دکتر مهرنوش شمس فرد، دانش آموخته‌ی کارشناسی و کارشناسی ارشد رشته‌ی مهندسی نرم افزار از دانشگاه شریف و دکتری هوش مصنوعی از دانشگاه امیرکبیر است. وی از سال ۱۳۸۳ با عنوان استادیار و مسئول آزمایشگاه پردازش زبان طبیعی در دانشگاه شهید بهشتی مشغول به فعالیتهای آموزشی- پژوهشی در زمینه‌های پردازش زبان طبیعی، مهندسی هستان‌شناسی ،کاوش متن و شبکه‌های معنایی است.

مهسا عرب یارمحمدی، دانش‌آموخته‌ی مهندسی نرم افزار از دانشگاه صنعتی امیرکبیر و کارشناسی‌ارشد همان رشته از دانشگاه شهید بهشتی است. عمده فعالیت پژوهشی او در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و آزمایشگاه سیستمهای هوشمند دانشگاه صنعتی امیرکبیر بوده است که حاصل آن چاپ چند مقاله داخلی و بین‌المللی می‌باشد.

[10] P. Vossen, "Euro WordNet General Document" Euro WordNet Project LE2-4003 & LE4-8328 Report, University of Amsterdam, 2002••

[11] D. Tufis, D. Cristea,and S. Stamou ,"BalkaNet: aims,methods,results and perspectives: A general overview," Romanian Journal on Information science and Technology, Vol. 7,Nos.1-2 ,pp.9-43, 2004.

[12] J. Morato, M. Á. Marzal, J. Lloréns, and J. Moreiro,"WordNet Applications," Petr Sojka, Karel Pala, Pavel Smr,Christiane Fellbaum, Piek Vossen (eds.) Proceedings of 2nd GWC, Brno, Masaryk University, pp. 270-278, 2004.

[13] K. Lind´en, Word Sense Discovery and Disambiguation. University of Helsinki, Department of General Linguistics, Finland, 2005.

[14] D• Jurafsky, and J• H. Martin, Speech and Language processing, Prentice-Hall, Inc. USA,2000.

[15] D. Tufiş, R. Ion, L. Bozianu, A. Ceauşu , and D. Ştefănescu "Romanian WordNet: Current State, New Applications and Prospects," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp.441-452, 2008.

[16] K. Kerner," Proposing Methods of Improving Word Sense Disambiguation for Estonian," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp. 229 -238, 2008.

[17] M. A. Yarmohammadi, M. Shamsfard, M. A. Yarmohammadi, and M. Rouhizadeh, "Using WordNet in Extracting the Final Answer from Retrieved Documents in a Question Answering System," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp. 520-•••, 2008.

[18] P. Clark, C. Fellbaum, and J. Hobbs, "Using and Extending WordNet to Support Question-Answering," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp. 111-119, 2008.

[19] B. J. Dorr, and K. Maria, "Lexical Selection for Cross-Language Application: Combining LCS with WordNet," Proceedings of the Third Conference of the Association for Machine Translation in the Americas, Langhorne ,pp.438-447, 1998

[20] W. Kang, S. Jungun, and K. Gilchang,"The semantic analysis of prepositional phrases in English-to- Korean machine translation using neural network," Journal of the Korean Information Science Society.21(11) pp.2118-2125. 1994a

[21] W. Kang, S. Jungun, K. Gilchang and C. Keysun, "A neural network method for the semantic analysis of prepositional phrases in English-to-Korean machine translation," Processing of Chinese and Oriental Languages, 8(2), pp.163-175, 1994b.

[22] K..Knight, and S.K. Luk," Building a large ontology for machine translation," Proceedings of the ARPA Human Language Technology Workshop, Princeton, USA, 1993.

[23] B. J. Dorr, "Large-scale dictionary construction for foreign language tutoring and interlingua machine translation," Machine Translation,12(1) ,pp.1-55, 1997.

[24] X. Hu,and A. Graesser, " Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialogue," Proceedings of ICCE'98, 2, Beijing, China Higher Education Press,pp.337- 341, 1998.

[25] R. Murphy, "A Bulglur in the Treasure House: Plumb Design's ' Visual Thesaurus'," Intelligent Agent, 2, P.3, 1998.

[26] J. Namrava, Using WordNet Glosses to Refine Google Queries, Dateso. April 26th-28th. 2006

[27] S.M. Assi,"Farsi Linguistic Database(FLDB) ," International Journal of Lexicography, Vol.10, No.3, p.6, Oxford University Press, 1997.

[28] M. Dabir-Moghadam, "Compound Verbs in Persian," Studies in the Linguistic Sciences, 27.2, pp. 25-59, 1997.

[۲۹] م. باطنی،و ف. آذر مهر،و م. مهاجر،و م. نبوی، فرهنگ معاصر انگلیسی به فارسی، ویراست دوم، تهران، موسسه‌ی فرهنگ معاصر، ۱۳۷۸.

[۳۰] ج. فراروی، فرهنگ طیفی طبقه بندی لغات و اصطلاحات فارسی، تهران، جمشید فراروی، ۱۳۷۸.

[۳۱] ف. خداپرستی، فرهنگ جامع واژگان مترادف و متضاد زبان فارسی، شیراز،دانشنامه‌ی فارس،۱۳۷۶.

[۳۲] ف. سلیمان پور، فرهنگ لغات فارسی به انگلیسی (الکترونیکی)، ویراست ۲.

[۳۳] ح. عمید، فرهنگ فارسی عمید، تهران، انتشارات امیر کبیر، ۱۳۶۲.

[۳۴] س. حییم، فرهنگ کوچک انگلیسی- فارسی سلیمان حییم، تهران، فرهنگ معاصر، ۱۳۷۵.

اکبر حسابی کارشناسی‌ارشد خود را در رشته‌ی آموزش زبان فارسی به غیرفارسی‌زبانان از دانشگاه علامه طباطبایی دریافت نمود. وی دانشجوی مقطع دکتری رشته‌ی زبان‌شناسی همگانی دانشگاه علامه طباطبایی است. فعالیتهای پژوهشی او در زمینه‌ی زبان‌شناسی , عصب‌شناسی زبان, ترجمه‌ی ماشینی و شبکه‌ی واژگانی است.

دکتر مصطفی عاصی، کارشناس‌ارشد زبان‌شناسی همگانی از دانشگاه تهران و دکتری زبان‌شناسی با گرایش رایانه و فرهنگ نگاری از دانشگاه اکستر انگلستان است. وی عضو هیات علمی فرهنگستان زبان ایران و پژوهشگاه علوم انسانی از سال ۱۳۵۰ تا کنون بوده و دانشیار و مدیر گروه زبان‌شناسی همگانی پژوهشگاه علوم انسانی و مطالعات فرهنگی است.

استفاده گردید (این کار در ساخت شبکه واژگانی آلمانی مکرر رخ داده است)[۳]. در بعضی موارد برای دسته‌ی هم معنا معادلی یافت نشد مانند {garmineous} یا { protoctist} که این موارد از خلاءهای واژگانی در زبان فارسی می‌باشد (البته ممکن است در آینده این خلاءها برطرف گردند).

۴. نکته‌ی بعد استفاده از گروه اسمی در ساخت دسته‌های هم معناست. در دسته‌های هم معنای شبکه‌ی واژگانی پرینستون در موارد بسیاری از گروه اسمی استفاده شده است (خصوصا اسم و صفت) و این مسئله در دسته‌های هم معنای زبان فارسی نیز به تبع اعمال گردید مثلا در دسته‌ی هم معنای{division, air division} که در فارسی به صورت { یگان، یگان هوایی} معادل سازی گردیده است و یا{ pressure, level,force per unit area pressure} که به صورت {فشار،سطح فشار،میزان فشار} معادل سازی گردید.

۵. یکی از نکات دیگر رابطه‌ی چند به یک بین دسته‌های هم معنای فارسی و انگلیسی است. مثلا دسته‌ی هم معنای {uncle} با دو دسته‌ی هم معنای {عمو،عم} و {دایی، خال، خالو} معادل سازی می‌شود و همچنین {aunt, auntie, aunty} با چهار دسته‌ی هم معنای {خاله}و{عمه}و{زن عمو} و{زن دایی} معادل سازی می‌گردد. از طرف دیگر دسته‌ی هم معنای { رئیس، رییس، روسا} با دسته‌های هم معنای{president}و{ president, chairman, chairwoman, chair, chairperson} و{persident,prexy} و{ chancellor} و{ chancellor, premier, prime minister} و { head, chief, top dog} و... معادل است. برای گروه اول تنها یک کد در زبان انگلیسی وجود دارد و در زبان فارسی چندین دسته‌ی هم معنا وجود دارد که باید همه با همان کد مشخص گردند. چون این مسئله منجر به ابهام می‌شود تنها راه حل در نظر گرفتن معادلهای مختلفی برای آن در زبان فارسی است یعنی دسته‌های هم معنای متفاوتی با کدهای متفاوتی در زبان فارسی که با یک دسته‌ی هم معنا انگلیسی معادل است. یعنی رابطه‌ی یک به یک تبدیل به رابطه‌ی چند به یک شود. برای این کار به کد انگلیسی حروف کوچک الفبا افزوده شد. برای مثال دسته‌ی هم معنای{uncle} در انگلیسی دارای کد ENG20-10035451 است که در فارسی برای دسته هم معنای {عمو،عم} کد ۱۰۰۳۵۴۵۱a و برای دسته‌ی هم معنای {دایی، خال، خالو} کد 10035451b در نظر گرفته شد.

۱۰- جمع بندی

از ابتدا طراحی و ایجاد شبکه‌ی واژگانی اسامی زبان فارسی این نکته مد نظر قرار داشت که کدهای شناسایی یکسانی برای دسته‌های هم معنای شبکه‌ی واژگانی اسامی زبان فارسی با دسته‌های هم معنای شبکه‌ی واژگانی پرینستون مورد استفاده قرار گیرد تا بدین طریق امکان تطبیق و مقایسه‌ی هر چه دقیقتر دسته‌های هم معنا با یکدیگر امکان پذیر باشد که بدین منظور از همان کدهای دسته‌های هم معنای شبکه‌ی واژگانی پرینستون بهره گیری شد. از این ویژگی می‌توان به عنوان نقطه‌ی قوتی جهت معادل یابی دقیق برای پروژه‌های ترجمه ماشینی بهره گرفت. از طرف دیگر با توجه به بهره گیری از ویرایشگر تطبیق یافته‌ی VisDic برای زبان فارسی و قابلیت صدور اطلاعات ذخیره شده در این ابزار در قالب XML، تعامل کاربردی بین این شبکه و دیگر شبکه‌های واژگانی (فعل و صفت فارسی و شبکه‌های واژگانی سایر زبانها) تضمین می‌گردد. با توجه به اینکه قسمتهایی از روابط میان دسته‌های هم معنا در پایگاه داده درج شده است، با تکمیل این روابط در کنار تعاریف و مثالهای مربوط به دسته‌های هم معنا، و همچنین افزودن حوزه و هستان‌شناسی رده بالا هسته‌ی یاد شده کاملتر می‌گردد. انتظار می‌رود گسترش (نیمه)خودکار با استفاده از منابع موجود زبان فارسی گامی بعدی در تحقق شبکه‌ی واژگانی کاربردی اسامی زبان فارسی باشد و سپس با الحاق شبکه‌های واژگانی صفات و افعال زبان فارسی، شبکه‌ی واژگانی زبان فارسی تحقق یابد. این شبکه با اتصال به شبکه‌های واژگانی زبانهای دیگر به ابزار مفیدی جهت انجام پژوهشهای پردازش زبان فارسی و پژوهشهای میان زبانی تبدیل خواهد شد.

تشکر و قدردانی

بخشی از این پژوهش تحت حمایت مالی مرکز تحقیقات مخابرات با قرارداد شماره ۵۰۰/۱۹۲۳۱/ت انجام شده است.

مراجع

[1] F. Keyvan, and H. Borjian, and M. Kasheff, and C. Fellbaum,"Developing PersiaNet: The Persian WordNet," Proceedings of the 3rd Global WordNet Conference, South Korea, pp.315-318, 2006.

[۲] ع. فامیان، بررسی و تحلیل روابط معنایی صفت برای طراحی شبکه‌ی واژگانی صفات زبان فارسی. رساله‌ی دکتری زبان‌شناسی همگانی، دانشگاه تربیت مدرس،۱۳۸۶.

[۳] م. روحی زاده، طبقه بندی افعال فارسی برای کاربرد در شبکه‌ی واژگانی زبان فارسی، پایان نامه‌ی کارشناسی ارشد زبان‌شناسی همگانی، دانشگاه علامه طباطبایی،۱۳۸۶.

[4] N. Mansoory and M.. Bijankhan " The Possible Effects of Persian Light Verb Constructions on Persian WordNet," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp. 297-303, 2008

[5] M.Shamsfard," Developing FarsNet: A Lexical Ontology for Persian," A. Tanács, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (eds.) Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, pp. 413-418, 2008a.

[6] M.Shamsfard, "Towards Semi Automatic Construction of a Lexical Ontology for Persian," LREC 2008 Proceedings, Morocco,2008b.

[7] A. Kilgarriff, "Review of WordNet: An electronic lexical database," Language, 76:706-708, 2000.

[8] G. Miller, R. Beckwith , C. Fellbaum, D. Gross, and K. Miller, "Five Papers on WordNet" CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[9] G. Miller,"Forward," WordNet: an electronic lexical database. Christiane Fellbaum (ed.),MIT Press, pp. xv-xxii, 1998.

[3] تبادل نظر با سازندگان شبکه‌ی واژگانی زبان آلمانی

که در آنها از مفهوم "دست" استفاده شده آورده می‌شود. این قسمت زمینه‌ی ارتباط میان مقوله‌های مختلف را فراهم می‌نماید.

صفات برجسته

در این رابطه برجسته ترین صفت مربوط به دسته‌ی هم معنا که در هنگام ساخت دسته‌ی هم معنا به ذهن زبانشناس می‌آمد آورده شده است، برای مثال برای دسته‌ی هم معنای { نمک} "شور" در نظر گرفته شد. این رابطه نیز زمینه‌ی ارتباط میان مقوله‌های اسم و صفت را فراهم می‌آورد.

نامعلوم

گاهی اوقات هیچ یک از رابطه‌های ذکر شده در بالا بین دو واژه یا دسته‌ی هم معنا وجود ندارد، که در آن صورت رابطه را به صورت نامعلوم ذکر می‌نماییم. می‌توان این رابطه را با هم آیی نیز نامید.

با ورود دسته‌های هم معنای کنونی شامل (معادلهای BC1,BC2 و مفاهیم پایه‌ی زبان فارسی) وافزودن روابط شمول و زیر شمول و بعضی از روابط دیگر،هسته‌ی شبکه‌ی واژگانی اسامی فارسی در حال کامل شدن است.

۹- نتایج و ارزیابی

بررسی کار انجام شده تاکنون به دو صورت کمی و کیفی امکان پذیر است. برای ارزیابی کمی می‌توان به اطلاعات آماری زیر در مورد شبکه‌ی واژگانی اسامی زبان فارسی که در زمان نوشتن این مقاله موجود بوده‌اند اشاره نمود:

جدول ۳: اطلاعات مربوط به شبکه‌ی واژگانی اسامی زبان فارسی

مقوله واژگانی	تعداد دسته‌های هم معنا	واژه ها	تعداد روابط
اسم	۳۸۳۱	۸۱۲۳	۸۵۵۸

جدول ۴: اطلاعات مربوط به روابط موجود بین دسته‌های هم معنا در شبکه‌ی واژگانی اسامی زبان فارسی

رابطه	تعداد
نزدیک به متضاد	۷۴
نزدیک به مترادف	۱
...جنس	۲
بخش واژه	۷۵
عضو واژه	۳۰
جزء واژه	۳۰
دارای بخش	۹۴
دارای واحد...	۱
دارای جنس...	۱
دارای عضو	۲۹
دارای جزء	۴۰
شامل	۴۲۷۳
زیرشمول	۳۹۰۷
تعدادکل روابط	۸۵۵۸

همانطور که دیده می‌شود تاکنون بیش از ۸۰۰۰ اسم در بیش از ۳۸۰۰ دسته هم معنا قرار گرفته اند. همانطور که گفته شد ادامه‌ی این پروژه به فاز اول پروژه فارس نت متصل خواهد شد. در این فاز انتظار می‌رود در حدود ۵۰۰۰ دسته هم معنا برای اسامی زبان فارسی فراهم آید تا با درکل ۱۰۰۰۰ دسته هم معنا برای مقوله‌های اسم، فعل و صفت از جهت تعداد در ردیف وردنت‌های متوسطی مثل وردنت عربی قرار بگیریم.

برای ارزیابی کیفی ساختار و ویژگی‌های درنظر گرفته شده نیز می‌توان به این نکته توجه نمود که از آنجا که به منظور بالابردن قدرت تطابق با وردنت‌های دیگر جهان و برقراری ارتباط میان زبان فارسی با زبانهای دیگر خصوصا انگلیسی طراحی شبکه واژگانی فارسی بسیار به طراحی شبکه پرینستون نزدیک بوده است، لذا کیفیت ساختار و ویژگی‌های طراحی شده برای شبکه اسامی فارسی به جهت تطابق با وردنت پرینستون مورد تایید است.

همچنین از آنجا که برای تهیه محتوای این شبکه (تعیین واژه ها، معادل‌های انگلیسی، واژه‌های مترادف، تعریف همراه هر دسته‌ی هم معنا و مثالهای ذکر شده برای آنها و ...) از منابع معتبر زبان‌شناسی همچون پیکره‌های فارسی و فرهنگ‌های دوزبانه و تک زبانه بهره گرفته شده صحت محتوایی شبکه منوط به صحت محتوایی این منابع است که همگی مورد تایید عامه متخصصان این حوزه هستند.

در پایان لازم است به نکاتی چند که در این پروژه یافت شد و بعضا مختص شبکه‌ی واژگانی اسامی زبان فارسی هستند اشاره نماییم:

۱.در زبان فارسی تا آنجاییکه نگارندگان مطالعه نموده‌اند رابطه‌ی شمول معنایی بین اسامی فارسی به صورت دقیق معین نگردیده است و این پژوهش برای اولین بار این روابط را معین کرده است. برای مثال روابط بین واژه‌های " ابزار، آلت، دستگاه، وسیله وعامل " به صورت سلسله مراتب شمول معنایی در نظر گرفته شده است (ابزار زیر شمول آلت که خود زیرشمول دستگاه که خود زیر شمول وسیله که خود زیر شمول عامل می‌باشد) و هم شمولهای[2] آلت عبارتند از "افزار،ادوات" وهم شمولهای دستگاه عبارتند از" اسباب و سازوکار " و هم شمول" وسیله "،" تجهیزات" می‌باشد.

۲. یکی از موارد مهم در روابط شمول معنایی دیدگاههای معرفت‌شناختی اعمال شده در این روابط است مثلا دسته‌ی هم معنای {انسان، بشر} در دسته‌های هم معنای BC1 زیر شمول نخستی‌ها در نظر گرفته شده است که اعمال آن با دیدگاههای معرفت شناختی ما موافق نمی باشد و بنابر این، این بخش از روابط واژگانی به صورت ویژه‌ی زبان فارسی در نظر گرفته شده است.

۳. نکته‌ی دیگرعدم تطبیق واژگانی و خلاءهای واژگانی موجود میان دسته‌های هم معنای پرینستون و دسته‌های هم معنای شبکه‌ی واژگانی زبان فارسی است. در هنگامی که برای دسته‌های هم معنای انگلیسی (مفاهیم پایه‌ی گروه اول) معادلسازی انجام می‌پذیرفت برای بعضی از کلمات معادل مناسب واژگانی در فارسی یافت نشد که البته قابل پیش بینی بود. در بعضی از این موارد اقدام به ساخت واژه‌های معادل شده است همانطور که در ساخت شبکه‌های واژگانی دیگر زبانها این امر صورت پذیرفته بود مثلا برای دسته‌ی هم معنای {headdress,headgear} معادل {سر آذین، سر افزار، زینت سر}

[2] دو واژه که هر دو زیرشمول یک واژه باشند را هم شمول گویند

داده‌های مربوط به یکی از این روابط صورت معکوس آن نیز ایجاد می‌شود.

دارای عضو

بعضی از دسته‌های هم معنا خصوصا آنهایی که به صورت مجموعه می‌باشند دارای این رابطه می‌باشند مانند: دسته‌ی هم معنای {بدن، پیکر،تن، جثه، کالبد، تنه} که دارای رابطه‌ی دارای عضو با دسته‌های هم معنای {دست،ید}، {پا،لنگ، رجل}، {گوش،اذن}،{چشم، دیده، عین}،{بینی، دماغ، غنه، خیشوم، پوز}، {سر،کله، راس} و... است.

عضو واژه

عده‌ای از دسته‌های هم معنا خود عضوی از دسته‌های هم معنای دیگر می‌باشند مانند: دسته‌های هم معنای {دست،ید}، {پا،لنگ، رجل}، {گوش،اذن}،{چشم، دیده، عین}،{بینی، دماغ، غنه، خیشوم، پوز}، {سر،کله، راس} که عضوی از دسته‌ی هم معنای{بدن، پیکر،تن، جثه، کالبد، تنه} می‌باشند. رابطه‌های "عضوی از" و "دارای عضو" نیز معکوس می‌باشند.

دارای بخش

عده‌ای از دسته‌های هم معنای دارای بخش یا بخشهایی می‌باشند مانند دسته‌ی هم معنای {نامه،رقعه،نوشته} که دارای بخش‌های {خطوط} و{آدرس،نشانی} و{پاکت} می‌باشد.

بخش واژه

تعدادی از دسته‌های هم معنا بخشی از دسته‌های هم معنای دیگر می‌باشند, برای مثال {آدرس،نشانی} و{پاکت}بخشی از دسته‌ی هم معنای{نامه،رقعه،نوشته}می باشند. این رابطه صورت معکوس رابطه‌ی "دارای بخش" می‌باشد.

دارای واحد

این رابطه میان بعضی دسته‌های هم معنا و دسته‌های هم معنای دیگرکه واحد شمارش انها هستند برقرار است مانند: دسته‌ی هم معنای {آب، مایه‌ی حیات،ماء} و دسته‌ی هم معنای {قطره،چکه،ذره،قطرات،ذرات} ویا دسته‌ی هم معنای {چک} که دارای رابطه‌ی "دارای واحد" با {فقره} است.

واحد واژه

تعدادی از دسته‌های هم معنا واحدی از دسته‌های هم معنای دیگر می‌باشند, برای مثال دسته‌ی هم معنای {فقره} واحد دسته‌ی هم معنای {چک}است. این رابطه صورت معکوس رابطه‌ی "دارای واحد" است.

دارای جنس

این رابطه میان گروهی از دسته‌های هم معنا و دسته‌های هم معنای دیگر که نشان دهنده‌ی جنس سازنده‌ی آنهاست برقرار است مانند دسته‌ی هم معنای {شمشیر، تیغ، دشنه} که دارای رابطه‌ی "دارای جنس" با دسته‌ی هم معنای {آهن، فولاد، پولاد، استیل} می‌باشد.

جنس

این رابطه اشاره به دسته‌های هم معنایی که از طریق جنس خاصی به یکدیگر مربوط می‌شوند دارد. مثلا دسته‌ی هم معنای {آهن،فولاد،پولاد،استیل} جنس دسته‌های هم معنایی مانند {شمشیر،خنجر، تیغ، دشنه} و{قاشق} و {چنگال} و... را تشکیل می‌دهد.

نزدیک به مترادف

این رابطه در حقیقت بین دسته‌های هم معنایی برقرار است که به جهت هم معنایی بسیار به هم نزدیکند اما نمی توان آنها را در بافت مشخص بجای یکدیگر بکار برد و بنابراین در یک دسته‌ی هم معنا جا نگرفته و در دو دسته‌ی هم معنا قرار می‌گیرند. برای مثال در بافتی که سخن از موسیقی است، بجای آلات می‌توان از ابزار و وسایل استفاده نمود اما کاربرد تجهیزات یا امکانات موسیقی صحیح نمی باشد. بدین دلیل دسته‌ی هم معنای {آلات،آلات، افزار، ادوات} و دسته‌ی هم معنای {تجهیزات،امکانات} از یکدیگر مجزا می‌گردند. با این وجود دسته‌ی هم معنای {آلات،آلات، افزار، ادوات} به دسته‌ی هم معنای {تجهیزات،امکانات} نزدیکی زیادی دارد و در بافتهای دیگری ممکن است بجای یکدیگر به کار روند تا به دسته‌ی هم معنای {ماشین،اتومبیل،خودرو،وسیله‌ی نقلیه} که هر سه در طبقه‌ی مصنوعات آورده شده‌اند و بنابراین رابطه‌ی نزدیک به مترادف برای آن دو در نظر گرفته می‌شود.

نزدیک به متضاد

در روابط شبکه‌ی واژگانی پرینستون تضاد رابطه‌ی بین صورت واژه هاست و نه معنای آنها. اما اگر کلیه‌ی واژه‌های یک دسته‌ی هم معنا با همه‌ی واژه‌های دسته‌ی هم معنای دیگری متقابل باشند برای آنها رابطه‌ی نزدیک به متضاد در نظر گرفته می‌شود, برای مثال رابطه‌ی دسته‌های هم معنای {ناتوانی، معلولیت، عجز} و{ توانایی، قدرت، قدرت ذهنی}.

سبب شدن

هرگاه دسته‌ی هم معنایی منجر به ایجاد دسته‌ی هم معنای دیگر گردد بین دسته‌ی هم معنای اول و دوم رابطه‌ی باعث شدن وجود دارد. مانند رابطه‌ی دسته‌ی هم معنای {زلزله، زمین لرزه} و{ ویرانی، خرابی، تخریب، انهدام }.

منبعث شدن

هرگاه دسته‌ی هم معنایی ناشی از دسته‌ی هم معنای دیگر گردد این رابطه بین آنها برقرار است مانند رابطه‌ی میان دسته‌ی هم معنای {ویرانی، خرابی، تخریب، انهدام } و دسته‌های هم معنای {سیل، سیلاب} یا {زلزله، زمین لرزه}.

دارای زیررویداد

عبارت است از زمانی که یک دسته‌ی هم معنا رویداد دیگری را در درون خود دارد؛ برای مثال دسته‌ی هم معنایی مانند {مسابقه، رقابت} دارای زیر رویدادی است که همان دسته‌ی هم معنای {برد، پیروزی، موفقیت،غلبه} است.

زیررویداد

این رابطه معکوس رابطه‌ی " دارای زیر رویداد" می‌باشد و عبارتست از رابطه یمیان دسته‌های هم معنایی که در درون دسته‌ی هم معنای دیگری قرار دارند.

اشتقاق

برای این رابطه در مرحله‌ی کنونی، دو واژه‌ی مشتق از یکی از اعضای دسته‌ی هم معنا آورده می‌شود. در صورت وجود داشتن فعل مرکب که از اسم مشتق شده باشد، ابتدا آن فعل و سپس واژه‌ی دیگری که از آن مشتق شده آورده خواهد شد, برای مثال برای دسته‌ی هم معنای {دست, ید} افعال "دست دادن" و "دست گذاشتن" به عنوان افعالی

معنا اثبات گردد. در شبکه‌ی واژگانی اروپا منظور از معادل- معنایی (هم معنایی) این است که دو واژه بدون در نظر گرفتن تفاوتهای ساختواژی- نحوی، سیاقی، سبکی، گویشی و یا کاربردشناختی شان به هستینه‌های مشابهی دلالت نمایند. نکته‌ی کاربردی دیگر این است که هم معناها نمی توانند به وسیله‌ی هیچ رابطه‌ی معنایی تعریف شده‌ی دیگری به هم مربوط گردند.

آزمون مورد استفاده در شبکه‌ی واژگانی اسامی زبان فارسی برای مشخص نمودن اعضای دسته‌ی هم معنا در زیر آورده شده است.

آزمون ۱ رابطه‌ی ترادف بین اسامی

بله	(A)	آن (یک) الف است بنابراین آن (یک) ب است
بله	(B)	آن (یک) ب است بنابراین آن (یک) الف است
شرایط:		الف و ب اسامی مفرد یا جمع می‌باشند
مثال:	(A)	آن رستنی است بنابر این آن گیاه است
	(B)	آن گیاه است بنابراین آن رستنی است
نتیجه :		{ رستنی، گیاه} اعضای یک دسته‌ی هم معنا

سایر اعضای دسته‌ی هم معنای { گیاه، نبات، رستنی}را نیز می‌توان آزمود و صحت جمله‌های ساخته شده دلیل قرار گرفتن اسامی گیاه، نبات و رستنی در یک دسته‌ی هم معناست.

با توجه به اینکه دسته‌های هم معنا در شبکه‌ی واژگانی پرینستون بسیار جزئی و ظریف و با دقت انتخاب شده‌اند برای معادل یابی فارسی نیز باید همان دقت اعمال گردد و این مسئله کار ساخت دسته‌های هم معنا را بسیار وقت گیر نموده و می‌نماید چون این دقت درتدوین فرهنگهای انگلیسی به فارسی موجود اعمال نگردیده است و در بسیاری از موارد حتی بعضی از معادلها در لغت نامه‌ی انگلیسی- فارسی آورده نشده است (اگر معادل یابی به صورت خودکار انجام می‌شد مسلما یا معادل مناسبی برای بعضی واژه‌ها یافت نمی گردید و یا واژه‌ی غلطی به عنوان معادل در نظر گرفته می‌شد). همزمان با معادل سازی BC1 و BC2، جهت یافت مفاهیم پایه‌ی ویژه‌ی زبان فارسی با استفاده از پیکره‌ی الکترونیکی زبان فارسی[۲۷] که پایگاه داده‌ای برخط حاوی انواع متون نوشتاری و گفتاری زبان فارسی است و در زمان انجام پروژه حاوی بیش از ۵۰،۰۰۰،۰۰۰ واژه بود، ابتدا واژه‌های پربسامد (واژه هایی که در پیکره دارای بالاترین تعداد رخداد بوده اند) و سپس از بین آنها اسامی پربسامد فارسی به صورت غیرخودکار (به علت وجود نداشتن لیست مجزایی از اسامی زبان فارسی و نبود برچسب در پیکره‌ی یاد شده) استخراج گردید و با مقایسه‌ی اسامی با دسته‌های هم معنای گروه اول و دوم شبکه‌ی واژگانی پرینستون و مشخص نمودن اسامی که در آن دسته‌های هم معنا وجود نداشت، دسته‌های هم معنای اسامی ویژه‌ی زبان فارسی با استفاده از منابعی که برای ساخت معادلهای دسته‌های هم معنای پرینستون مورد استفاده قرار گرفت ساخته شد.

پس از یافتن هم معناها و ساخت دسته‌های هم معنا نوبت به ورود داده‌ها و تعیین روابط میان دسته‌های هم معنا با استفاده از ویرایشگر بود. در شبکه‌ی واژگانی اسامی زبان فارسی علاوه بر رابطه‌ی معنایی هم معنایی یا ترادف (بین اعضای یک دسته‌ی هم معنا)، ۲۱ رابطه‌ی دیگر نیز برای دسته‌های هم معنا در نظر گرفته شده است. لازم به ذکر است که برای هر رابطه آزمونی طراحی شده است که برای نمونه با آزمون ۱ که مربوط به ترادف بود آشنا شدید در زیر مروری اجمالی بر این روابط همراه با مثالهایی برای هر کدام خواهیم داشت:

ترادف

برای هر اسم در دسته‌ی هم معنای مربوطه مترادفهای آن درج شده است. در شبکه‌ی واژگانی اسامی زبان فارسی مانند سایر شبکه‌های واژگانی تفاوتهای سیاقی، سبکی، گویشی و یا کاربردشناختی عاملی برای متفاوت دانستن مترادفهای یک کلمه در نظر گرفته نشده است (هر چند تفاوتهای ساختواژی • نحوی برخلاف شبکه‌ی واژگانی اروپا به خاطر اختصاص شبکه‌ی حاضر به مقوله‌ی اسم در نظر گرفته شده است)؛برای مثال واژه‌های { الله، اله، خدا، خداوند، خداوندگار، ایزد، یزدان، پروردگار، رب،کردگار،دادار، آفریدگار، اهورامزدا، خالق} دسته‌ی هم معنایی را شکل می‌دهند که در آن واژه های‌ی یزدان و دادار و اهورامزدا که در سیاق ادبی بکار می‌رود, در کنار سایر واژه‌ها و به صورت یک دسته‌ی هم معنا درج شده است و یا واژه‌ی ادبی گردن فرازی در دسته‌ی هم معنای { تکبر، خود بزرگ بینی، نخوت، گردن فرازی}.

دربردارنده یا شامل

برای هر دسته‌ی هم معنا، دسته یا دسته‌های هم معنایی را که آن شامل می‌گردد مشخص می‌نماید برای مثال دسته‌ی هم معنای {فصل، فصول، موسم}شامل دسته‌های هم معنای زیر می‌باشد:

{بهار، بهاران، ربیع}

{تابستان، تموز، صیف}

{پاییز،پائیز، برگریزان، خزان، خریف، مهرگان}

زیرشمول

هر دسته‌ی هم معنا خود زیر شمول دسته‌ی هم معنای دیگری است و البته این زیر شمول بودن تا بالاترین سطح که "هستینه " می‌باشد ادامه می‌یابد. جهت ساخت هسته‌ی شبکه‌ی واژگانی زبان فارسی فعلا تنها یک سطح بالاتر ذکر می‌گردد. برای مثال دسته‌ی هم معنای {فصل، فصول، موسم} زیر شمول {دوره، زمان، روزگار} می‌باشد ویا {دستگاه، اسباب} زیرشمول {وسیله} و {وسیله} زیرشمول{واسطه، عامل} و{واسطه، عامل} زیر شمول{کل، کلیت، تمامیت}و{کل، کلیت، تمامیت}زیرشمول{شی، شی فیزیکی}و{شی، شی فیزیکی}زیرشمول{موجود فیزیکی، هستینه‌ی فیزیکی}و{هستینه} است. این رابطه‌ی سلسله مراتبی با ورود داده‌های مربوط به همه‌ی دسته‌های هم معنا به صورت خودکار بین تمامی دسته‌های هم معنا شکل می‌گیرد.

کل واژه

برای هر دسته‌ی هم معنا با توجه به اینکه واجد اجزاء باشد، رابطه‌ی دارای جزء در نظر گرفته شده است. برای مثال دسته‌ی هم معنای {کتاب، مجلد} دارای جزءهای {شیرازه}، {اوراق، ورق} و {جلد} است.

جزء واژه

بعضی از دسته‌های هم معنا خود جزئی از یک دسته‌ی هم معنای دیگرند مانند دسته‌های هم معنای {شیرازه}، {اوراق، ورق} و {جلد} که جزءهای دسته‌ی هم معنای {کتاب،مجلد} می‌باشند. رابطه‌های "جزئی از" و "دارای جزء" معکوس می‌باشند و به همین دلیل در شبکه‌ی واژگانی اسامی زبان فارسی دارای برچسبی می‌باشند که با ورود

واژگانی افعال و صفات زمینه‌ی تلفیق این سه شبکه‌ی واژگانی امکان پذیر می‌باشد که با این کار شبکه‌ی واژگانی زبان فارسی عینیت می‌یابد و پس از ارزیابی‌های بعدی از آن می‌توان در پروژه‌های پردازش زبان طبیعی بهره گرفت. در ادامه ابتدا با ویرایشگر و نرم افزار و بدنبال آن با روش ساخت و در انتها با منابع زبانی و فرایند ایجاد شبکه‌ی واژگانی اسامی زبان فارسی آشنا خواهید شد.

۸-۱- ویرایشگر و نرم افزار

ویرایشگر: ویرایشگر بکار برده شده برای ساخت شبکه‌ی واژگانی اسامی زبان فارسی، نسخه‌ی تنظیم شده‌ی ویرایشگر VisDicاست که ضمن اینکه در آن برای نخستین بار جهت گیری چپ به راست تعریف شده است، برای این شبکه تنظیم گردیده است. همچنین به واسطه‌ی ساختار پرونده‌های این ویرایشگر که در قالب XML می‌باشند، قابلیت اتصال به دیگر شبکه‌های موجود را نیز دارد.

پیش نیازها: برای اجرای نرم افزار شبکه‌ی واژگانی اسامی زبان فارسی نیاز به سیستم عامل ویندوز ۲۰۰۰ یا بالاتر است و مدیر پایگاه داده‌ی مورد نیاز برای اجرا، زبان پرس و جوی ساخت یافته‌ی مایکروسافت نسخه‌ی ANSI 92 یا بالاتر است.

سکوهای قابل اجرا : سکوی اجرایی نرم افزار حاضر، سیستم عامل ویندوز XP یا بالاتر است و به صورت نسخه‌ی رومیزی قابل اجراست.

روش شناسی و فناوری: در طراحی، تحلیل و پیاده سازی شبکه‌ی واژگانی اسامی زبان فارسی از روش نمونه سازی استفاده شده است. زبان برنامه نویسی مورد استفاده نیز ویژوال بیسیک می‌باشد. معماری پیاده سازی به روش سه لایه صورت گرفته است که شامل الف) لایه‌ی دسترسی به داده ها، ب) لایه‌ی منطق و عملکرد و پ) لایه‌ی نمایش می‌باشد. بین لایه‌ی اول و سوم از یک همسان ساز برای حفظ سازگاری با سایر سیستم‌های مدیریت داده‌های آینده استفاده شده است. این سه هر کدام وظایف جداگانه و خاص خود را دارند. در سیستم‌های پیاده سازی شده با معماری سه لایه کارشناسان می‌توانند در صورت نیاز برای ایجاد تغییر که می‌تواند کوچک یا بزرگ باشد در هر لایه این تغییرات را اعمال نمایند بدون اینکه نیاز به تغییر در لایه‌های دیگر باشد یا کمترین تغییرات متوجه لایه‌ی دیگر شود. برای تولید نسخه‌ی وب می‌توان از لایه‌های اول و دوم نسخه‌ی رومیزی نرم افزار موجود استفاده کرد و تنها لایه‌ی سوم را طراحی کرد.

معماری پایگاه داده ها: پایگاه داده شبکه‌ی واژگانی اسامی زبان فارسی شامل چهار جدول Nouns , NSynsets , Synset-Relations, Synset-synonyms است که حاوی اسامی، دسته‌های هم معنا و اطلاعات مربوط به آنها می‌باشند. در جدول Nouns تا کنون۸۱۲۳ اسم فارسی وارد گردیده است. در جدول NSynsets دسته‌های هم معنا، جزء کلامی آنها، تعریف، کاربرد و نوع مفهوم پایه (گروه ۱ یا ۲) نگهداری می‌شود. در جدول Synset-Relations روابط تعریف شده برای دسته‌های هم معنا که حدود ۸۵۵۸ رابطه را دربرمی گیرد و درجدولSynset-synonyms واژگان هم معنا ذخیره گردیده اند.

۸-۲- روش ساخت

در ساخت شبکه‌های واژگانی سه روش متداول است. در روش اول که به رهیافت تلفیقی موسوم است شبکه‌ی واژگانی به صورت مستقل از سایر شبکه‌های واژگانی و با استفاده از منابع زبانی زبان مورد نظر ساخته می‌شود. در این روش عدم وجود سوگیری نسبت به زبانی خارجی خاص جزء مزیتهای شبکه‌ی واژگانی ایجاد شده می‌باشد؛ اما روش یاد شده بسیار کند، وقت گیر و پیچیده است.

روش دوم که به رهیافت توسیعی مشهور است با استفاده ازترجمه‌ی دسته‌های هم معنای شبکه ی واژگانی پرینستون انجام می‌پذیرد و ساختار موجود در آن شبکه در زبان مقصد پیاده سازی می‌گردد. این روش نسبت به روش اول نیازمند نیرو و وقت کمتری بوده (هر چند نیرو و وقت بسیار زیادی می‌طلبد) لیکن شبکه‌ی واژگانی ایجاد شده سوگیری بسیاری به شبکه‌ی واژگانی انگلیسی دارد.

روش سوم رهیافت ساخت بالا به پایین است که در آن سعی می‌شود نقاط قوت روشهای فوق اعمال و نقاط ضعف آنها برطرف گردد. در این رهیافت هسته‌ی شبکه‌ی واژگانی با استفاده از مفاهیم پایه‌ی مشترک زبانها، در کنار مفاهیم ویژه‌ی زبان شکل می‌گیرد و سپس بر این اساس شبکه‌ی واژگانی گسترش یافته تا عینیت یابد. این روش در حالیکه حافظ ساختار ویژه‌ی زبان است، امکان تطابق هر چه بیشتر شبکه‌ی واژگانی ایجاد شده با شبکه‌های واژگانی زبانهای دیگر را پدید می‌آورد.

جهت ساخت شبکه‌ی واژگانی اسامی زبان فارسی از روش سوم یعنی رهیافت ساخت بالا به پایین بهره گرفته شده است.

۸-۳- فرایند ساخت و منابع زبانی

در این پروژه بخاطر دردسترس نبودن فرهنگ واژه‌های به روز قابل خواندن برای رایانه، معادل یابی بخش اول و غالب بخش دوم مفاهیم پایه مشترک زبانها (بخش اول دسته‌های هم معنای بکاربرده شده در شبکه‌ی واژگانی اروپا موسوم به BC1 وبخش دوم موسوم به BC2 است)، با استفاده از فرهنگ‌های غیر الکترونیکی معتبر انجام پذیرفت.

جهت اطمینان از صحت انتخاب معادلهای اعضای دسته‌های هم معنای گروه اول و دوم، ابتدا دسته‌ی هم معنا و تعریف و مثالهای آن در کنار حوزه و هستان‌شناسی مقوله بالایی آن به دقت مطالعه شده و با مراجعه به لغت نامه‌ی باطنی [۲۹] دقیقترین معادل فارسی برای آن با توجه به حوزه و تعریف و مثالهای آن درBC1 و BC2 انتخاب شده است. نقطه‌ی قوت فرهنگ یاد شده دسته بندی معانی واژه‌ها بر اساس حوزه‌ی معنایی است (هر چند حوزه‌ها مشخص نشده‌اند)سپس با استفاده از فرهنگ طیفی فراروی [۳۰] و فرهنگ جامع مترادف و متضاد خداپرستی [۳۱] و همچنین گاهی اوقات فرهنگهای دیگر مانند فرهنگ رایانه‌ای سلیمان پور [۳۲]، فرهنگ عمید [۳۳] و فرهنگ حییم [۳۴] و شم زبانی زبانشناس سایر اعضای دسته‌ی هم معنا معین گردیده و اعضای دسته‌ی هم معنا قطعی شده است. اعضای دسته‌ی هم معنا با استفاده از آزمون مورد بررسی قرار گرفته تا مانند آنچه در شبکه‌ی واژگانی اروپا انجام گرفته بود، قابلیت قرار گرفتنشان در یک دسته‌ی هم

ابزاری برای تولید سلسله مراتب مفهومی برای بازیابی دو سویه‌ی اطلاعات و ساخت چند بعدی‌های منتظم برای داده‌های چند رسانه‌ای تاکید دارند. علاوه بر این به شاخص گذاری نسخه‌های متنی برنامه‌های رادیویی با هدف بازیابی اطلاعات از رسانه‌های رادیویی با استفاده از شبکه‌ی واژگانی اشاره می‌نمایند.

۶-۷- فرهنگ نگاری

روحی زاده [۳] به نقل از مورفی [۲۵] از شبکه‌ی واژگانی به عنوان منبع ارزشمندی برای انواع طرحهای فرهنگ نگاری یاد می‌نماید و به گنج واژه‌ای تصویری برای زبان انگلیسی که بر پایه‌ی شبکه‌ی واژگانی پرینستون طراحی شده اشاره می‌نماید. در این فرهنگ معنایی با درج واژه توسط کاربر، تصویری ارائه می‌گردد که واژه‌ی مورد نظر در وسط و مجموعه‌ی واژه‌های مرتبط با آن از جمله هم معنا ها، متقابلها، زیر شمولها و غیره با رنگهای متفاوت در اطرافش نمایش داده می‌شوند. فاصله‌ی کلمات مرتبط با واژه‌ی میانی نیز تداعی کننده‌ی فاصله‌ی معنایی واژه‌ها با یکدیگر است. این فرهنگ معنایی را در وبگاه http://www.visualthesaurus.com می‌توان مشاهده کرد.

۶-۸- جستجوهای اینترنتی

جستجوی واژه‌ها در اینترنت بر اساس ملاک صورت واژه استوار است و این یکی از نقاط ضعف جستجو خصوصا در مواردی که با ابهام واژگانی روبرو می‌باشیم است. روحی زاده [۳] به طرحی اشاره می‌نماید که از سوی نمراوا [۲۶] ارائه شده و در آن از تعاریف دسته‌های هم معنای شبکه‌ی واژگانی برای پالایش معنایی نتایج جستجوی عبارتهای هم معنا در گوگل استفاده می‌شود. در این طرح نتایج جستجو بر اساس بافت تقسیم بندی و به کاربر ارائه می‌شود و او بافت مورد نظر را انتخاب می‌نماید.

پس از آگاهی از کاربردهای متنوع شبکه‌ی واژگانی به شبکه‌ی واژگانی زبان فارسی و شبکه‌ی واژگانی اسامی زبان فارسی به عنوان پروژه‌های در جریان می‌پردازیم.

۷- شبکه‌ی واژگانی زبان فارسی

جهت ساخت شبکه‌های واژگانی دو روش دستی و نیمه خودکار پیشنهاد شده است و بنابراین برای ساخت شبکه‌ی واژگانی زبان فارسی نیز به دو روش دستی و نیمه خودکار می‌توان عمل نمود. روشهای بکار گرفته شده توسط فامیان [۲] جهت ساخت شبکه واژگانی صفات زبان فارسی، روحی زاده [۳] جهت شبکه‌ی واژگانی افعال زبان فارسی، طرح کیوان و همکاران [۱] که در سومین کنفرانس جهانی شبکه‌ی واژگانی گزارشی از آن ارائه گردید و طرح پیشنهادی منصوری و بیجن خان [۴] برای افعال زبان فارسی و همچنین روش بکار گرفته شده در پروژه‌ی جاری برای اسامی جزء روشهای دستی ساخت شبکه‌ی واژگانی می‌باشند. لازم به ذکر است کیوان و همکاران [۱] در طرح خود تنها به سه حوزه‌ی ورزش، حمل ونقل و جغرافیا پرداخته بودند و در گزارش ارائه شده به عدم پیشرفت کار به علت نبود مشارکت داوطلبانه اشاره شده است و تا بحال اطلاعاتی در مورد کاربرد عملی آن در پردازش زبان طبیعی منتشر نگردیده است.

شمس فرد [۵] در چهارمین کنفرانس جهانی شبکه واژگانی و همچنین در کنفرانسLREC [۶] روشی نیمه خودکار جهت ساخت شبکه‌ی واژگانی روشی نیمه خودکار را جهت ساخت شبکه‌ی واژگانی معرفی نموده است.

با توجه به اینکه روش به کار گرفته شده در این پروژه دستی است به بررسی اجمالی روشهای دستی به کار رفته‌ی پیشین می‌پردازیم و سپس به تحلیل و بررسی روش به کار گرفته شده در پروژه‌ی اخیر خواهیم پرداخت.

در زمینه‌ی شبکه‌ی واژگانی صفات زبان فارسی رساله‌ی فامیان [۲] به انجام رسیده که طی آن طراحی شبکه‌ی واژگانی صفات زبان فارسی انجام پذیرفته است. دراین طرح صفات فارسی به پانزده طبقه‌ی اصلی و و بیش از هفتاد طبقه‌ی فرعی تقسیم شده‌اند و برای هر صفت ده رابطه یا اطلاع معنایی پیش بینی شده، که وی سه رابطه‌ی "مشخصه‌ی تعریفی برجسته"، "مشخصه‌ی تعریفی بالقوه" و "نیاز دارد به" را به عنوان روابط جدید در شبکه‌ی واژگانی صفات فارسی گنجانده است. در این پژوهش دسته‌های هم معنای صفات به صورت دستی ساخته شده‌اند اما نگارنده در مورد فرایند معادل سازی و روش ساخت دسته‌های هم معنا توضیح دقیقی نداده است و همچنین اینکه آیا از کدهای یکسانی جهت معادل سازی با شبکه‌ی واژگانی پرینستون استفاده نموده است یا خیر.

در زمینه‌ی افعال زبان فارسی پایان نامه‌ی روحی زاده [۳] به انجام رسیده که طی آن با استفاده از روش «ساخت بالا به پایین» دسته‌های هم معنا شکل گرفته است. مفاهیم پایه‌ی شبکه‌ی واژگانی بالکان به صورت نیمه خودکار و با نظارت زبانشناس معادل سازی شده اند. سپس با استفاده از پیکره زبان فارسی [۲۷] افعال فارسی دارای بسامد بالا که در این مجموعه نیامده اند، اضافه گردیده است. در نهایت با افزودن واژه‌های شامل و یک سطح از واژگان زیر شمول هسته‌ی شبکه‌ی واژگانی افعال زبان فارسی مشتمل بر۱۵۰۰ فعل شکل گرفته است. گسترش این هسته در مراحل بعدی مد نظر است. لازم به ذکر است که افعال مرکب زبان فارسی در این طرح با استفاده از دیدگاه دبیرمقدم [۲۸] مورد تحلیل قرار گرفته اند. در طرح فوق نیز فرایند معادل سازی و روش ساخت دسته‌های هم معنا و استفاده از کدهای یکسان با کدهای شبکه‌ی واژگانی پرینستون مشخص نگردیده است.

درحال حاضر پروژه‌ای در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی در حال انجام است که هدف آن ساخت شبکه‌ی واژگانی فارسی به عنوان فاز اول از پروژه فارس‌نت است. در این پروژه علاوه بر مجتمع سازی فعالیت‌های گسسته گذشته، ساخت شبکه واژگانی فارسی به روشی نیمه خودکار تا تکمیل حدود ۱۰۰۰۰ دسته‌ی هم معنا برای مقوله‌های اسم، فعل و صفت ادامه می‌یابد. شبکه واژگانی اسامی مورد نظر در این مقاله نیز بخشی از این پروژه خواهد بود

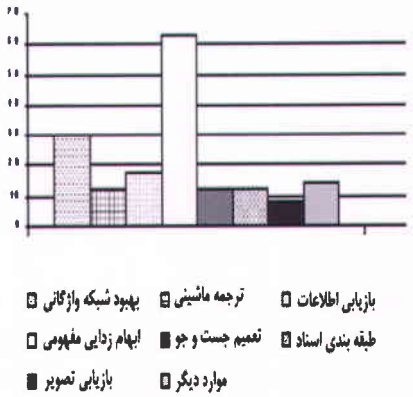
۰- شبکه‌ی واژگانی اسامی زبان فارسی

شبکه‌ی واژگانی اسامی فارسی با توجه به آخرین معیارهای طراحی و ایجاد که برای شبکه‌های واژگانی بالکان مورد استفاده قرار گرفته است در حال ایجاد است تا اینکه حداکثر تطابق را با شبکه‌های یاد شده جهت بررسی‌ها و کاربردهای میان زبانی داشته باشد. همچنین بواسطه‌ی اشتراک میان ابزارهای طراحی این شبکه و ابزارهای ایجاد شبکه‌ی

را نیز در قالب پرونده‌ی XML عرضه می‌نماید. ساختار پرونده‌های یاد شده با همه‌ی شبکه‌های موجود تطابق دارد و بنابراین اتصال هر شبکه‌ی واژگانی که بر این مبنا ساخته شده باشد را با شبکه‌های واژگانی موجود فراهم می‌آورد.

۶- کاربردهای شبکه‌ی واژگانی

دسترسی آسان و کیفیت و توانایی بالقوه‌ی شبکه‌ی واژگانی در زمینه‌ی پردازش زبان طبیعی عامل گسترش و موفقیت روز افزون آن بوده است. شکل ۲ توزیع موضوعی مقالات و طرحهای تحقیقاتی را طی سالهای ۱۹۹۴ تا ۲۰۰۳ نشان می‌دهد [۱۲]. در این نمودار بازیابی تصویر با کمتر از ۸ مورد و ابهام زدایی مفهومی با بیش از ۶۸ مورد رکورد داران استفاده از شبکه‌ی واژگانی بوده اند.



شکل ۲: توزیع موضوعی مقالات و طرحهای مربوط به شبکه‌ی واژگانی [۱۲]

آنچه از شکل۲ می‌توان دریافت کاربردهای متنوع شبکه‌ی واژگانی است که در زیر دقیقتر بدانها پرداخته شده است.

۶-۱- ابهام زدایی واژگانی یا مفهومی

در شکل۲، شبکه‌ی واژگانی شترین موفقیت را در زمینه‌ی ابهام زدایی واژگانی یا مفهومی کسب نموده است وآن عبارتست از « انتخاب معنای مناسب برای کلمه در بافتی مشخص» [۱۳].

یورافسکی و مارتین [۱۴] ابهام زدایی واژگانی را عبارت از بررسی نمودهای واژه در بافت و مشخص نمودن دقیق اینکه کدام معنی واژه مورد استفاده قرار گرفته است، می‌دانند.

موراتو و همکاران [۱۲] ابهام زدایی را پر وفورترین ومتنوعترین کاربرد شبکه‌ی واژگانی دانسته و به طرحها و پروژه‌های متنوعی که با استفاده از شبکه‌ی واژگانی به انجام رسیده اشاره نموده اند. از آن جمله طرح موسوم به‌ای دبلیو ای/ اچ که هدفش طراحی چارچوبی هستان شناختی برای ابهام زدایی معیارهای جستجوی اینترنتی است و طرحهای اوینگو و سیمپل فایند به عنوان دو محصول اینترنتی است که از ایجاد ابهام در جستجوی منابع زبان طبیعی دراینترنت جلوگیری می‌نماید. آنها (همانجا) به چندین طرح دیگر در این زمینه اشاره نموده اند. همچنین توفیس و همکاران [۱۵] از شبکه‌ی واژگانی رومانیایی به عنوان ابزاری مهم در ابهام زدایی واژه در این زبان استفاده نموده اند. علاوه بر این کرنر [۱۶] با استفاده از شبکه‌ی واژگانی زبان استونیایی اقدام به پیشنهاد طرحهایی جهت ابهام زدایی واژگانی در این زبان نموده است.

۶-۲- بازیابی و استخراج اطلاعات

یورافسکی و مارتین [۱۴] بازیابی اطلاعات را حوزه‌ای بسیار گسترده می‌دانند که طیف وسیعی از موضوعات وابسته به ذخیره سازی،تحلیل و بازیابی همه‌ی روشهای رسانه‌ای را شامل می‌شود.

این زمینه به بازنمایی و سامان دهی دانش موجود بر روی اینترنت مربوط می‌شود. به فاصله‌ی کوتاهی پس از ارائه‌ی شبکه‌ی واژگانی روشهای تلفیق منطقی و استنتاجی بر روی آن متمرکز شد. شبکه‌ی واژگانی به عنوان یک واژگان جامع معنایی در بازیابی اطلاعات و پس از آن به عنوان یک ابزار دانش زبانی برای بازنمایی و تفسیر معنای اطلاعات به کار گرفته شد. [۳]

توفیس و همکاران [۱۵] به استفاده از شبکه‌ی واژگانی رومانیایی در بازیابی و استخراج اطلاعات پرداخته اند. همچنین یارمحمدی و همکاران [۱۷] به سیستم پرسش و پاسخی که با استفاده از شبکه‌ی واژگانی اقدام به این عمل می‌نماید و کلارک و همکاران [۱۸] نیز به طرحی دیگری جهت پرسش و پاسخ با استفاده از شبکه‌ی واژگانی پرداخته اند.

۶-۳- ترجمه ماشینی

در فرایند ترجمه‌ی ماشینی با ابهام زدایی واژگانی و بازیابی و استخراج اطلاعات سر وکار داریم بنابراین به طور غیر مستقیم از شبکه‌ی واژگانی بهره گیری می‌نماییم. روحی زاده [۳] به طرحهای زیر که در آنها از شبکه‌ی واژگانی به طور مستقیم در سامانه‌های ترجمه استفاده گردیده اشاره نموده است، دور و ماریا [۱۹]، کانگ و همکاران [۲۰]، کانگ و همکاران [۲۱]، نایت [۲۲] و دور [۲۳]. همچنین توفیس و همکاران [۱۵] به استفاده از شبکه‌ی واژگانی رومانیایی برای ترجمه ماشینی در این زبان پرداخته اند.

۶-۴- طبقه بندی اسناد

در این زمینه تمرکز بر روی جنبه هایی از ابزار شبکه‌ی واژگانی است که برای مقوله بندی اسناد بکار می‌رود یعنی طبقه بندی معنایی اسناد با توجه به دسته بندی اسامی، افعال و صفات به کار رفته در آنها [۱۲]. به عبارت دیگر با توجه به کاربرد و بسامد اسامی، افعال و صفات خاصی در سند می‌توان آن را متعلق به طبقه‌ی خاصی از متون در نظر گرفت (البته ابتدا لازم است در این زمینه معیار سازی صورت پذیرد و مشخصات متنهای مختلف از لحاظ میزان استفاده از اسامی، افعال و صفات مشخص گردد).

۶-۵- آموزش زبان

موراتو و همکاران [۱۲] به نقل از هو و گراسر [۲۴] به طرحی اشاره می‌نمایند که برای ارزیابی تسلط دانش آموزان بر زبان با استفاده از شبکه‌ی واژگانی ارائه داده اند. همچنین برای آموزش واژه‌ها و روابط بین آنها در زبان آموزی می‌توان از شبکه‌ی واژگانی استفاده نمود.

۶-۶- بازیابی صدا و تصویر

موراتو و همکاران [۱۲] با اشاره به کاوشگر چند رسانه‌ای آنرا به عنوان نمونه‌ای شناخته شده برای استخراج اطلاعات و دانش چند رسانه‌ای از وب معرفی می‌نماید و بر نقش شبکه‌ی واژگانی در این طرح به عنوان

•- شبکه‌ی واژگانی اروپا

شبکه‌ی واژگانی اروپا متشکل از چندین شبکه‌ی واژگانی از زبانهای مختلف اروپایی است که هر کدام خود ساختاری شبیه به شبکه‌ی واژگانی پرینستون دارند. در درون هر شبکه‌ی واژگانی روابطی مانند روابط ذکر شده برای شبکه‌ی واژگانی پرینستون وجود دارد که با نام روابط واژگانی درون زبانی شناخته می‌شوند. علاوه بر روابط درون زبانی هر دسته‌ی هم معنا به نزدیکترین دسته‌ی هم معنا در شبکه‌ی واژگانی پرینستون ۱،۵متصل گردید. با ذخیره سازی شبکه‌های واژگانی در یک سیستم مرکزی پایگاه داده‌ی واژگانی، اقدام به ایجاد پایگاه داده‌ای چند زبانه گردیده است که دسته‌های هم معنای شبکه‌ی واژگانی پرینستون در آن نقش رابط میان زبانی را بازی می‌نمایند. بدین طریق از دسته‌ای هم معنا در یک زبان می‌توان به دسته‌ی هم معنای هم ارز در زبانی دیگر رفت. چنین پایگاه داده‌ای برای بازیابی اطلاعات میان زبانهای مختلف و یا مقایسه‌ی میان شبکه‌های واژگانی مفید است. با مقایسه می‌توان میزان ثبات روابط در شبکه‌های واژگانی را مشخص نموده و یا به ویژگیهای زبانی خاص پی برد. همچنین از آن به عنوان ابزاری قدرتمند برای مطالعه‌ی منابع معنایی واژگانی و ویژگیهای خاص زبانی استفاده نمود.

شبکه‌های واژگانی این پروژه در وهله‌ی اول تنها به مقولات اسامی و افعال پرداخته‌اند و مقولات صفت وقید را تا آنجائیکه به دو مقوله‌ی پیشین مرتبط بودند را تحت پوشش قرار داده اند. کلمات در نظر گرفته شده، متشکل از همه‌ی کلمات عام و اساسی زبانهاست که برای ارتباط معانی خاصتر مورد نیازند و همچنین کلماتی که بالاترین بسامد را در پیکره‌های عمومی داشته اند.

۴-۱- مفاهیم پایه

این مفاهیم که برای اولین بار در ساخت شبکه‌ی واژگانی اروپا مطرح گردید با هدف دستیابی به حداکثر هم پوشی و سازگاری میان شبکه‌های واژگانی موجود و همچنین حفظ ساختارهای ویژه و نظام واژگانی زبانها به کار گرفته شدند و مهمترین مشخصه آنها به نقل از وسن [۱۰] اهمیت آنها در شبکه‌ی واژگانی بود و این اهمیت ناشی از استفاده‌ی گسترده از آنها است، به صورت مستقیم یا برای ارجاع به مفاهیمی دیگری که به صورت گسترده مورد استفاده قرار می‌گرفتند. بنابر این براساس معیارهای تعداد روابط (به طور کلی یا منحصر به شمول معنایی) و جایگاه بالای آن در سلسله مراتب معنایی (در شبکه‌ی واژگانی ۱،۵ پرینستون یا هر طبقه بندی ویژه‌ی دیگری) می‌توان به انتخاب آنها دست زد. انتخاب مفاهیم پایه برای هر زبان در شبکه‌ی واژگانی اروپایی ابتدا به صورت مستقل انجام پذیرفت تا از سوگیری به زبان یا منبعی خاص جلوگیری گردد و سپس انتخابها به نزدیکترین معادل در شبکه‌ی واژگانی ۱،۵ ترجمه گردید و با مقایسه‌ی دسته‌های هم معنای زبانهای موجود در شبکه‌ی واژگانی اروپا نهایتا با محاسباتی که به صورت مفصل در وسن [۱۰] آورده شده است، ۱۳۱۰ دسته‌ی هم معنا شامل ۱۰۱۰ دسته‌ی هم معنای اسم و ۳۰۰ دسته‌ی هم معنای فعل به عنوان مفاهیم پایه مستقل از زبان (مشترک) انتخاب شدند. سپس هر زبان دسته‌های هم معنایی که در زبان خود مهم می‌دانست ولی آنها را در بین مفاهیم پایه‌ی مستقل از زبان نمی یافت، تحت عنوان مفاهیم پایه‌ی ویژه‌ی زبان به آنها افزوده و هسته‌ی شبکه‌ی واژگانی زبان خود را با استفاده از این دو مجموعه (مفاهیم پایه‌ی مستقل از زبان+ مفاهیم پایه‌ی ویژه‌ی زبان) و با اضافه نمودن دسته‌های هم معنای شامل و زیر شمول به مفاهیم پایه ایجاد نمود.

۵- شبکه‌ی واژگانی بالکان

شبکه‌ی واژگانی بالکان شبیه به شبکه‌ی واژگانی اروپا یک پایگاه داده‌ی چند زبانه است که شامل شبکه‌های واژگانی زبانهای بلغاری، یونانی، رومانیایی، صربی، ترکی و صورت گسترش یافته‌ی زبان چکی که در شبکه‌ی اروپایی آغاز گردیده، است. در طراحی وایجاد این شبکه‌ی واژگانی از اصول و روشهای بکار رفته در شبکه‌ی اروپایی استفاده گردیده است [۱۱].

مفاهیم بنیادی وابزارهای ساخت شبکه‌ی واژگانی بالکان به عنوان ابزارها و مواد استاندارد ساخت شبکه‌های واژگانی جدید بکار می‌رود.

۵-۱- ویرایشگر VisDic

پس از اینکه در شبکه‌ی واژگانی اروپا از ویرایشگر پولاریس استفاده شد و اشکالات و معایب فنی در کنار هزینه‌ی بالای مجوز استفاده از آن مشخص گردید طراحان شبکه‌ی واژگانی بالکان به فکر ساخت ویرایشگر جدیدی برای پروژه‌ی خود افتادند. این ویرایشگر به عنوان ابزار مشاهده و ویرایش پایگاههای داده‌ای لغت نامه‌ای (در درجه‌ی اول شبکه‌های واژگانی)ذخیره شده در قالب XML ساخته شد. از ویژگیهای این ویرایشگر قابلیت انعطاف بالا جهت کار با انواع مختلف لغت نامه‌های تک زبانه، دو زبانه و گنج واژه‌ها است. در پنجره‌ی اصلی ویرایش می‌توان چند شبکه‌ی واژگانی را به صورت همزمان مشاهده و ویرایش نمود. همچنین می‌توان نتیجه‌ی جستجو را با اشکال مختلف مشاهده کرد. برای دیدهای متفاوت از جمله دید متنی، ویرایش، نمایش درختی و نمایش درختی معکوس، نتیجه‌ی جستجو،فهرست خارجی پرونده‌ها و نمایش XML می‌توان برنامه را تنظیم نمود. محتوای نمایش متنی کاملا از تعریفهای کاربرساخته می‌شود و بدین جهت دید ارائه شده به سادگی قابل خواندن است و بخش‌های مهم محتوای مدخل در آن مشخص است. برای مثال به نمایش متنی دسته‌ی هم معنای {sunset,sundown} در زیر توجه نمایید.

POS: n ID: ENG171-12836307-n
Synonyms: sunset:1, sundown:1
Definition: the time in the evening at which the sun begins to fall below the horizon
- ->> [hypernym] * [n] hour:2, time of day:1
- ->> [holo_part] * [n] evening:1, eve:4, eventide:1
- ->> [near_antonym] [n] dawn:1, dawning:1, morning:3, aurora:1, _rst light:1, daybreak:1, break of day:1, break of the day:1, dayspring:1, sunrise:1, sunup:1, cockcrow:1

کاربر می‌تواند با استفاده از امکانات ویرایشگر بر محتوای یک مدخل وارتباطات آن در ساختار سلسله مراتبی شبکه‌ی واژگانی نظارت کامل نماید. بسیاری از کارکردهای ویرایشگر توسط پرونده‌ی پیکربندی آن قابل انطباق با نیازهای موجود است. همه‌ی تنظیمات برنامه در چند پرونده‌ی XMLذخیره می‌شوند. همچنین این ویرایشگر خروجی خود

مختلف است که منجر به عدم استقلال واژه نگارهای فوق می‌شود. اما چالش پیش رو چگونگی انتخاب آغازکننده هاست. افراد مختلف می‌توانند انتخابهای مختلفی داشته باشند اما نکته‌ی اساسی تحت پوشش قرار دادن کلیه‌ی اسامی زبان هدف است. در شبکه‌ی واژگانی پرینستون ۲۵ آغازکننده در نظر گرفته شده‌اند که فهرست آنها در جدول ۱ آمده است.

جدول۱: فهرست ۲۵ آغازکننده‌ی شبکه‌ی واژگانی پرینستون [۸]

۱.کنش	۱۰.دارایی	۱۸.شکل
۲.شی طبیعی	۱۱.شناخت،دانش	۱۹.غذا
۳.جانور	۱۲.فرایند	۲۰.حالت
۴.پدیده‌ی طبیعی	۱۳.خبر،ارتباط	۲۱.گروه،مجموعه
۵.مصنوع	۱۴.کمیت،مقدار	۲۲.جوهر
۶.انسان،شخص	۱۵.رویداد،رخداد	۲۳.مکان
۷.مشخصه	۱۶.رابطه	۲۴.زمان
۸.گیاه	۱۷.احساس	۲۵.انگیزه
۹.بدن		

۳-۴- روابط معنایی

ساختار شبکه‌ی واژگانی بر اساس روابط معنایی میان واژه‌ها در درون دسته‌ی هم معنا (ترادف یا هم معنایی) و روابط معنایی میان دسته‌های هم معنا (شمول معنایی، جزء واژگی، تقابل معنایی و شمول همزمان) شکل گرفته است. میان دسته‌های هم معنای اسامی، شمول معنایی و جزء واژگی مهمترین روابط را تشکیل می‌دهند. نمونه‌ای از روابط معنایی در دسته‌های هم معنای که از نسخه‌ی ۱.۵ شبکه‌ی واژگانی پرینستون گرفته شده است، در شکل ۱ نشان داده شده است [۱۰]:

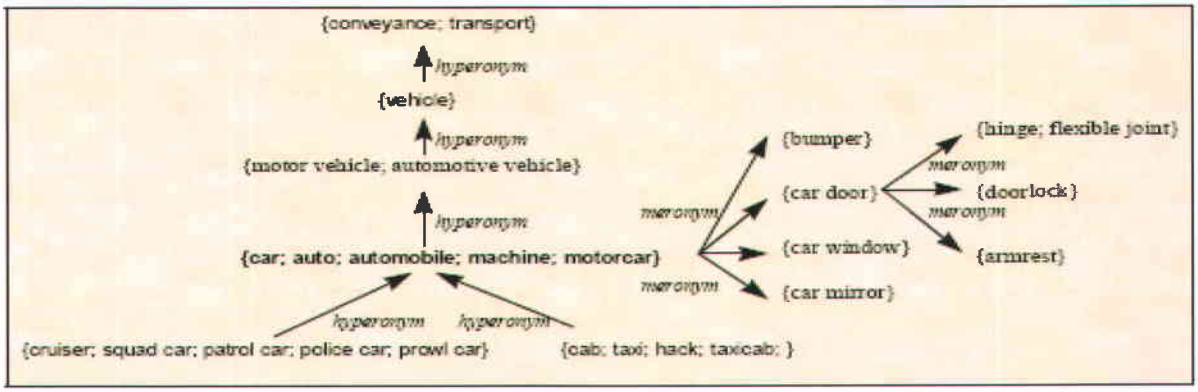
در شکل ۱ بین دسته‌ی هم معنای {car; auto; automobile; machine; motorcar} و سایر دسته‌های هم معنا روابط زیر برقرار است: دسته‌ی هم معنای {motor vehicle ,automotive vehicle} که مفهومی کلی تر رادر بر دارد شامل آن است (شمول معنایی) دسته‌های هم معنای {cruiser; squad car; patrol car; police car; prowl car}و {cab; taxi; hack; taxicab} که مفاهیمی خاصتر را در بر دارند دسته‌های هم معنای زیر شمول را تشکیل می‌دهند. دسته‌های هم معنای تک عضوی {car door}, {car bumper}; {car window} و {car mirror} که اجزای آن را نشان می‌دهند با آن رابطه‌ی جزءواژگی دارند.

۳-۵- اطلاعات آماری شبکه‌ی واژگانی پرینستون

اطلاعات آماری نسخه‌ی سوم شبکه‌ی واژگانی پرینستون به نقل از وبگاه این شبکه[1] در جدول ۲ آمده است:

جدول ۲: آمار واژه ها، دسته‌های هم معنا و معانی

مقوله واژگانی	واژه ها	دسته‌های هم معنا	معانی
اسم	117798	82115	146312
فعل	11529	13767	25047
صفت	21479	18156	30002
قید	4481	3621	5580
مجموع	155287	117659	206941



شکل ۱: روابط معنایی بین دسته های هم معنای مربوط به مفهوم خودرو در نسخه ی ۱.۵ شبکه ی واژگانی پرینستون [۱۰]

[1] http://www.globalwordnet.org/gwa/
Fellbaum@clarity.Princeton.edu

و به بار نشسته است. این طرح ایجاد شبکه‌ی واژگانی زبان فارسی است که گام هایی جهت ایجاد آن برداشته شده است [۱،۲،۳،۴،۵،۶] وایجاد شبکه‌ی واژگانی اسامی زبان فارسی را گامی دیگر در جهت تکمیل این پروژه (که می‌توان آن را ملی ویا حتی فراملی تعریف نمود) دانست. دراین مقاله پس از بررسی شبکه‌های واژگانی مطرح در جهان و بررسی کاربردهای این شبکه‌ها به شبکه‌ی واژگانی فارسی به طور عام و شبکه‌ی واژگانی اسامی زبان فارسی به طور خاص می‌پردازیم.

۲- شبکه‌ی واژگانی

شبکه‌ی واژگانی واژهستان‌شناسی است که حاصل تلاشی در حوزه‌ی روانشناسی زبان است برای بازنمایی آنچه تصور می‌شود در واژگان ذهنی انسان به صورت واژه‌ها و روابط میان آنها وجود دارد. این کار که با استفاده از دسته‌های هم معنا و روابط میان آنها در سال ۱۹۸۵ در دانشگاه پرینستون آغاز گردید، منجر به دادگانی شد که به شبکه‌ی واژگانی پرینستون معروف گردید و نسخه‌ی ۱،۰ آن در سال ۱۹۹۱ عرضه گردید. کاربردهای موفقیت آمیز و روز افزون آن در عرصه‌ی پردازش زبان طبیعی و کمک به زبانشناسان رایانه‌ای در حل بعضی معضلات که مدتها در صدد یافتن راه حلی برای آن بودند، منجر به بیشتر شناخته شدن آن در میان زبانشناسان رایانه‌ای تا روان‌شناسان زبان گردید. فامیان [۲] به نقل از کیلگاریف [۷] آن را موفقیتی عظیم می‌داند که استفاده نکردن از آن نیاز به دلیل و توجیه دارد.

موفقیت شبکه‌ی واژگانی پرینستون چنان تاثیرگذار بود که منجر به ایجاد انجمن جهانی شبکه‌ی واژگانی گردید و وبگاه این انجمن شرایط تبادل نظر در بین علاقه مندان و کاربران و زمینه‌ی اطلاع رسانی از یافته‌ها وکاربردهای جدید این ابزار ارزشمند را فراهم نمود که منجر به بهبود روشهای طراحی وایجاد شبکه‌های واژگانی جدید و همچنین راهکارهایی برای ساخت سریعتر این ابزار ارزشمند برای زبانهای مختلف گردید. در کنار این فعالیتها انجمن درصدد فراهم آوردن زمینه‌ای مناسب برای ایجاد یک پایگاه داده چند زبانه جهانی است. اطلاعات جدید در مورد شبکه‌های واژگانی موجود را توسط لینکی در این وبگاه می‌توان مشاهده نمود که نشان می‌دهد هم اینک برای حدود ۴۰ زبان دنیا شبکه‌ی واژگانی طراحی شده ویا در حال طراحی است. هم اینک می‌توان بر روی این وبگاه ازآخرین نسخه‌های شبکه‌ی واژگانی پرینستون استفاده و یا آن را به صورت برخط دریافت نمود. در ادامه به معرفی مختصر سه شبکه‌ی واژگانی مطرح در جهان شامل شبکه‌ی پرینستون, اروپا و بالکان خواهیم پرداخت.

۳- شبکه‌ی واژگانی پرینستون

این شبکه که پایگاه داده‌ای واژگانی است به سرپرستی جرج ا. میلر بر پایه‌ی یافته‌های متعدد حوزه‌ی روان شناسی زبان درباره‌ی واژگان ذهنی انسان، طراحی گردید. از جمله‌ی این یافته‌ها مطالعات انجام شده درباره‌ی لغزشهای زبانی است که فرد مثلا واژه‌ی "هفته" را بجای "روز" و یا "امروز" را بجای "دیروز" بکار می‌برد و یا جایگزینی هایی مانند "صندلی" بجای "میز" ویا "زانو" بجای "بازو" که نمونه هایی معمول در زبان پریشی می‌باشند (واژه‌های دچار اختلال در شبکه‌ی واژگانی هم شمول خوانده می‌شوند). یافته‌های فوق ویافته هایی از این دست که نشان دهنده‌ی حوزه بندی واژگان می‌باشد و همچنین یافته هایی که نشان دهنده‌ی جدایی مقولات اسم و صفت و فعل و غیره می‌باشند بسیار متعددند و اساس ایجاد شبکه‌ی واژگانی پرینستون قرار گرفته و روابط معنایی میان واحدهای سازنده‌ی شبکه‌ی واژگانی را شکل دادند [۸].

فامیان [۲] به نقل از جی. ای. میلر [۹] طراحی این شبکه را بر پایه‌ی سه فرضیه می‌داند:

الف. جدایی پذیری: واژه‌ها را می‌توان به طور مستقل و جدا از سایر بخشهای زبان بررسی کرد.

ب. الگوپذیری:در ذهن انسان ارتباط واژه‌ها و معنای آنها به گونه‌ای قاعده مند کد گذاری شده است.

ج. جامعیت: پایگاه‌های دانش واژگانی مورد استفاده در زبان‌شناسی رایانه‌ای بایستی به گستردگی واژگان ذهنی انسان باشد تا ابزار مناسبی برای پردازش زبان طبیعی به شمار آید.

میلر و همکاران [۸] شبکه‌ی واژگانی را از جهاتی شبیه و از جهات دیگر متفاوت از فرهنگ واژه‌ها و گنج واژه‌ای موجود می‌دانند. از جمله شباهتها ارائه‌ی تعریف و مثال برای هر دسته‌ی هم معناست که شبیه فرهنگ واژه هاست. اما از جهت دیگر واژه‌ها به صورت دسته‌های هم معنا و شبیه به گنج واژه‌ها سامان دهی شده اند. در نتیجه آنرا نه فرهنگ واژه ونه گنج واژه می‌توان نامید.

اساس دسته بندی واژه‌ها در شبکه‌ی واژگانی دسته‌های هم معناست.

۳-۱- دسته‌های هم معنا یا مجموعه‌های ترادف

اساس نظم دهی اطلاعات در شبکه واژگانی، گروه‌های واژگانی با نام دسته‌های هم معنا یا مجموعه‌های ترادف می‌باشد که در بافتهای خاصی می‌توان آنها را بجای یکدیگر استفاده نمود. برای مثال واژه‌های " اتومبیل "، "خودرو"، "ماشین "، "وسیله‌ی نقلیه" هم معنایند و یک دسته‌ی هم معنا به صورت {اتومبیل،خودرو،ماشین،وسیله‌ی نقلیه} را تشکیل می‌دهند که این دسته‌های هم معنا غالبا همراه با تعریف و مثال یا مثالهایی آورده می‌شود. در کنار این موارد میان دسته‌های هم معنا نیز روابطی مانند شمول و زیر شمول و ... نیز وجود دارد که در قسمت مربوط به روابط معنایی مورد بحث قرار خواهند گرفت.

۳-۲- مقولات واژگانی

در شبکه‌ی واژگانی پرینستون مقولات واژگانی اسم، فعل، صفت و قید به صورت مجزا در نظر گرفته شده‌اند و مقولاتی مانند ضمایر، حروف اضافه، حروف تعریف و موارد مشابه با این فرض که مربوط به بخش نحو زبان بوده و شبکه‌ی واژگانی پرینستون تنها برای مقولاتی که متعلق به طبقه باز می‌باشند تدوین شده است در نظر گرفته نشده اند.

۳-۳- طبقه بندی اسامی

جهت تقسیم بندی اسامی از مجموعه‌ای از آغازکننده‌های معنایی که هر کدام دارای سلسله مراتبی در زیر خود می‌باشند استفاده شده است. سلسله مراتبهای یاد شده هر کدام حوزه‌ی معنایی مجزایی با واژه‌های خود را شامل می‌شوند واز آنجاییکه ویژگی‌های هر آغازکننده به وسیله‌ی کلیه‌ی زیرشمولهایش به ارث برده می‌شود، هر آغازکننده به عنوان جزء اولیه‌ی معنایی برای کلیه‌ی واژه‌های حوزه‌ی معنایی یاد شده در نظر گرفته می‌شود. نقطه‌ی مثبت این روش جهت ساخت شبکه‌ی واژگانی در عمل این است که هر فرد واژه نگار می‌تواند یکی از آغازکننده‌ها را تکمیل نماید که منجر به بالا رفتن سرعت ساخت شبکه‌ی واژگانی می‌گردد. البته نکته‌ی دیگر ارتباط میان مفاهیم تحت آغازکننده‌های

مجله بین المللی
ارتباطات وفناوری اطلاعات

ITRC

جلد یکم - شماره ۳ - تیر ماه ۱۳۸۸ (ش ص . ۱۳-

طراحی و ایجاد شبکه‌ی واژگانی اسامی زبان فارسی

اکبر حسابی
دانشکده‌ی ادبیات و زبانهای خارجی
دانشگاه علامه‌طباطبایی
a.hesabi11@yahoo.com

سید مصطفی عاصی
پژوهشگاه علوم انسانی و مطالعات فرهنگی
s_m_assi@ihcs.ac.ir

مهرنوش شمس فرد
آزمایشگاه پردازش زبان طبیعی
دانشکده‌ی مهندسی برق و کامپیوتر
دانشگاه شهید بهشتی
m-shams@sbu.ac.ir

مهسا عرب یارمحمدی
آزمایشگاه پردازش زبان طبیعی
دانشکده‌ی مهندسی برق و کامپیوتر
دانشگاه شهید بهشتی
m_yarmohammadi@std.sbu.ac.ir

تاریخ دریافت: ۱۳۸۷/۱۱/۱۹- تاریخ پذیرش: ۱۳۸۸/۳/۲۴

چکیده: در این پژوهش طراحی و ایجاد شبکه‌ی واژگانی اسامی زبان فارسی به عنوان یکی از ابزارهای مهم جهت پردازش زبان فارسی در نظر بوده است. در ابتدا ضمن اشاره به تاریخچه و کاربردهای شبکه‌ی واژگانی، بعضی از مهمترین طرحهای شبکه‌ی واژگانی که برای ساخت سایر شبکه‌های واژگانی الگو قرار گرفته‌اند معرفی شده‌اند و برجسته ترین ویژگی(های) آنها بررسی شده است. در انتها به مراحل و چگونگی طراحی و ایجاد هسته‌ی شبکه‌ی واژگانی اسامی زبان فارسی مشتمل بر نوع ویرایشگر، روش ساخت و فرایند ساخت و منابع زبانی بکار رفته پرداخته شده است. در انتها قابلیتهای شبکه‌ی واژگانی اسامی از جمله قابلیت ادغام با شبکه‌های واژگانی صفات و افعال زبان فارسی که پیش از این طراحی شده‌اند و همچنین قابلیت اتصال به سایر شبکه‌های واژگانی زبانهای دیگر مطرح گردیده است.

واژه های کلیدی: شبکه‌ی واژگانی، پردازش زبان طبیعی، روابط واژگانی، زبان‌شناسی رایانه ای، زبان فارسی

Abstract: **This paper discusses the process of designing and developing the Persian noun Wordnet as an important tool for natural language processing. First, an introduction to the most significant projects for building WordNets around the world and their prominent features and applications is provided. Then different stages of designing and developing of Persian noun WordNet including editor, building process and language resources will be discussed. At the end, we will show the capability of connecting the developed noun Wordnet to other Persian WordNets (adjective and verb WordNets) and the WordNets of other languages.**

۱- مقدمه

در دنیای رایانه‌ای امروز پردازش زبان طبیعی افراد زیادی را در حوزه‌های مختلف علم از جمله زبان‌شناسی ، هوش مصنوعی، علوم پزشکی و صنعت به خود مشغول داشته و امروزه متخصصان زبانهای بسیاری به این امر مشغولند. یکی از نیازهای ضروری امروز جهت پردازش زبان فارسی با میلیونها گویشور در کشورهای فارسی زبان ایران،تاجیکستان، افغانستان وسایر کشورها طرحی است که سالهاست در کشورهای دیگر آغاز گردیده