای سلسله مراتب سایک سه شاخه دارد. در بالای سلسله مراتب، مفهوم "چیزها" وجود دارد که ویژگی‌هایی از آن خود ندارند. سلسله مراتب زیر چیزها کاملاً پیچیده است. البته تمام این زیرمقوله‌ها بسته و انحصاری نیستند. به طور کلی، چیزها در سه شاخه زیر تقسیم شده‌اند:



شاخه اول اشیای باز نموده است (represented things) که در نقطه مقابل اشیای ماشین درونی قرار می‌گیرند. هر مقوله سایک باید مصداقی از یک و فقط یکی از این مجموعه‌ها باشد. اشیای ماشین درونی همه آن چیزهایی هستند که نسبت به زیر ساخت سایک، بومی (local) به حساب می‌آیند (رشته‌ها، اعداد و مانند آن) اشیای باز نموده همه چیزهای دیگر هستند. سوا (Sowa) [16] این را رد می‌کند و معتقد است که سایک نباید از بازنمایی اشیا در ماشینی که خود ساخته است، معاف شود. شاخه دوم، شامل اشیای فردی است که در نقطه مقابل مجموعه (collection) قرار می‌گیرند. مجموعه‌ها شامل تمام مقوله‌های مطرح در سایک است. بنابراین، مجموعه جرم ندارد و محسوس و ملموس نیست. سوا در اینجا این مسأله را طرح می‌کند که به این ترتیب روشن نیست، جایگاه چیزی مانند دسته پرنده‌گان کجاست. چرا که یک مجموعه به حساب می‌آید و به روشنی، محسوس و ملموس نیز هست. شاخه سوم اشیای نامشهود

	عنوان پروژه:			
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی			
	عنوان زیر پروژه:			
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18

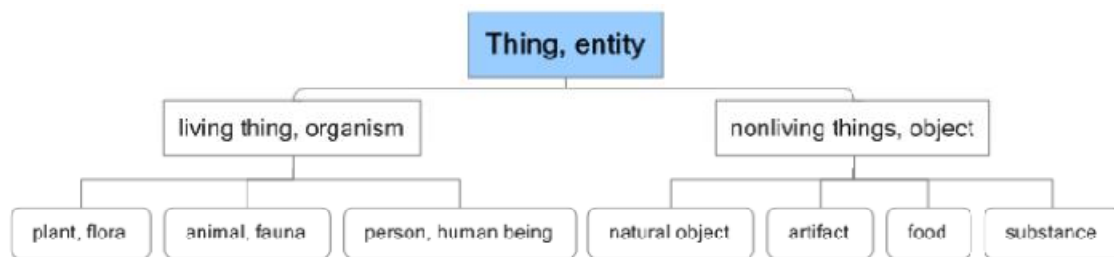
در مقابل اشیای مشهود و همچنین در مقابل مفاهیم مرکب اشیای مشهود و نامشهود است. هر واحد در سایک مصداقی از دقیقاً یکی از این سه مقوله است. نامشهود هر چیزی است که جرم ندارد (مجموعه تمام مردم، شماره هفتاد و دو و مانند آن) در حالی که شیء مشهود هر چیزی است که جرم و انرژی دارد (یک تکه سنگ، بدن انسان). اشیای مرکب مشهود و نامشهود چیزهایی هستند که هم گستره‌ای مادی و فیزیکی دارند و هم گستره‌ای غیرمادی و نامشهود. به طور مثال، هر شخص معین، دارای جسمی است که گستره فیزیکی اوست و دارای ذهنی است که گستره نامشهود او به حساب می‌آید.

WordNet - 3-7

WordNet [10] یکی از توسعه‌یافته‌ترین هستان‌شناسی‌های قاموسی است. هستان‌شناسی WordNet یک سیستم مرجع قاموسی است که به صورت دستی ایجاد شده است. اشیای قاموسی WordNet به صورت معنایی سازمان دهی شده‌اند که در آن تمایز اساسی روی مقوله‌های دستوری اسم، فعل، صفت، و قید تعریف شده است مفهوم مرکزی در طراحی ساختار WordNet، مجموعه‌های هم معنایی است که مجموعه‌هایی از واژه‌های هم‌معنا یا مترادف هستند. اگر یک واژه بیش از یک حوزه معنایی (Sense) داشته باشد، آن واژه در بیش از یک مجموعه هم معنایی ظاهر خواهد شد. در WordNet، بیش از هفتاد هزار مجموعه هم معنایی داریم. مجموعه‌های هم معنایی در سلسله مراتبی از روابط فرارده - فرورده سازمان دهی می‌شوند. این رابطه در WordNet، فراواژگانی - فراواژگانی نامیده می‌شود. برای هر مفهوم (که مترادف آن یک مجموعه هم معنایی تعریف شده است)، یک نشانگر به اسم‌هایی داریم که اجزای آن را نشان می‌دهند. به طور مثال، اجزای مفهوم "پرنده" نوک و بال‌ها هستند. در WordNet این امکان وجود دارد که انواع دیگری از نشانگرها نیز تعریف شوند. به طور مثال از اسم به فعل برای نمایش کارکردها یا از اسم به صفت برای معرفی خواص (اما این نوع نشانگرها به صورت کاملاً محدودی پیاده سازی شده‌اند). هستان‌شناسی WordNet یک رده‌بندی است؛ مفاهیم یا اصول بدیهی در آن ساخت یافته نیستند. با این که WordNet از سلسله مراتبی ساده برای مجموعه‌های هم معنایی اسمی استفاده می‌کند، سازمان دهی مجموعه‌های هم معنایی آن برای افعال و صفات، متفاوت است. صفات توصیفی در آن، در قالب خوشه‌های دو قطبی مبتنی بر تضاد سازمان دهی شده‌اند. به طور مثال، دو صفت خشک و تر، خوشه‌ای دو قطبی از مترادف‌های این دو صفت را می‌سازند که هر کدام در شاخه متناظر خود قرار می‌گیرند. صفات‌های رابطه‌ای مانند «دوستانه و صمیمی» فقط در مجموعه‌های هم معنایی معینی قرار می‌گیرند. که نشانگرهایی به اسم‌های متناظر آنها دارند. افعال در WordNet به

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			



پانزده خوشه تقسیم می‌شوند؛ خوشه‌های افعال براساس معانی آنها تعریف می‌شوند. استلزام یا تضمن رابطه اولیه بین افعال در یک خوشه است. اغلب این خوشه‌ها متناظر حوزه‌های معنایی هستند: افعال کارکردها و مراقبت‌های فیزیکی، افعال تغییر و تحول، افعال شناختی، افعال ارتباطی، افعال رقابت و مانند آنها. افعالی مانند کفایت کردن، تعلق داشتن، و شباهت داشتن، که به هیچ یک از حوزه‌های معنایی تعلق ندارند و فقط به حالت‌ها ارجاع می‌دهند، گروهی جداگانه تشکیل می‌دهند. شکل 7 سطوح اولیه سلسله مراتب را در WordNet نشان می‌دهد.



شکل 7- سطوح اولیه در WordNet

7-4- مدل عمومیت یافته سطح بالا

مدل عمومیت یافته سطح بالا یا GUM هستان‌شناسی عمومی و مستقل از حوزه‌های تخصصی کاربردی و مبتنی بر رویکرد زبان شناختی است که پردازش‌های پیچیده زبان‌های طبیعی را برای انگلیسی، آلمانی و ایتالیایی، پشتیبانی می‌کند. سطح تجرید آن بین دانش قاموسی و دانش مفهومی قرار دارد. ادعای این هستان‌شناسی، ساده سازی فاصل بین دانش تخصصی و منابع زبان شناختی عمومی است. این مدل، دارای نوعی سازمان دهی در حوزه‌های دانش است و فقط یک رده بندی دارد؛ این ساختار با این فرض طراحی شده است که ابزارهای پردازش زبان‌های طبیعی کاربر این هستان‌شناسی، خود می‌توانند اطلاعات بدیهی (مرتبط با اصول بدیهی) را برای پردازش رایانه‌ای رمزگذاری کنند. در این هستان‌شناسی، سلسله مراتب بزرگی از مفاهیم (در حدود 250 مفهوم) و همچنین سلسله مراتب جداگانه‌ای برای روابط مدل (سطح بالا) وجود دارد. در رده بالای سلسله مراتب مفاهیم، مفهوم چیز یا "شیء" مدل سطح بالا قرار دارد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0	تاریخ: 1388/04/18

این مفهوم به سه زیرگونه عمده که معرف عمومی‌ترین پدیده یا وضعیت است، تقسیم می‌شود که عبارتند از:



- 1) پیکربندی، به معنی پیکربندی تمام عناصری که در فعالیت یا وضعیتی شرکت دارند.
- 2) عنصر، یک اصطلاح مفهومی منفرد.
- 3) دنباله، به معنی وضعیت پیچیده‌ای که در آن فعالیت‌ها یا پیکربندی‌های مختلف به وسیله رابطه‌ای به یکدیگر پیوند خورده‌اند تا یک دنباله را بسازند.

پس از آن، از روابط برای وصل کردن عناصر به پیکربندی‌ها و دنباله‌ها استفاده می‌شود. اغلب روابط، بین یک فرایند و شرکت کنندگان در آن فرآیند، چگونگی انجام یافتن آن فرایند و مانند آن برقرار می‌شوند. به این ترتیب، مقوله رابطه مدل سطح بالا به 3 زیر مقوله تقسیم می‌شود. 1) فرایند در پیکربندی، 2) مقتضیات در پیکربندی و 3) شرکت کنندگان در پیکربندی. به طور مثال، رابطه علی، یکی از مقتضیات در روابط پیکربندی به حساب می‌آید؛ به همین ترتیب، هر صفت یکی از عوامل مؤثر در روابط پیکربندی محسوب می‌شود

7-5- هستان‌شناسی سوا

سوا اصول اساسی طراحی هستان‌شناسی خود را تمایزها، ترکیب‌ها، و قیدها می‌داند. مقوله‌بندی او در هستان‌شناسی سوا بیشتر با انگیزه فلسفی شناخته می‌شود. در این هستان‌شناسی سه تمایز سطح بالا وجود دارند:

- اولین تمایز، مادی یا فیزیکی در مقابل اطلاعات، یا عینی در مقابل مجرد است. این تمایز تقسیمی نه چندان منسجم است که برای تمام مقولات هستان‌شناسی مطرح است.
- دومین تمایز، اولویت ترتیبی است؛ مثلاً اول بودن در مقابل دوم بودن در مقابل سوم بودن و مانند آن؛ یا فرم در مقابل نقش در مقابل وساطت. این مقوله‌ها متقابلاً مستقل و جدا از هم نیستند. به طور مثال یک زن فرم (اول بودن) در نظر گرفته شده است، زیرا بدون در نظر گرفتن چیزی دیگر جدای از شخص، قابل تعریف است. به عنوان یک مادر، یک آموزگار، یا یک کارمند، همان شخص می‌تواند نمونه‌ای از یک نقش (دوم بودن) باشد. این نقش‌ها، هر فرد را در رابطه با گونه‌ای دیگر



	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			

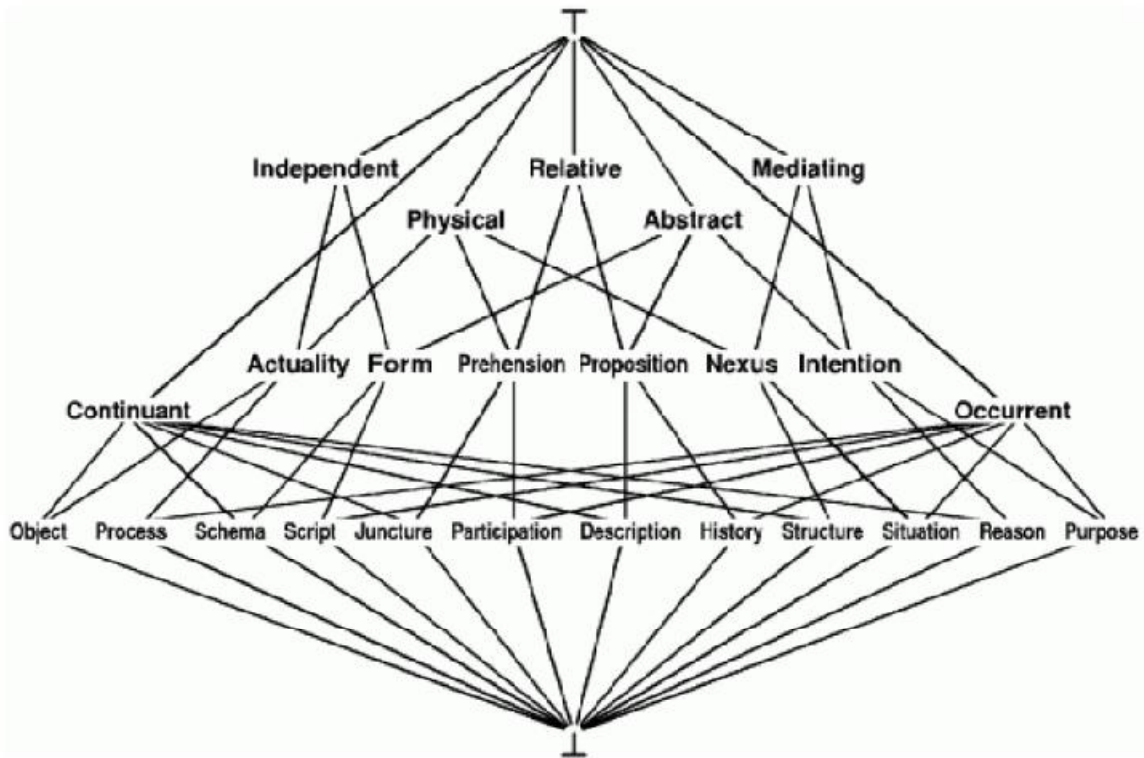
(کودک، دانش آموز، یا کارفرما) معرفی می‌کنند. ازدواج نوعی عامل میانجی یا مقوله‌ی وساطت (سوم بودن) است زیرا چندین گونه و در اینجا، دو گونه را به یکدیگر ربط می‌دهد.

- سومین تمایز، پیوسته در مقابل اتفاقی یا رخدادی، یا شیء در مقابل فرایند است. پیوسته‌ها اشیایی هستند که هویت خود را در یک دوره‌ی زمانی حفظ می‌کنند و مستمر هستند؛ اما اتفاقی‌ها، فرایندهایی هستند که فرم آنها در حال تغییر و در جریان است. به طور مثال، بهمن یک فرایند است و یخچال یک شیء. توجه به این نکته مهم است که این تمایز به مقیاس زمانی وابسته است. در مقیاس بزرگ چند قرن، یخچال هم یک فرایند خواهد بود.

این تمایزها در هستان‌شناسی سوا ترکیب می‌شوند تا مقولات جدیدی را بسازند. در سطحی پایین تر، به طور مثال، یک نوشته (مثلاً یک برنامه‌ی کامپیوتری یا دستورالعملی برای پخت نان) یک فرم است که دنباله‌هایی را معرفی می‌کند و در نتیجه به عنوان مجرد، فرم، و فرایند معرفی می‌شود. همچنین، تاریخ (مثلاً اجرای برنامه‌ای کامپیوتری) که گزاره‌ای در توصیف دنباله‌ای از فرایندها است، با این مقدمات، مجرد، فرم و شیء در نظر گرفته می‌شود.

در سطح بالا، برای هر ترکیب ممکن از تمایزها یک مقوله وجود دارد. اما برای جلوگیری از ترکیبات بیش از حد پر تعداد، در سطوح پایین تر محدودیت‌هایی را تعریف کرده‌اند تا مقوله‌هایی که نمی‌توانند وجود داشته باشند، قابل تعریف نباشند. محدودیت‌هایی منطقی می‌توانند مصادیقی مانند مثلث‌های دارای چهار ضلع را باطل سازند، محدودیت‌های تجربی، نیز می‌توانند مصادیقی مانند پرندگان سخنگو را منتفی سازند. محدودیت‌ها به صورت اصول بدیهی یا متعارف معرفی می‌شوند و از طریق سلسله مراتب به سطوح پایین تر منتقل می‌شوند. این انتقال از نوع وراثتی است. شکل 8- سلسله مراتب مفاهیم سطح بالا در سوا را نشان می‌دهد.



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



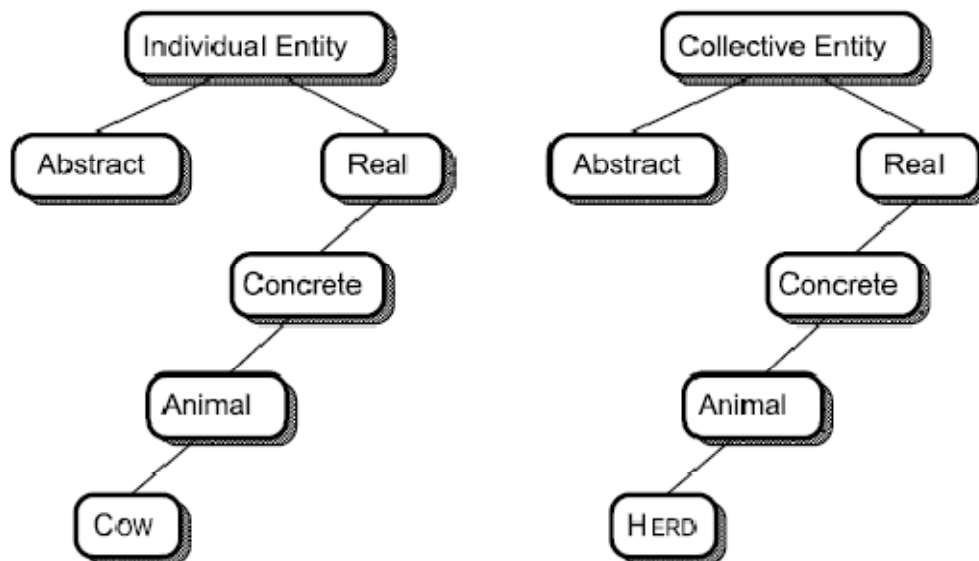
شکل 8- سلسله مراتب مفاهیم سطح بالا در سوا

7-6- هستان‌شناسی دالگرن

کاتلین دالگرن [51] هستان‌شناسی خود را براساس انگیزه زبان شناختی طراحی کرد. یکی از نکات عمده‌ای که دالگرن در کار خود مطرح کرد این است که هستان‌شناسی باید ارتباطات نزدیکی با زبان داشته باشد. هستان‌شناسی او نیز از همین رو، مبتنی بر تحقیقات گسترده زبان شناختی است. اصل ساختاری اولیه در هستان‌شناسی او، سلسله مراتب طبقات متقاطع (طبقه بندی متداخل) است؛ این تقاطع و تداخل به این معنی است که در هر گره از رده بندی این هستان‌شناسی، شاخه سازی در چندین بعد صورت می‌پذیرد. به طور مثال در سطح بالا، هر موجودیتی یا مجرد است یا واقعی و هر کدام از آنها نیز یا فردی است یا جمعی. این شاخه سازی همان طبقه بندی متداخل بالا است. این تقسیمات را اساساً موازی یکدیگر فرض می‌کنند. بنابراین، بعضی چیزها فردی و واقعی هستند (مثل یک حیوان، مثلاً گوسفند) و بعضی چیزها نیز جمعی و واقعی هستند (مثل گل). لازم است به این نکته توجه شود که تمام ارتباطات در

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



این هستان‌شناسی از یک نوع نیستند؛ همین موضوع باعث می‌شود این ساختار در هستان‌شناسی تا حدی گیج‌کننده و ضد و نقیض باشد؛ به طور مثال، اغلب رابطه‌ها از نوع "هست یک" هستند، اما بعضی از روابط در بخش جمعی سلسله مراتب اینگونه نیستند. به عنوان مثال گله زیر حیوان قرار گرفته است اما رابطه آن از نوع "تشکیل می‌شود از" است. در سطح زیرین سلسله مراتب گره‌های پایانی قرار می‌گیرند که اهمیت نظری ندارند و به صورت متقاطع طبقه بندی نشده‌اند. به طور مثال، مفاهیمی مانند حیوان، نقش، شخص، همگی جزوه گره‌های پایانی هستند. پس از آن، اسم‌های خاص به گره‌های پایانی ربط داده می‌شوند. شاید از آنجا که تمامی اسم‌ها براساس کاربرد آنها در جملات مشخص (خاص) می‌شوند، طبقه بندی گره‌های پایانی، در بعضی موارد با ویژگی‌های معنایی واژه‌ها در تناقض قرار می‌گیرد. شکل 9- ساختار سلسله مراتب سطح بالای دالگرن را نشان می‌دهد.



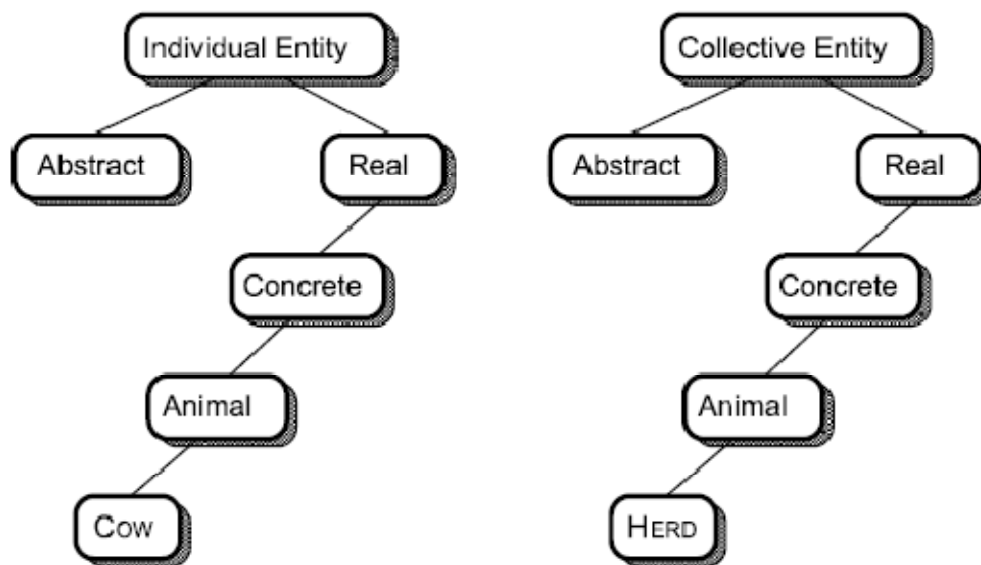
شکل 9- ساختار سلسله مراتب سطح بالای دالگرن

7-7 - سیستم زبانی واحد پزشکی

سیستم زبانی واحد پزشکی در کتابخانه ملی پزشکی، نمونه‌ای از سیستم‌های هستان‌شناسی تخصصی است. این سیستم برای تسهیل بازیابی اطلاعات و همچنین یکپارچه سازی اطلاعات منابع مختلف و



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

متعدد پزشکی - زیست شناختی طراحی شد. هدف پروژه، تسهیل و تقویت رابطه بین منابع مختلف و کاربرانی بود که از اصطلاحات متفاوتی استفاده می‌کردند. سیستم UMLS دارای سلسله مراتبی از مفاهیم است که شامل 135 مفهوم پزشکی است. این سیستم همچنین دارای شبکه‌ای معنایی است که روابط اضافه دیگری بین مقولات را معرفی می‌کند. این روابط اضافه از نوع رده بندی‌ها نیستند. سلسله مراتب مفاهیم در این هستان‌شناسی هم شامل موجودیت‌های مختلف (موجودیت‌های فیزیکی و مفهومی) است و هم رخدادهای را در برمی‌گیرد؛ رخدادهایی نظیر فعالیت، فرایند، یا مواردی مانند مصدومیت یا مسمومیت. در شکل 10، سطوح بالایی سلسله مراتب موجودیت‌ها در این هستان‌شناسی نشان داده شده است.



شکل 10 - سلسله مراتب در UMLS

روابط نمایش داده شده در این سیستم شامل روابطی معنایی هستند که نمونه‌هایی از آنها عبارتند از: به صورت فیزیکی مرتبط با، به صورت فضایی مرتبط با، به صورت کارکردی مرتبط با، به صورت زمانی مرتبط با و به صورت مفهومی مرتبط با. در این هستان‌شناسی، 51 رابطه مختلف مشابه موارد فوق تعریف شده‌اند، شبکه‌ای معنایی مفاهیم و تمام روابط امکان پذیر بین آنها بسیار بسیار بزرگ است. یکی از کاستی‌های عمده این هستان‌شناسی، این است که در آن دینامیک فرایندها در نظر گرفته نشده است و روابط دینامیک در مقایسه با روابط استاتیک بین فرایندها دیده نشده‌اند.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		کد زیر پروژه: پیک‌متن‌فارس - 3 - ت	ویرایش: 1/0
تاریخ: 1388/04/18			



7-8 - Ontoliguia

هستان‌شناسی ایجاد شده در دانشگاه اسنفورد می‌باشد. سرور این هستان‌شناسی، محل ذخیره هستان‌شناسی‌ها می‌باشد و این امکان برای کاربر مهیا می‌باشد که هستان‌شناسی جدید ایجاد و یا هستان‌شناسی موجود را توسعه بخشند. هستان‌شناسی ذخیره شده در سرور به فرمت‌های مختلف می‌تواند تبدیل شود. قابلیت استفاده مجدد از هستان‌شناسی و تغییر نام نشانه‌های استفاده شده در این هستان‌شناسی موجود می‌باشد [9]. یکی از مزایای سرور Ontoliguia دسترسی به کتابخانه‌ای از هستان‌شناسی‌های از قبل ایجاد شده می‌باشد. این کتابخانه با ایجاد هستان‌شناسی‌های جدید توسعه می‌یابد. بنابراین ایجاد هستان‌شناسی به صورت ماجولار می‌باشد. قابلیت استفاده مجدد از هستان‌شناسی‌های دیگر نیز وجود دارد.

7-9 - MESA

ابزاری برای ترکیب چندین منبع اطلاعاتی برای به وجود آوردن یک هستان‌شناسی واحد می‌باشد. کلمه‌های این هستان‌شناسی، کلمات ابتدایی و اصلی در هر حوزه‌ای می‌باشند. به عنوان مثال، در حوزه اقتصادی، کلمات مهم و ابتدایی شامل سهام و قیمت می‌باشد. این کلمات ابتدایی می‌توانند با استفاده از عملگرهای ترکیبی تشکیل کلمات پیچیده‌تر را نمایند. تمامی عملگرها به صورت منطقی می‌توانند نشان داده شوند. بنابراین امکان مقایسه کلمات پیچیده وجود دارد. به عنوان مثال می‌توان مشخص کرد که یک کلمه خاص شده کلمه دیگر می‌باشد. عملگرها توسط [52] پیشنهاد شده است. مقایسه بین دو مفهوم و توسعه داده‌شده آن دو مفهوم که بر اساس اعمال عملگرها به وجود آمده است، با استفاده از سیستم خبره انجام می‌شود. 4 ارتباط بین دو مفهوم در نظر گرفته شده است که به قرار ذیل می‌باشد:

- تساوی
- عمومیت بخشی
- اشتراک
- عدم تجانس

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

بعد از این کار، می‌بایست درستی مفاهیم و ارتباط بین مفاهیم به نحوی چک شود و اصلاحاتی انجام شود. اصلاحات از طریق خاص کردن یا عمومیت‌بخشی انجام می‌شود. خاص کردن از طریق انتخاب مفهوم دیگر از میان اصطلاحات موجود و یا به وجود آوردن اصطلاح جدید انجام می‌شود. مفاهیم جدید یا جایگزین دیگر مفاهیم می‌شوند و یا دوباره با عملگرها ترکیب می‌شوند و توسعه خواهند یافت. این کار چک کردن و اصلاح مفاهیم تکرار می‌شود تا زمانی که مفاهیم و ارتباط بین آن‌ها به هماهنگی برسد. جدول 1 نحوه چک کردن ارتباط بین مفاهیم را نشان می‌دهد.



جدول 1- چک کردن ارتباط بین مفاهیم

Semantic Inter-Correspondences				
	$C_1 = C_2$	$C_1 \subset C_2$	$C_1 \cap C_2$	$C_1 \neq C_2$
$L(C_1) \equiv L(C_2)$	OK	specialize $L(C_1)$ or generalize $L(C_2)$	specialize $L(C_1)$ and $L(C_2)$	error
$L(C_1) \subset L(C_2)$	specialize $L(C_2)$ or generalize $L(C_1)$	OK	specialize $L(C_2)$ or generalize $L(C_1)$	error
$L(C_1) \cap L(C_2)$	specialize $L(C_1)$ and $L(C_2)$	specialize $L(C_1)$ or generalize $L(C_2)$	OK	disjoint $L(C_1)$ and $L(C_2)$
$L(C_1) \cap L(C_2)$	error	error	disjoint $L(C_1)$ and $L(C_2)$	OK

7-10- سازمان مجازی تورنتو

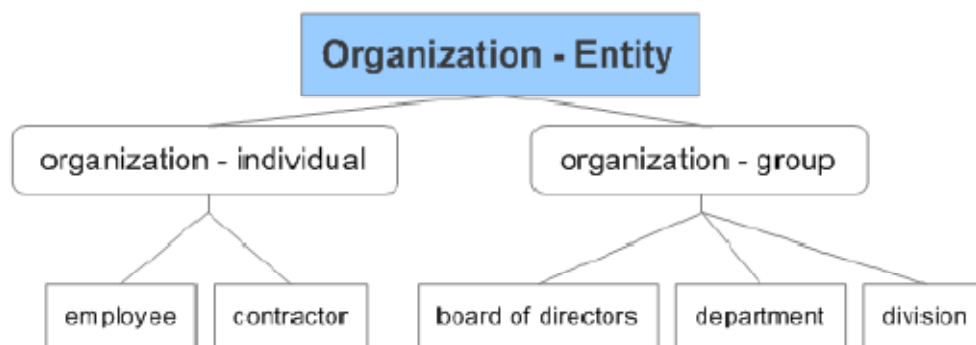
نمونه دیگری از هستان‌شناسی‌های تخصصی (TOVE) [15] پروژه سازمان مجازی تورنتو است که برای کاری خاص طراحی شده است. هدف این پروژه ایجاد مدل‌هایی سازمانی است که می‌توانند به پرسش‌های مربوط به اطلاعات صریح موجود در مدل پاسخ دهند و همچنین می‌توانند برای پاسخ به پرسش‌ها استنتاج کنند. TOVE مراحل ذیل را در مهندسی هستان‌شناسی انجام داده است:

- سناریوی هدفمند: نقطه شروع تولید هستان‌شناسی مجموعه مسائل و مشکلات مورد مواجهه می‌باشد.
- مطرح کردن سوالات غیر فرمال: نیازهای تولید هستان‌شناسی بر اساس سناریوها به صورت سوال‌هایی مطرح می‌شود و هستان‌شناسی می‌بایست به این سوالات بتواند پاسخ دهد.
- توصیف اصطلاحات: اشیاء، خصوصیات و ارتباطات بین آن‌ها به صورت فرمال توصیف می‌شود.



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- مطرح کردن سوالات فرمال: نیازهای تولید هستان‌شناسی بر اساس توصیف فرمالی که از اصطلاحات و لغات انجام شده‌است، در قالب سوالات فرمال انجام می‌شود.
- توصیف منطق: تعریف منطق‌هایی (منطق مرتبه اول) که تعریف لغات و تفسیر آن‌ها را انجام می‌دهد. از مرحله 4 برای ایجاد این منطق‌ها استفاده می‌شود.

پروژه TOVE از رویکردی فرمال (صوری) به فرایند مهندسی هستان‌شناسی و فرایند ارزیابی هستان‌شناسی، استفاده می‌کند. اول آن که از طریق تعامل با شرکای صنعتی خود، مهندسان هستان‌شناسی مسائلی را که در سازمان‌های واقعی مطرح می‌شوند، تعیین می‌کنند. به این ترتیب، پرسش‌هایی که هستان‌شناسی باید قادر به پاسخ‌گویی به آنها باشد، شکل می‌گیرد؛ همین پرسش‌ها، انتخاب مفاهیم و روابط در هستان‌شناسی، از بین گزینه‌ها، را توجیه می‌کنند. پس از آن، پرسش‌های توانمندی و تمام مفاهیم و اصول بدیهی در قالب منطق رتبه اول، فرمال می‌شوند؛ همچنین، قضایای کامل بودن برای پرسش‌های توانمندی، براساس همین شناخت واژگانی از اصطلاحات اثبات می‌شوند. در این سیستم، یک هستان‌شناسی منفرد مطرح نیست، بلکه مجموعه‌ای از چندین هستان‌شناسی برای بخش‌های منطقی مختلف در مدل‌سازی وجود دارند که به وسیله روابطی به یکدیگر، متقابلاً مرتبط می‌گردند. این هستان‌شناسی‌ها برای فعالیت‌ها، حالت‌ها (که نمایشی از زمان را نیز در خود دارند)، محصولات، سازمان، و مدیریت هزینه‌ها براساس فعالیت، هستند. در داخل هر یک از این هستان‌شناسی‌ها، تعدادی سلسله مراتب کوچک (با عمقی در حد دو یا سه سطح) وجود دارند که خوشه‌های کوچک دانش را معرفی می‌کنند. اصول بدیهی و روابط در این هستان‌شناسی برای مرتبط ساختن دانش از خوشه‌های مختلف به کار می‌روند. شکل 11- سلسله مراتب موجودیت - سازمان در TOVE را نشان می‌دهد.



شکل 11- سلسله مراتب موجودیت - سازمان در TOVE

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

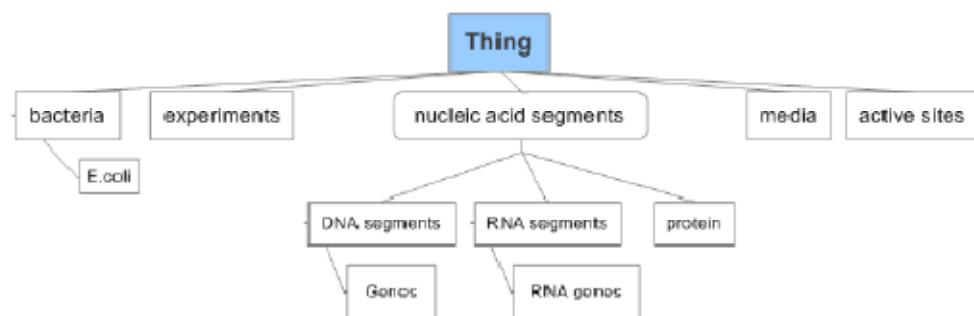
GENSIM -11-7

سیستمی برای شبیه سازی ژنتیک است که واکنش‌های بیوشیمیایی کاتالیز شده با آنزیم‌ها را مدل می‌کند؛ این واکنش‌ها تحت تأثیر مولکول‌های بزرگی انجام می‌پذیرند که دارای ساختارهای درونی پیچیده‌ای هستند DNA و RNA از این نوع مولکول‌ها هستند GENSIM دارای دو زیرهستان‌شناسی است که به پایگاه‌های دانش مشهورند:

(1) پایگاه دانش طبقات



(2) پایگاه دانش فرایندها

سطح بالای سلسله مراتب پایگاه دانش طبقات در این هستان‌شناسی در شکل زیر نشان داده شده است. این سطح شامل مفاهیم (شیای) بیوشیمیایی نظیر ژن‌ها، پروتئین‌ها، و پیوندگاه‌های بیوشیمیایی است طبقات عمومی مفاهیم نیز در شبیه سازی‌ها به کار می‌روند. در GENSIM، هر مفهوم یا شیء، مانند پروتئین‌ها، معرف یک مولکول واحد نیست، بلکه جمعیتی متجانس از این نوع مولکول‌ها را نشان می‌دهد. شکل 12 ساختار سطح بایلی GENSIM را نشان می‌دهد.



شکل 12- هستان‌شناسی سطح بالای GENSIM

پایگاه دانش فرایندها رفتار بالقوه مفاهیم در سیستم را توصیف می‌کند. طراح هستان‌شناسی از یک نظریه فرایندی کیفی برای معرفی فرایندها استفاده کرده است و در آن هر فرایند، دارای چهار چوبی مفصل است که نه تنها پارامترهایی نظیر ورودی و خروجی واکنش‌ها را شامل می‌شود، بلکه شامل شرایط اولیه و

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

مقدماتی برای آغاز یک واکنش و تأثیرات آن واکنش نیز هست. پس از آن، این اطلاعات در شبیه‌سازی واکنش‌های مبتنی بر شرایط مشخص آغازی به کار می‌روند.

SENSUS -12-7

این هستان‌شناسی بر اساس این فرضیه می‌باشد که اگر دو پایگاه دانش روی یک هستان‌شناسی ایجاد شده باشند، دانش بین آن‌ها می‌تواند به صورت خواناتری به اشتراک گذاشته شود، زیرا دارای زیر ساخت مشترکی می‌باشند. ابتدا دو هستان‌شناسی مطابق ذیل تشخیص داده می‌شود [21].



1) هستان‌شناسی حوزه: مجموعه‌ای از لغات برای توصیف حوزه خاص می‌باشد.

2) هستان‌شناسی تئوری: مجموعه‌ای از مفاهیم برای نمایش بعضی از جنبه‌های دنیا (همانند زمان، فضا و علت - معلول‌ها می‌باشد)

سیستم SENSUS هستان‌شناسی با پوشش بالا می‌باشد که شامل عبارات سطح بالا و میانی می‌باشد. از ادغام WordNet و Onto و طبقه‌بندی‌های معنایی در لغت‌نامه‌های الکترونیکی به وجود آمده‌است. در طراحی هستان‌شناسی تخصصی بعضی از عبارات انتخاب می‌شوند تا نمایانگر مفاهیم مربوط به یک حوزه خاص باشند. تمامی مفاهیم موجود از این عبارات تا ریشه SENSUS در هستان‌شناسی نهایی موجود می‌باشند. دیگر کلمات همانند طبقه‌بندی‌های مفاهیم مرتبط و تمامی گره‌های زیر درخت مربوط به عبارات خاص یک حوزه تا ریشه نیز در هستان‌شناسی نهایی موجود می‌باشند.

7-13- پلینیوس

پروژه پلینیوس [53] با هدف استخراج نیمه خودکار دانش از متن در حوزه علم مواد، و بخصوص مواد سرامیک، طراحی شد. مفهوم سازی ترکیب شیمیایی مواد در هسته و مرکز هستان‌شناسی پلینیوس قرار دارد. هستان‌شناسی فرایندها و خواص، در این مجموعه جدا از هستان‌شناسی مواد طراحی شده است. طراحان پلینیوس رویکردی به سازمان دهی هستان‌شناسی داشته‌اند که با رویکرد سنتی سلسله مراتبی - بدیهی (اصل موضوعی) متفاوت است. آنها این مجموعه را مجموعه ساخت مفهومی می‌نامند: هستان‌شناسی آنها با مجموعه‌هایی از مفاهیم تجزیه‌ناپذیر آغاز می‌شود. مفاهیمی مانند عناصر شیمیایی،



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

اعداد حقیقی، و حالت‌های مواد (گاز، مایع و مانند آن) که به عنوان مفاهیم بدوی عمل می‌کنند. پس از آن، عناصر این مجموعه‌ها را ترکیب می‌کنند تا تمام مفاهیم دیگر تعریف شوند. این ترکیب با قواعدی در ساخت تعریف می‌کنند. تا تمام مفاهیم دیگر تعریف شوند. این ترکیب با قواعدی در ساخت تعریف می‌شود که عملاً راه‌های عملی ترکیب سازی‌ها را تحت کنترل خود دارند. قواعد ساخت، چگونگی ساخته شدن مواد و ساختهای پیچیده از مفاهیم تجزیه ناپذیر را تعریف می‌کنند. برای گروه‌ها، مواد شیمیایی، و مراحل ساخت مواد شیمیایی با نسبت‌های ترکیبی نسبی از عناصر سازنده آنها قواعدی تعریف شده‌اند. پس از آن نظامی برای رده بندی تعریف می‌شود که به صورت ضمنی و با استقراء چگونگی گنجاندن مفاهیم در رده‌های خاص را معین می‌سازد. نمی‌توان به سادگی درباره چگونگی بسط این ساختار در حوزه‌هایی غیر از شیمی درباره چگونگی نمایش فرآیندها و خواص در این ساختار اظهار نظر کرد.

7-14- قالب تبادل دانش

KIF [54] زبانی برای تعریف هستان‌شناسی‌ها است. این زبان امکان تعریف اشیاء، کارکردها، و روابط را فراهم می‌آورد قالب مبادله دانش (KIF) معنا شناسی اخباری است و مبتنی بر حساب محمولات مرتبه اول طراحی شده است. این زبان می‌تواند نمایش فرادانش را میسر سازد و به کاربر اجازه می‌دهد که قواعد استدلال غیر یکنواخت را نمایش دهد. KIF را می‌توان یک هستان‌شناسی نیز در نظر گرفت، زیرا در بردارنده دیدگاهی معین از جهان نیز هست. با این که سطح بازنمایی آن از نظر جزئیات در ردیف هستان‌شناسی‌های دیگری که در اینجا معرفی کردیم، نیست. اما به طور مثال، جزئیاتی نیز مانند تئوری‌هایی در سطح خرد برای اعداد، مجموعه‌ها، و فهرست‌ها در آن به صورت اصول بدیهی پرورانده شده‌اند.



به عنوان یک قالب تبادل، KIF از نظر ارائه ویژگی‌های هستان‌شناسی‌ها، دشواری‌هایی در کاربرد دارد. اما سیستم‌هایی در سطوح کلان آن ساخته شده‌اند که اجازه می‌دهند هر هستان‌شناسی، با اصطلاحات آشنایی نظیر طبقات، روابط و، مانند آنها، توصیف و تبیین شود. سیستم Ontolingua نمونه‌ای از این سیستم‌ها است که مجموعه‌ای از ابزارها برای تجزیه و تحلیل و ترجمه هستان‌شناسی‌ها است. آزمایشگاه سیستم‌های دانش در دانشگاه استنفورد که این سیستم را به وجود آورده شامل تعدادی هستان‌شناسی است که با Ontolingua طراحی شده‌اند. همین وبگاه ویراستاری ساده نیز دارد که با استفاده از آن می‌توان

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

هستان‌شناسی‌های مورد نیاز را تولید کرد. این ویراستار، همچنین به کاربر اجازه می‌دهد، هستان‌شناسی‌های موجود را از دادگان دانشگاه انتخاب کرده و در هستان‌شناسی مورد نظر خود بگنجانند.

OntoLearn -15-7

در این روش، ابتدا اصطلاحات و لغات از روی مجموعه‌ای از متون استخراج می‌شود [55]. این مجموعه از متون شامل صفحات وب خاص و دیگر اسناد می‌باشد. پس از آن، این مجموعه اصطلاحات با استفاده از روش‌های پردازش زبان طبیعی و تکنیک‌های آماری که تحلیل‌های مقایسه‌های روی حوزه‌های مختلف انجام می‌دهند، فیلتر می‌شود. پس از آن از پایگاه دانش لغوی همانند WordNet استفاده می‌نماید تا کلمات را از نظر معنایی تفسیر نماید. پس از آن، مجموعه مفاهیم را بر اساس انواع ارتباطات موجود به یکدیگر مرتبط می‌نماید و جنگلی از مفاهیم مربوط به یک حوزه ایجاد می‌شود. از روش‌های بر پایه قاعده نیز برای استخراج این ارتباطات استفاده می‌شود. پس از آن این جنگل مفاهیم با WordNet ترکیب شده و یک هستان‌شناسی خاص یک حوزه به وجود می‌آورد.

	عنوان پروژه:		
	فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه:		
	تاریخ: 1388/04/18	ویرایش: 1/0	کد زیر پروژه: پیک متن فارسی - 3 - ت



8 - نحوه ارزیابی هستان‌شناسی

برای ارزیابی سیستم‌های تولید هستان‌شناسی خودکار، دو روش پیشنهاد شده است:

- ارزیابی روش یادگیری
- ارزیابی نتیجه حاصله (کاربرد)

معمولاً ارزیابی نتیجه از طریق مقایسه دو یا چند هستان‌شناسی مدل‌سازی شده در یک قلمروی خاص با تکنیک‌های ارزیابی ضربدری [31] و یا از طریق کاربردی که مورد استفاده قرار می‌گیرد، انجام می‌شود. ارزیابی مبتنی بر کاربرد بیشتر برای هستان‌شناسی‌های خاص منظوره مناسب می‌باشد. اکثر سیستم‌های یادگیر، عمل ارزیابی را از طریق مقایسه حاصل کار با نتایج مورد نظر فرد خبره و یا با محاسبه معیارهای ارزیابی متداول در استخراج اطلاعات (مثل دقت و توجه) انجام می‌دهند. معیار دقت، نمایشگر نسبت نتایج (مثلاً مفاهیم استخراج شده) صحیح به کل نتایج موجود در مجموعه آزمون، و معیار توجه، نشانگر نسبت نتایج صحیح به کل نتایج استخراج شده است.



پوشش مفهومی، میزان مفید بودن هستان‌شناسی در کاربردهای مختلف آن جزو شاخص‌های مهم ارزیابی هستان‌شناسی هستند. یکی از راه‌های ارزیابی هستان‌شناسی‌ها نیز انتخاب پیکره‌ای از پیش تعیین شده است که می‌توان با استفاده از آن تعیین کرد که آیا تمامی اطلاعات موجود در پیکره با هستان‌شناسی ایجاد شده پوشش داده می‌شوند یا خیر؟ TOVE [56] فرایندی رسمی را برای ارزیابی هستان‌شناسی خود اجرا کرد. این فرایند شامل مراحل به شرح زیر بود: پرسش‌های توانمندی به صورت فرمال معرفی شدند. پس از آن قضایای کامل بودن با توجه به این پرسش‌ها ثابت شد. برای اثبات قضایا از منطق مرتبه اول و مرتبه دوم در بازنمایی مفاهیم، ویژگی‌ها و روابط استفاده شد. برای GENSIM که برای شبیه‌سازی‌ها طراحی شده است، ارزیابی به وسیله مقایسه واکنش‌های از پیش دانسته با خروجی‌های پیش‌بینی شده توسط برنامه، انجام داده شد. براساس گزارش مؤلفان، پیش‌بینی‌ها بی‌نقص بودند. البته، اغلب پروژه‌ها برنامه‌های کاربردی مختلفی را در نظر دارند که از هستان‌شناسی آنها استفاده می‌کنند و در نتیجه همین کاربران می‌توانند پوشش مفهومی و فایده عملی هستان‌شناسی را بیازمایند. لازم به ذکر است که GENSIM سیستم دارای شبیه‌سازی ژنتیک است که واکنش‌های زیست‌پزشکی کاتالیز شده با آنیزم‌ها را بازنمایی و مدل‌سازی می‌کند و TOVE هستان‌شناسی طراحی شده برای مدل‌سازی شرکت‌ها است به

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

صورتی که قابلیت استنتاج پاسخ به پرسش‌هایی درباره اطلاعات موجود در مدل رداشته باشد که در فصل 7 توضیح داده شد.

بنابراین به طور خلاصه، روشی که برای ارزیابی وجود دارد، بررسی کامل انجام دادن وظیفه محوله به هستان‌شناسی می‌باشد. شرط کامل بودن شامل دو زیر شرط ذیل می‌باشد:

- پوشش (آیا تمامی مفاهیم دلخواه پوشانده شده‌است؟)
- درشتی (granularity) (آیا تمامی تمایزهای مربوطه ایجاد شده است؟)

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



9 - نتیجه‌گیری

در طراحی هستان‌شناسی فارسی استناد به زبان طبیعی جزو مسائلی است که بخصوص در زبان فارسی حساسیت بالایی می‌یابد، چرا که منابع واژگانی محدودی برای تعریف به زبان طبیعی داریم و اغلب منابع موجود فارسی در حوزه‌های دانشی مختلف، از نظر توصیف و تعریف واژگان و اصطلاحات ضعف دارند یا اساساً موجود نیست. حتی برای هستان‌شناسی‌های عمومی نیز، دسترسی به منابع توصیفی اصطلاحات به زبان طبیعی به خوبی میسر نیست. به این ترتیب یکی از پیش نیازهای تولید هستان‌شناسی‌ها تدارک همین مرجع استنادی برای تعریف و توصیف واژه‌ها به زبان فارسی است. همین ضعف باعث شده است که در کارهای انجام شده در زبان فارسی، روی طیف دادگان تا هستان‌شناسی، بیشتر در حوزه دادگان بمانیم و چندان موفقیتی در تولید هستان‌شناسی فارسی به دست نیامده باشد، چرا که اجماع نظر اهل فن در آن حوزه که جزو شرایط لازم تحقق هستان‌شناسی است، بدون استفاده از تعریف و توصیف محکم در (زبان طبیعی) فارسی، دشوار است و معمولاً تیم‌های کارشناس و مهندسی دست اندرکار تهیه تعریف‌های فرمال، یا نمی‌توانند توصیف و تعریف‌های زبان فارسی را تهیه کنند و اگر در مواردی هم این کار را انجام داده باشند، به مجموعه‌ای از تعاریف قابل استناد و قابل اجماع نرسیده‌اند.

بررسی هستان‌شناسی‌های موجود نشان می‌دهد که همگی در موارد ذیل با یکدیگر اشتراک دارند [57]:



- تمامی هستان‌شناسی‌ها یک کاری را به عنوان نقطه شروع انجام می‌دهند. به عنوان مثال کار جمع‌آوری داده را انجام می‌دهند، شرحی از توانایی‌های هستان‌شناسی مشخص می‌نمایند
- بیشتر هستان‌شناسی‌ها مساله فابلیت استفاده مجدد را در نظر نگرفته‌اند و در این زمینه محدودیت‌هایی دارند. سوالاتی همانند اینکه آیا می‌توان هستان‌شناسی را به اشتراک گذاشت مناسب می‌باشد.
- این انتظار در مورد هستان‌شناسی‌ها وجود دارد که ابتدا یک هستان‌شناسی پایه ایجاد شود و سپس آن توسعه داده شود و دوباره استفاده شود.

اصول تولید هستان‌شناسی، شفافیت، انسجام، توسعه‌پذیری و تولید هستان‌شناسی کمینه است که می‌بایست در نظر گرفته شود. ایجاد هستان‌شناسی بیشتر از آن که مهندسی باشد یک مهارت می‌باشد. در زبان فارسی، هستان‌شناسی که به طور فراگیر و جامع باشد وجود ندارد. تولید هستان‌شناسی به

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	



روش‌های اتوماتیک و نیمه اتوماتیک مستلزم انجام پردازش روی متون فارسی است که با چالش‌های مختص زبان فارسی روبرو می‌باشیم. علاوه بر آن وجود استثنائات فراوان در ترکیب جمله، ایجاد الگوهای معنایی کامل برای استخراج مفاهیم از متن را دچار مشکل نموده‌است.

با توجه به آن‌چه که در فصول قبل انجام شد، نیاز به زیر ساخت‌های لازم همانند ریشه‌یاب، تحلیل‌گر نحوی، تحلیل‌گر لغوی و معنایی برای ایجاد خودکار هستان‌شناسی از روی متون می‌تواند جزو ابتدایی‌ترین فعالیت‌های لازم در امر هستان‌شناسی باشد. مساله دیگر وجود مجموعه متون پایه و استاندارد می‌باشد که بتوان بر اساس آن هستان‌شناسی پایه‌ای در حال حاضر موجود نیست. به نظر می‌رسد که تا کنون تولید هستان‌شناسی فارسی مورد توجه چندانی نبوده‌است و راه زیادی در تولید هستان‌شناسی فارسی باقی است. به عنوان قدم اول می‌توان اقدام به تهیه مرجع استنادی برای تعریف و توصیف واژه‌ها به زبان فارسی و ایجاد توصیفات فرمال برای اصطلاحات نمود. علاوه بر آن برای هستان‌شناسی‌های خاص می‌بایست حوزه هستان‌شناسی مشخص شود و نیازمندی‌هایی که هستان‌شناسی می‌بایست به آن‌ها پاسخ دهد تعیین شود.



	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیر پروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

مراجع



- [1] ش. ف. م, "هستان شناسی: نمونه سازی در یک محیط درک متن فارسی, رساله دکتراي مهندسی کامپیوتر طراحی مدل یادگیر, "هوش مصنوعی, دانشکده مهندسی کامپیوتر و فناوری اطلاعات, دانشگاه صنعتی امیرکبیر, 1381
- [2] A. Sheth, and C. Ramakrishnan, "Semantic Web Technology in Action: Ontology Driven Information Systems for Search, Integration and Analysis," *IEEE Computer Society Technical Communittee On Data Engineering*, 2003.
- [3] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," in International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation, Italy, 1993.
- [4] P. E. V. D. VET, and N. J. I. MARS, *Bottom-up Construction of Ontologies: The Case of an Ontology of Pure Substances*, Department of Computer Science, University of Twente, Netherlands, 1995.
- [5] R. Mizoguchi, J. Vanwelkenhuysen, and M. Ikeda, "Task Ontology for reuse of problem solving knowledge, Towards Very Large Knowledge Bases," in Knowledge Building and Knowledge Sharing (KBKS'95), 1995, pp. 46-57.
- [6] A. Maedche, and S. Staab, "Semi-automatic Engineering of Ontologies from Text," in 12th International Conference on Software Engineering and Knowledge Engineering (SEKE'2000), 2000.
- [7] L. Gillam, M. Tariq, and K. Ahmad, "Terminology and the construction of ontology," *Terminology*, pp. 55-81, 2005.
- [8] K. Latifur, and F. Luo, "Ontology Construction for Information Selection," in IJDM, 2002.
- [9] T. Gruber, *Ontolingua: A Mechanism to Support Portable Ontologies*, Stanford University, Stanford, California, 1992.
- [10] D. Miller, "WordNet: An On-Line Lexical Database," *International Journal of Lexicography*, vol. 3, 1990.
- [11] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in ICDM, 2003, pp. 541-544.
- [12] S. Staab, J. Angele, S. Decker *et al.*, "AI for the web - ontology-based community web portals," in AAI, 2000, pp. 1034-1039.
- [13] "UMLS," <http://www.nlm.nih.gov/research/umls>.
- [14] D. B. Lenat, "CYC: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 8-33, 1995.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- [15] M. Fox, F. Belli, and F. J. Radermacher, "The TOVE Project: A Common-sense Model of the Enterprise," *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 25-34, 1992.
- [16] J. Sowa, *Knowledge Representation: Logical, Philosophical and Computational Foundations*, 1999.
- [17] M. Grüninger, and M. Fox, "Methodology for the design and evaluation of ontologies," in Skuce D (ed) *IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995, pp. 1-10.
- [18] M. Uschold, and M. Grüninger, "Ontologies: Principles, Methods and Applications," *Knowledge Engineering Review*, vol. 11, no. 2, pp. 93-115, 1996.
- [19] S. Nirenburg, V. Raskin, and B. Onyshkevych, "Apologiae Ontologiae," in *Memoranda in computer and cognitive science*, MCCS, 1995, pp. 95-281.
- [20] H. Eriksson, R. W. Ferguson, Y. Shahar *et al.*, "Automatic generation of ontology editors for knowledge-based system," 1999.
- [21] B. Swartout, R. Patil, K. Knight *et al.*, "Toward distributed use of large-scale ontologies," *Symposium on Ontological Engineering, Stanford*, 1997.
- [22] A. Farquhar, R. Fikes, and J. Rice, "The ontolingua server: Tool for collaborative ontology construction," *IJHCS*, vol. 46, no. 6, pp. 707-728, 1997.
- [23] J. Domingue, and O. Tadzebao, "WebOnto: Discussing, browsing, and editing ontologies on the Web," in *Proceedings of the Eleventh Workshop on Knowledge Acquisition, Modeling and Management, KAW*, 1998.
- [24] T. Yamaguchi, "Acquiring coceptual relations from domain-specific texts," in *IJCAI Workshop on ontology learning*, 2001.
- [25] G. Chalendar, and B. Grau, "SVETLAN: System to classify nouns," in *First Workshop on Ontology Learning (OL 2000)*, In Conjunction with the 14th European Conference on Artificial Inteligence, 2000.
- [26] N. F. Noy, and M. A. Musen, "PROMPT: Algorithm and tool for automated ontology merging and alignment," in *Seventeenth National Conference on Artificial Intelligence*, 2000.
- [27] H. Chapulsky, E. Hovy, and T. Russ, "Progress on an automatic ontology alignment methodology," 1997.
- [28] I. Ryutaro, T. Hideaki, and H. Shinichi, "Rule induction for concept hierarchy alignment," in *Proceedings of the 2nd Workshop on Ontology Learning at the 17th Int. Joint Conf. On AI(IJCAI)*. 2001.
- [29] D. Faure, C. Nedellec, and C. Rouveirol, *Acquisition of semantic knowledge using machine learning methods: The system ASIUM*, Universite Paris-Sud, 1998.
- [30] C. H. Hwang, "Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information," in *6th International Workshop on Knowledge Representation Meets Databases(KRDB,99)*, 1999.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- [31] A. Maedche, and S. Staab, "Ontology learning for the semantic Web," *IEEE Intelligent System*, vol. 16, no. 2, pp. 72-9, 2001.
- [32] M. Shamsfard, and A. A. Barforoush, "An introduction to hasti: An ontology learning system," in 6th IASTED International Conference on Artificial Intelligence and Soft Computing, (ASC,2002), 2002.
- [33] V. Kashyap, "Design and creation of ontologies for environmental information retrieval," in Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW, 99), 1999.
- [34] A. B. Williams, and C. Tsatsoulis, "An instancebased approach for identifying candidate ontology relations within a multi-agent system," in First Workshop on Ontology Learning (OL-2000), In Conjunction with the 14th European Conference on Artificial Intelligence (ECAI, 2000), 2000.
- [35] H. Suryanto, and P. Compton, "Learning classification taxonomies from a classification knowledge based system," *Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI, 2000*, 2000.
- [36] N. Pernelle, M. C. Rousset, and V. Ventos, "Automatic construction and refinement of a class hierarchy over semi-structured data," *Workshop on Ontology Learning(OL,2001)*, 2001.
- [37] G. Heyer, M. Lauter, U. Quasthoff *et al.*, "Learning relations using collocations," *2001, Workshop on Ontology Learning*, 2001.
- [38] L. Gillam, and M. Tariq, "Ontology Via Terminology," 2003.
- [39] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," in Fourteenth International Conference on Computational Linguistics, 1992.
- [40] h. Assadi, "Knowledge acquisition from texts: Using an automatic clustering method based on nounmodifier relationship," *35th Annual Meeting of the Association for Computational Linguistic*, 1997.
- [41] J. M. Anderson, "The grammar of case: Towards a localistic theory," *Cambridge: Cambridge University Press*, 1971.
- [42] C. Pollard, and I. A. Sag, *Head-driven phrase structure grammar*: Chicago: University of Chicago Press, 1994.
- [43] F. R. Palmer, *Grammatical roles and relations* Cambridge, Cambridge University Press, 1994.
- [44] B. Aarts, *English syntax and argumentation*: London: Macmillan Press, 1997.
- [45] A. Radford, "Syntactic theory and the structure of english: A minimalist approach," *Cambridge Text books in Linguistics Cambridge, Cambridge University Press*, 1997.
- [46] M. Baker, and L. Haegemann, "Thematic roles and syntactic structure, Elements of grammar," *Handbook in generative syntax, Dordresht; Kluwer*, pp. 73-137, 1997.

	عنوان پروژه: فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی		
	عنوان زیرپروژه: تحلیل نیازمندی‌های تولید هستان‌شناسی‌های عمومی و تخصصی برای زبان فارسی		
	تاریخ: 1388/04/18	ویرایش: 1/0	

- [47] G. Smith, "Encoding thematic roles via syntactic functions in a german treebank,," *Workshop on Syntactic Annotation of Electronic Corpora*, 2000.
- [48] ش. ف. م.، ع. ز. ب. ا.، " واژگان محاسباتی : ساختار مرکزی در سیستم های پردازش زبان طبیعی " مجله امیرکبیر، سال 12، شماره 48، 1380
- [49] م. ش. فرد،، ا. ع. ز. بارفروش، "استخراج دانش مفهومی از متن با استفاده از الگوهای زبانی و معنایی،" تازه های علوم شناختی. 1381، سال 1، شماره 4.
- [50] M. Hearst, "Automatic acquisition of hyponyms from large text corpora,," in *International Conference on Computational Linguistics*, 1992.
- [51] K. Dahlgren, "Naive Semantics for Natural Language Understanding,," Kluwer Academic, 1988.
- [52] H. Wache, "An Integration method for the specification of rule-oriented mediators,," in *International Symposium on Database Applications in Non Traditional Environment*, 1999.
- [53] v. d. Vet, P. E, P. H. Speel *et al.*, "The PLINIUS Ontology of Ceramic Materials,," in *Eleventh European Conference on Artificial Intelligence (ECAI'94) Workshop on Comparison of Implemented Ontologies*, 1994, pp. 8-12.
- [54] NCITS. "Draft proposed American National standard for Knowledge Interchange Format. ," <http://logic.stanford.edu/kif/dpans.html>.
- [55] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and its application to Automated Terminology Translation,," *IEEE Intelligent System*, pp. 22-31, 2004.
- [56] TOVE. "TOVE Manual,," www.ie.utoronto.ca/EIL/tove/ontoTOC.html.
- [57] D. JONES, T. BENCH-CAPON, and P. VISSER, "methodologies for ontology development,," in *IFIP*, 1998.