وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

شناسایی طرح های مرتبط با تحقق مترجم ماشینی در زبان فارسی

(گزارش نهایی)

سید مهدی سمائی

تیر ۱۳۹۰

فهرست مطالب

فصل اول: سابقه پژوهش

۱-۱-مقدمه

زبان‌شناسی رایانه‌ای یکی از شاخه های زبان‌شناسی است که در آن فنون و مفاهیم علم رایانه را در تحقیقات حوزه‌های مختلف به کار می برند. ترجمه ماشینی یکی از زیرشاخه-های زبان‌شناسی رایانه‌ای است که نام‌هایی چون ترجمه رایانه‌ای[۱] و ترجمه تعاملی[۲] نیز به آن اطلاق می‌شود. در این شاخه تلاش می‌شود چگونگی به‌کارگیری نرم‌افزار رایانه‌ای برای ترجمه متن یا گفتار از یک زبان طبیعی به زبان دیگر مورد مطالعه و بررسی قرار گیرد.

علیرغم همه محدودیت‌هایی که وجود دارد برنامه‌های ترجمه ماشینی در سراسر دنیا کاربرد دارند. از بزرگ‌ترین کاربران ترجمه ماشینی اتحادیه اروپا است. پروژه مولتو که توسط دانشگاه گوتنبرگ هدایت می‌شود، بالغ بر ۲,۳۷۵ میلیون یورو از اتحادیه اروپا بابت ایجاد یک ابزار ترجمه که بیشتر زبان‌های اروپایی EU را پوشش دهد، دریافت می کند.

گوگل مدعی است که موتور ترجمه جدیدش نتایج امیدبخشی به‌دست خواهد داد. این موتور با روش آماری در گوگل لنگوئج تولز برای ترجمه عربی – انگلیسی و

---

[1] computer-aided translation

[2] interactive translation

چینی- انگلیسی، از طرف موسسه ملی استاندارد و فن‌آوری نمره ۴۲۸۱/ ۰را دریافت کرده است.

نیروهای نظامی ایالات متحده امریکا نیز سرمایه‌گذاری‌های چشم‌گیری روی مهندسی زبان‌های طبیعی انجام داده‌اند. دفتر فن‌آوری پردازش اطلاعات در "دارپا[1]" از برنامه‌هایی چون تایدز و مترجم بابی‌لون استفاده می‌کند. نیروی هوایی امریکا قراردادی یک میلیون دلاری جهت ایجاد فناوری ترجمه امضا کرده است.

کاربردهای وسیع و متنوعی می‌توان برای ترجمه ماشینی تصور نمود. از شبکه‌های اجتماعی اینترنتی گرفته تا کاربرد آن در موبایل برای گویشوران زبان‌های مختلف که با یکدیگر تجارت می‌کنند و مسافرانی که زبان کشوری که به آن سفر کرده‌اند را نمی‌دانند.

[1] DARPA

۱-۲- تعریف مسئله

در طرح حاضر تلاش می‌شود یک طرح کلی برای ساخت مترجم ماشینی در زبان فارسی ارائه شود. لازمه ساختن نرم افزار ترجمه ماشینی در اختیار داشتن تصویری کلی از منطق ترجمه ماشینی و مراحل انجام آن است. علاوه بر این باید مشخص شود که در هر مرحله کدام بخش از زبان مقصد و مبدا مورد تحلیل دستوری قرار خواهد گرفت.

با توجه به بررسی مقدماتی پیشینه مطالعات و تحقیقات انجام‌شده در این زمینه، به نظر می‌رسد که طرح از پیش‌تعیین‌شده و واحدی برای ساختن مترجم ماشینی برای زبان فارسی وجود ندارد که در آن اجزای دستوری زبان فارسی مشخص شده باشد. با دردست داشتن دورنمایی از ترجمه ماشینی می‌توان ارزیابی درستی از پژوهش‌های موجود در این زمینه کرد.

این طرح در دو مرحله ارائه می‌شود. در مرحله اول ضمن یک مرور کلی بر سابقه پژوهش‌هایی که دیگر متخصصان این حوزه انجام داده‌اند به معرفی برخی نرم‌افزارهای موجود پرداخته خواهد شد.

در مرحله دوم طرح کلی ساخت نرم‌افزار مترجم فارسی ارائه خواهد شد.

۱-۳- هدف

ارائه طرح کلی ساخت یک مترجم ماشین در زبان فارسی

۱-۴- پیشینه تحقیق

۱-۴-۱- یک مرور کلی

از سال ۱۹۵۴ به بعد بود که اصطلاح ماشین ترجمه به تدریج به کار رفت (دلونسی[1]، ۱۹۷۲)، ولی عبارت مترجم الکترونیکی و مترجم خودکار را هم گاهی به کار می‌بردند. اما تلاش برای ساختن مترجم ماشینی از دهه سی میلادی شروع شده بود. مترجم ماشینی را آن زمان وسیله ساده ای نظیر ماشین حساب می‌دانستند که به راحتی زبانی را به زبان دیگر ترجمه می‌کند. این ساده‌انگاری اندک اندک جای خود را به واقع بینی داد. یکی از محققان پس از دو دهه تلاش، ترجمه ماشینی را یکی از پیچیده ترین فعالیت هایی دانست که بشر تا آن زمان به آن دست زده است (دریفوس[2]:۱۹۸۹:۵۶). نشانه ها در ماشین های عددی و الفبایی، تک معنا و شفاف هستند و ابهام و شرح وتفسیردر آن‌ها جایی ندارد و اجتماع نشانه نیز ارزش معنایی واحد و مطلقی دارند. در صورتی که در ترجمه زبان بشر، ابهام واژگانی[3] و ابهام ساختاری[4] از عمده مشکلات محققان هستند.

اولین مترجم ماشینی ساده و ابتدایی را یک محقق روسی در سال ۱۹۳۳در مسکو ساخت و به ثبت رساند. در سال ۱۹۴۶ بوت[5] و ویور[6] به تحقیق درباره مترجم ماشینی پرداختند. ویور کار خود را بر اساس روش هایی که با آن، پیام های دشمن رادرجنگ

[1] Delavenay
[2] Dreyfus
[3] Lexical ambiguity
[4] Lexical ambiguity
[5] Boot
[6] Weaver

جهانی دوم رمزشکنی می‌کردند قرار داده بود. بوت ماشینی ساخته بود که کلمات زبان را به زبان دیگر ترجمه می کرد. در کار بوت نه به نحو پرداخته می شد نه به ترتیب کلمات[1]. مترجم ماشینی بوت کلمات مهم متن یا کلیدواژه ها[2]را به زبان دیگر ترجمه می کرد و فهم و تفسیرمتن را به عهده کاربر وا می گذاشت.

ریچنز[3] انگلیسی در سال ۱۹۴۸ فکر تجزیه دستوری خودکار پایانه کلمات[4] را مطرح کرد که نوعی ترجمه کلمه به کلمه بود. ویور در سال ۱۹۴۹ ضمن تایید کار بوت و ریچنز شباهت های زیر ساختی زبان ها با یکدیگر را مطرح کرد؛یعنی شباهت‌های معنایی و منطقی زبان ها که مربوط به ویژگی های مغز انسان هستند و در همه انسان ها مشترک اند. آنچه ویور در نهایت به کار بوت و ریچنز اضافه کرد پیشنهادهایی برای رفع ابهام معنایی بود.

در سال ۱۹۵۰ رایفلر[5] نوعی ترجمه را که انسان با کمک ماشین انجام می داد مطرح کرد.به عقیده او باید ابتدا متن ترجمه می شد و سپس ماشین آن را ویرایش می‌کرد. یعنی به عبارتی، ترجمه ماشینی باید یک پیش ویرایش[6] و یک پس ویرایش[7] می-داشت و تعاملی[8] می‌بود (بوت،۱۹۶۷:۵۳).

[1] Word Order
[2] Key words
[3] Richens
[4] Ending
[5] Reifler
[6] Pre-editing
[7] Post-editing
[8] Interactive

تحقیقات به صورتی گسترده ولی پراکنده ادامه داشت. اسوالد[1] و فلچر[2] در آلمان روی تجزیه نحوی جملات کار می کردند. بنیاد راکفلر در سال ۱۹۵۲ در نخستین همایش هجده نفری زبان شناسان و متخصصان رایانه، سرمایه قابل توجهی را برای ترجمه ماشینی به "موسسه فناوری ماساچوست" (ام آی تی) اهدا کرد. اعضای این همایش موافقت کردند که تحقیقات در دو مرحله انجام شوند: مرحله اول شامل تحقیق درباره بسامد کلمات و شیوه معادل‌یابی آن‌ها و چگونگی استفاده از حافظه رایانه و دیگر جنبه های فنی فرهنگ لغت‌های ماشینی بود؛ مرحله دوم مربوط به شیوه تحلیل های نحوی می‌شد. از سال ۱۹۵۲ تا ۱۹۵۵ محققان آمریکایی درباره ظرفیت حافظه رایانه و شناسایی خودکار معنای کلمات و شیوه تحلیل پایانه های کلمات (پسوندها) تحقیق کردند. در سال ۱۹۵۴ دوسترت[3] و گاروین[4] در دانشگاه جورج تاون به موفقیت‌هایی در ترجمه روسی به انگلیسی دست یافتند. آن ها با ۲۵۰ کلمه و ۶ قاعده، الگویی برای ترجمه ماشینی عرضه کردند. در همان سال ام آی تی مجله "*ماشین ترجمه*" را منتشر کرد.

در سال ۱۹۵۶ ام آی تی اولین نشست بین المللی ترجمه ماشینی را با حضور متخصصان انگلیسی و کانادایی و آمریکایی و روسی برگزار کرد. در این زمان، آمریکا و انگلستان و شوروی محل اصلی تحقیقات ترجمه ماشینی بودند، ولی در ایتالیا و کشورهای اسکاندیناوی هم تحقیقاتی انجام می شد. محققان در این سال ها دیگر تلاش نمی کردند

[1] Osvald
[2] Fletcher
[3] Dostert
[4] Garvin

که ثابت کنند ترجمه ماشینی کاری شدنی است، بلکه در حال نظم دادن و متمرکز کردن پژوهش های پراکنده ای بودند که در نقاط مختلف جهان انجام می‌شد.

درسال ۱۹۵۸ در آمریکا دوازده گروه در کار پژوهش های مربوط به ترجمه ماشینی بودند: دردانشگاه هاروارد فرهنگ لغت ماشینی روسی به انگلیسی را می ساختند؛ چامسکی در ام آی تی درباره ساخت‌های نحوی تحقیق می کرد؛ در دانشگاه جرج تاون در باره نحو و معناشناسی زبان روسی پژوهش‌هایی انجام می شد و در دانشگاه میشیگان تلاش می کردند تا قاعده های نحو روسی را استخراج، و مشکل چند معنایی[1] را حل کنند؛ در لس آنجلس و کالیفرنیا نیز محققان درباره مسائل روش شناختی ترجمه ماشینی تحقیق می کردند. در این بین به ترجمه ماشینی زبان روسی به انگلیسی بیشتر اهمیت داده می شد.

در انگلستان بیش‌تر درباره مسائل روش شناختی ترجمه ماشینی، تحقیقات دقیقی انجام شد. یکی از مهم ترین کارهای پژوهشگران آن کشور، درباره همکاری زبان شناسان و متخصصان رایانه بود. آنان تلاش بسیاری انجام دادند تا زبان شناسی، رشته ای از علوم دقیقه شود و در این راه از روش‌های ریاضی استفاده می کردند.

[1] Polysemy

از دهه پنجاه میلادی و با چامسکی، تحلیل های نحوی مستقل از معنا آغاز شـده بـود، ولـی معناشناسـی را کتـز و فـودور (۱۹۶۳) و کتـز و پسـتال (۱۹۶۴) وارد کـار ترجمـه ماشـینی کردند. در مرحله بعد نیز کاربردشناسی وارد حوزه پردازش زبان شد.

انسان دانشی از عالم خارج دارد که آن را در تعبیر متن به کار می بنـدد. از ایـن رو نظریه‌های گوناگونی درباره شکل تجربه انسان از عالم خارج عرضه شد. یعنی متخصصـان سعی کردند دانش عالم خارج را در حافظه رایانه بگنجانند و عاقبت به ایـن نتیجـه رسـیدند که بهتر است دانش گسترده عالم خارج را محدود کنند. به این سبب به جـای ذخیـره کـل دانش در حافظه رایانه، موضوع عالم فرعی[۱] و نمایشنامه[۲] و فیلمنامه[۳] مطرح شد که همگـی، شکل‌های مختلف ذخیره دانش عالم خارج در رایانه هستند.

۱-۴-۲- بررسی روش‌ها ورهیافت‌ها

همان‌طور که گفته شد در سطح ابتدایی ابتدا به جایگزین‌سازی[۴] واژه‌های زبان مبدا با واژه‌های زبان مقصد پرداخته می‌شود. اما این جایگزینی به تنهایی ترجمه مناسبی به‌دست نمی‌دهد چرا که تشخیص گروه‌ها و و اجزای آن‌ها ضروری است. یکـی از راه حـل‌هـایی متخصصان به‌کارگرفتند استفاده از پیکره[۵] و تکنیک‌های آماری بود.

[1] Subworld
[2] Scripts
[3] Scenario
[4] substitution
[5] corpus

با محدود کردن قلمرو، مثل ساخت یک مترجم در یک حرفه مشخص – مثل نرم افزار گزارش وضع آب‌وهوا- می‌توان امکان جایگزین‌سازی واژه‌ها را محدود کرد. این شیوه بویژه در مواردی که یک زبان رسمی یا منضبط[1] به‌کار می‌رود، مفید واقع می‌شود. به همین دلیل است که ترجمه ماشینی اسناد دولتی و حقوقی در مقایسه با گفتگوها یا متونی که کم‌تر استاندارد هستند، نتایج بهتری در بردارد.

هم‌چنین با دخالت انسان در ترجمه می‌توان انتظار نتایج بهتری داشت. برای مثال چنان‌چه کاربر بتواند دقیقا تعیین کند کدام واژه‌ها اسم هستند، نرم‌افزار قادر به ترجمه دقیق‌تری خواهد بود. با استفاده از چنین تکنیک‌هایی، ترجمه ماشینی ثابت کرده است که می‌تواند به عنوان یک ابزار مفید جهت کمک به ترجمه توسط انسان به کار گرفته شود و حتی می‌تواند خروجی‌های قابل استفاده‌ای تولید کند (برای مثال گزارش وضع آب و هوا).

می‌توان گفت که فرآیند ترجمه شامل دو مرحله رمزگشایی معنای متن از زبان مبدا و به رمز درآوردن همان معنا در زبان مقصد است. در پس این فرآیند به‌ظاهر ساده یک عملیات شناختی پیچیده نهفته است. مترجم به باید تمام جنبه‌های زبان را موردتحلیل و تفسیر قرار دهد. این فرآیند نیاز به برخورداری از دانشی عمیق درمورد جنبه‌های

[1] formulaic

مختلف هر دو زبان (زبان مبدا و مقصد)، مثل صرف و نحو، معنی‌شناسی، اصطلاحات و... و حتی مسائل فرهنگی گویشوران آن زبان‌ها دارد.

باتوجه به آن‌چه که گفته شد یک چالش مهم در ترجمه ماشینی این است که چگونه می‌توان برنامه‌ای طراحی کرد که رایانه متن را مثل یک انسان درک کند، و مثل یک انسان متنی جدید در زبان مقصد تولید کند. رهیافت‌های مختلفی برای حل این مسئله وجود دارد.

ترجمه ماشینی می‌تواند از روشی مبتنی بر قواعد زبان‌شناختی بهره بگیرد. بدین معنی که واژه‌ها به طریقی زبان‌شناختی ترجمه شوند- بهترین ترجمه در زبان مقصد جایگزین واژه‌ها در زبان مبدا شوند.

به عقیده برخی از متخصصان "درک متن" مسئله‌ای مقدم بر ترجمه ماشینی است که موفقیت ترجمه ماشینی در گرو حل آن است.

براساس روش‌های قاعده بنیان[1] ، یک میانجی[2] لازم است که متن ابتدا با آن بازنمایی شده و سپس به زبان مبدا ترجمه می‌شود. بر اساس ماهیت این میانجی، رهیافت "ترجمه ماشینی میان‌زبانی[3]" یا "ترجمه ماشینی انتقال بنیان[4]" نامیده می‌شود. این روش‌ها نیازمند واژگان وسیع، اطلاعات صرفی، نحوی و معنایی و نیز تعداد زیادی قاعده است.

[1] rule-based
[2] intermediary
[3] interlingua machine translation
[4] transfer-based machine translation

معمولا گویشور بومی یک زبان ،در صورتی که داده کافی وجود داشته باشد، از ترجمه یک مترجم‌ماشینی می‌تواند به طور تقریبی به منظور گویشور دیگر پی ببرد. در این روش مسئله فراهم کردن داده کافی و مناسب است. به طور مثال برای روش‌های آماری به حجم عظیمی از پیکره‌های چندزبانه نیاز است.

برای ترجمه زبان‌هایی که با یکدیگر ارتباط نزدیکی دارند می‌توان از روشی به نام "ترجمه ماشینی سطحی[1]" استفاده نمود.

روش‌های قاعده بنیان شامل سه روش انتقال بنیان، میان‌زبانی و فرهنگ بنیان[2] می‌شود.

ترجمه ماشینی میان‌زبانی : یکی از انواع ترجمه ماشینی قاعده بنیان، ترجمه ماشینی میان‌زبانی است. در این رهیافت زبان مبدا به یک زبان میانجی تبدیل می‌شود. سپس زبان مقصد از این زبان میانجی تولید می‌شود.

ترجمه ماشینی فرهنگ لغت بنیان: این ترجمه بر اساس مدخل‌های فرهنگ لغت است. در این روش واژه‌ها مطابق با مدخل فرهنگ لغت ترجمه می‌شوند. در این روش از پیکره‌های متنی دوزبانه مانند کنیدین هنسرد[3]گزارش انگلیسی- فرانسوی پارلمان کانادا، و یوروپال[4] گزارش پارلمانی اروپا استفاده می‌شود. درشرایطی که چنین پیکره‌هایی در اختیار باشد ترجمه متون مشابه نتایج قابل توجهی دربرخواهد داشت، اما به هر حال چنین

[1] shallow-transfer machine translation
[2] dictionary-based
[3] Canadian Hansard
[4] EUROPARL

منابعی بسیار محدود است. اولین مترجم ماشینی آماری کاندید[1] از آی بی ام[2] بود. گوگل برای سال‌های متمادی از سیسترن[3] استفاده کرد اما در اکتبر ۲۰۰۷ به یک روش ترجمه آماری روی آورد. گوگل با افزودن حدود ۲۰۰ بیلیون واژه دقت ترجمه خود را بهبود بخشید.

ترجمه ماشینی مثال بنیان[4]: این روش توسط ماکوتو ناگو در ۱۹۸۴ بنیان نهاده شد [iii]. در این روش ازیک پیکره دوزبانه به عنوان منبع دانش اصلی، استفاده می‌شود. در این روش ترجمه بر اساس قیاس[5] انجام می‌شود.

ترجمه ماشینی پیوندی [6]: این روش ترکیبی است از روش‌های قاعده بنیان و آماری [iii]. برخی شرکت‌های ترجمه ماشینی مانند ایژیا آنلاین[7]، لینگواسیس[8]، سیسترن و یوپی وی[9] مدعی استفاده از چنین روش‌هایی هستند.

این روش با رویکردهای مختلفی می‌تواند به کار رود:

پس‌ویرایش آماری قواعد[10] : ترجمه توسط یک موتور قاعده بنیان انجام می‌شود. سپس از آمار برای اصلاح خروجی این موتور استفاده می‌شود.

---

[1] CANDIDE
[2] IBM
[3] SYSTRAN
[4] Example-based machine translation
[5] analogy
[6] Hybrid machine translation
[7] Asia Online
[8] LinguaSys
[9] UPV
[10] Rules post-processed by statistics

ویرایش آمار با قواعد[1]: به منظور بهبود عملکرد موتور آماری، داده‌ها با استفاده از قواعد پیش‌پردازش می‌شوند. هم‌چنین خروجی موتورآماری نیز توسط قواعد پس‌پردازش می‌شود. این رهیافت از انعطاف و قدرت و کنترل بالایی برخوردار است.

۱-۴-۲-۱- برخی مسائل و مشکلات

امروزه با وجود پیشرفت هایی که در ترجمه ماشینی حاصل شده، هنوز مشکلات رایجی مثل هم معنایی[2]، چند معنایی، ابهام ساختاری، ابهام واژگانی، و حتی محدود کردن عالم خارج و مشخص کردن بافت[3] به جا مانده است. در ابهام‌زدایی از معنای واژه[4] (WSD) به دنبال یافتن مناسب‌ترین معنی برای واژه‌هایی که به چندمعنای مختلف دلالت می‌کنند، هستیم. این مسئله اولین بار در سال‌های ۱۹۵۰ توسط یهوشوآبارهیلل[5] مطرح شد[7i]. او اعلام کرد که بدون یک دائره‌المعارف جهانی[6] یک مترجم ماشینی هرگز قادر به تمیز دادن بین دو معنای مختلف یک واژه نخواهد بود[7]. امروزه رهیافت‌های گوناگونی برای حل این مشکل به کار بسته می‌شوند که به طور کلی می‌توان آن‌ها را به دو دسته سطحی[7] و عمیق[8] تقسیم کرد. به طور کلی می‌توان گفت در دسته اول با استفاده از روش‌های

[1] Statistics guided by rules
[2] Homonymy
[3] Context
[4] Word-sense disambiguation
[5] Yehoshua Bar-Hillel
[6] universal encyclopedia
[7] shallow
[8] deep

آماری، از طریق واژه‌های همسایه واژه مبهم عمل می‌شود و دانش متن در نظر گرفته نمی-شود. و این درست برعکس روش عمیق است که به میزان گسترده‌ای دانش متن درنظر گرفته می‌شود.

از دیگر مشکلات که به ویژه در زبانی چون فارسی مشکل‌آفرین اس ت، آزاد بودن ترتیب کلمات (آرایش واژه‌ها) است. به این معنی که فارسی زبانان مخیرند سازه‌های مستقل را جابه‌جا کنند بدون آنکه به ویژگی دستوری بودن کلمات آسیبی وارد شود . ضمنا رسم‌الخط فارسی به دلایل مختلف در کار ترجمه ماشینی ممکن است اخلال ایجاد کند.

همانطور که گفته شد برای ساختن مترجم ماشینی باید عالم خارج را محدود کرد. یکی از مشکلاتی که مشخص نبودن بافت بوجود می آورد ابهام واژگانی است. گاهی می‌توان با توجه به دانش‌زبانی ابهام را برطرف کرد. مانند جمله زیر:

۱- Jane will coach our team.

در جمله شماره ۱ لفظ coach فعل است و در این بافت فقط فعل می نشیند و مشکلی پیش نمی‌آید. اما اگر دو اسم هماوا[1] با دو معنی متفاوت در دو بافت متفاوت قرار گیرند ابهامی پیش می‌آید که نمی‌توان آن را با دانش زبانی[1] برطرف کرد:

[1] homophone

۲- Two teams were trained by the same coach.

۳- The coach was heading for the station.

در جمله شماره ۳ لفظ coach به معنی وسیله نقلیه است که جز با آگاهی از بافت نمی توان به معنی آن در جمله مذکور پی برد. (روزتا،۱۹۹۴:۳).

ابهام ساختاری نوعی دیگر از ابهام است که به دلیل مشخص نبودن بافت پیش می‌آید. در جمله شماره ۴ بدون کمک بافت نمی‌توان فهمید که آیا دختر دوربین دارد یا مرد:

۴- The girl followed the man with the camera.

یکی از کاربردهای دستورسنج برای امتحان صحت دستوری یک زبان واحد است. یک نمونه از این دستورسنج‌ها را دانشکده زبان‌های دانشگاه هایدلبرگ آلمان در چارچوب طرحی به نام ویرایشگر خودکار پلین[۲] ساخته است. این طرح از آوریل ۱۹۸۹ تا مارس ۱۹۹۴ طول کشیده و مجری آن پیترهلویگ[۳] بوده است (لیزا،۱۹۹۷).

نرم‌افزار دیگری به نام ورپرفکت[۴] وجود دارد که ترکیب‌های دستوری و غیردستوری زبان انگلیسی را از هم تمیز می‌دهد. نرم‌افزار مذکور الگوی این رساله است.

شیوه کار ورد پرفکت صوری و فارغ از بافت است. این نرم‌افزار می‌تواند ایرادهای نحوی و املائی را تشخیص دهد. در صورت مشاهده اشکال در ساخت دستوری جمله پیشنهاد

---

[1] linguistics knowledge
[2] Maschinelle Syntaxprufung und Korrekture PLAIN – C.
[3] Peter Hellwig
[4] word perfect

شده، نوع اشکال را مشخص می‌کند و اگر املای کلمات نادرست باشد تذکر می‌دهد و کلمه‌های هم‌خانواده با کلمه نادرست را جدولی فهرست می‌کند تا یکی از آن میان انتخاب شود. در این‌جا برای نشان دادن شیوه عملکرد نرم‌افزار چند نمونه از داده‌هایی که به برنامه ارائه شده جواب‌های مربوط به آن ذکر می‌شود.

پاسخ نرم‌افزار	داده‌ها
Grammatical	۱- John eats the apple
Grammatical	۲- The apple is going to school
A compound subject requires a plural verb.	۳- John and Mary eats the apple
"The" doesn't usually come before John.	۴- The John eats the apple
Grammatical	۵- John eats
The noun phrase appears incomplete.	۶- John eats the
This doesn't seem to be a complete sentence. An article or other modifiers usually precedes the word "book".	۷- In book
This noun phrase appears incomplete. Check for word formation and punctuation errors for missing word.	۸- I:my
Grammatical	۹- John gave
Grammatical	۱۰- Apple
The subject pronoun "who" should not be used in the object position.	۱۱- The man who you saw eats an apple
The singular subject "man" takes a singular verb not the plural verb "eat".	۱۲- The man who eats the apple eat the dog
Spelling error : eaten, eater, eat, …	۱۳- John eats the apple
Grammatical	۱۴- The man that

از داده‌های نرم‌افزار به نتایجی می توان دست یافت:

الف. از داده شماره ۲ درمی‌یابیم که نرم‌افزار ساخت نحوی ساخت نحوی را ملاک قرار می دهد نه معنی را.

ب. از داده شماره ۳ و ۱۲ می توان اولا به قابلیت نرم افزار در تشخیص نهاد مفرد از نهاد جمع پی برد و ثانیا به این نکته رسید که داده ها فقط با نظریه مارکف[1] به صورت خطی[2] تحلیل نمی شوند و ساختهایی نظیر جمله های پایه و پیرو از هم تمیز داده می شوند.

ج. از داده شماره ۶ و ۷و ۸ می‌توان دریافت که قاعده گروه ها و نرم افزار وجود دارد.

د. داده شماره ۱۱ ثابت می کند که نرم افزار نقش ساختاری فاعل و مفعول را می شناسد.

ه. از داه شماره ۸ می توان به این نکته نیز پی برد که قاعده های علائم سجاوندی[3] در نرم افزار تعبیه شده است.

و. داده شماره ۵ و ۹ ثابت می کند که قواعد زیرمقوله ای تعیین نشده است.

ز. از داه شماره ۱۰ و ۱۴ می توان فهمید که نرم افزار داده ها را با قاعده های جمله سازی انطباق نمی دهد.

لازم به ذکر است که در این رساله قاعده های جمله سازی و قاعده های زیرمقوله ای ارائه خواهد شد.

[1] Markov
[2] linear
[3] punctuation

۱-۴-۳- مترجم ماشینی

یکی از مترجم‌های ماشینی که در آن موضوع ترجمه محدود شده و در واقع بافت مشخص شده است برنامه تاوم متئو[1] است که در هواشناسی کانادا از آن استفاده می کنند. این برنامه که ساخت دستوری آن نیز محدود است می‌تواند اخبار هواشناسی را به صورت کاملا خودکار از زبان انگلیسی به زبان فرانسوی ترجمه می‌کند (روزتا: ۷).

نمونه دیگر وی. ال. اس[2] نام دارد که در سال ۱۹۹۲ ساخته شده و با پیش‌ویرایش عمل می‌کند. یعنی شخصی که محدودیت‌های مترجم خودکار را می‌شناسد متن را قبل از واردکردن در دستگاه تنظیم می‌کند و در واقع نوعی زبان مهارشده[3] بر اساس بافت می‌سازد (روزتا:۸).

از دیگر مترجم‌های خودکار یکی سیسترن است مه در سال ۱۹۹۲ و دیگری متال[4] است که در سال ۱۹۸۴ به کارگرفته شده است. این دو مترجم فقط قادر به ترجمه متون فنی هستند و هر دو به صورت پس ویرایش عمل می کنند. یعنی پس از ترجمه با دست اصلاح می‌شوند (روزتا: ۸).

روزتا دیگر مترجم خودکاری است که از سال ۱۹۸۵ تا ۱۹۹۲ در شهر اندهوون[5] و در پژوهشکده فیلیپس[6] ساخته شده است.

[1] TAUM METO
[2] V.L.S
[3] controlled language
[4] METAL
[5] Einhoven
[6] Philips Research Laboratories

نرم‌افزاری به نام پاور ترانس‌لیتور[1] نیز در ایالات متحده ساخته شده که در ایران نیز در دسترس هست و می‌تواند متون انگلیسی و فرانسوی و آلمانی و اسپانیایی را به یکدیگر تبدیل کند. این مترجم خودکار محدودیت بافت ندارد و از این رو ابهام واژگانی و ساختاری باعث اخلال در عملکرد آن می‌شوند.

همه این مترجم‌ها دارای واژگانی هستند که در آن‌ها ویژگی‌های صرفی و نحوی و معنایی هر واژه ثبت شده است. واژه‌ها بسته به مقولاتی که دارند دارای حوزه قواعد نیز هستند.

امروزه نرم‌افزارهای ترجمه زیادی وجود دارد که برخی از آن‌ها به صورت آن-لاین هم کار می‌کنند. برخی از آن‌ها عبارت‌اند از:

- آنوسارکا[2]: یک مترجم رایگان و متن‌باز[3] انگلیسی به هندی است که براساس دستور پانینی[4] کار می‌کند.
- آپ تک[5]: که سیستم ترجمه پیوندی را در ۲۰۰۹ عرضه کرد.
- عربیک ماشین ترانسلیشن[6]: یک مترجم چندزبانه است.
- آسیا آن‌لاین[7]: یک موتور ترجمه است که ابزارهایی برای ویرایش متن هم دارد.
- گوگل ترانسلیت[8]: یک مترجم آلاین از سایت گوگل است.

[1] Power Translator
[2] Anusaaraka
[3] open source
[4] Panini
[5] AppTek
[6] Arabic machine translation
[7] Asia Online
[8] Google Translate

- گوگل ترانسلیتور تول کیت[1]: وب‌سایتی برای ارائه خدمات جهت ویرایش متن‌هایی است که توسط گوگل ترجمه شده‌اند.

- مترجم ماشینی هندی به پنجابی[2]

- ایدیومکس[3]

- لینگواسیس[4]: یک مترجم با استفاده از روش پیوندی است که می‌تواند از هر زبانی به زبان دیگر برود.

- نروترن[5]: نرم‌افزاری است که قادر به ترجمه کتاب، صفحات وب، ایمیل،فکس، یادداشت، کتاب راهنما، گزارش، نامه و ... به زبان‌های مختلف است.

- پرمپت[6]

- شی شی ترا[7]: یک مترجم اسپانیایی- کاتالونیایی

- تاگل تکست[8]: مترجم انگلیسی- اندونزیایی با استفاده از روش انتقال بنیان

- ورلد لینگو[9]: هم آمار و هم قاعده بنیان است. در مایکروسافت ویندوز و مک آفیس[10] از آن استفاده می‌شود.

- دیلماج

[1] Google Translator Toolkit
[2] Hindi to Punjabi Machine Translation System
[3] IdiomaX
[4] LinguaSys
[5] NeuroTran
[6] Promt
[7] SiShiTra
[8] Toggletext
[9] Worldlingo
[10] Mac Office

- پدیده

هیچ‌یک از مترجم‌های موجود ترجمه کامل و بدون نقص وکاملا خودکار از متن‌های محدود نشده[1] ارائه نمی‌دهد [vi vii vi]. اما به هرحال کیفیت ترجمه وقتی قلمرو[2] محدود می‌شود به شکل قابل‌توجهی بهبود می‌یابد[ix].

[1] unrestricted
[2] domain

۱-۴-۴-پژوهش‌های مرتبط با ترجمه ماشینی

طیبی (۱۳۷۴) در رساله کارشناسی ارشد خود که عنوان آن "کاربرد دستور واژگانی نقشمند در ترجمه ماشینی" است، پژوهشی نظری برای ترجمه برخی تلکس های موجود در سازمان هواپیمایی کشوری کرده است. تحقیق او بر اساس نظریه برزنن[1] یعنی دستور واژگانی نقشمند[2] بوده است.

ثانی (۱۳۷۸) در مقاله‌ای با نام" ترجمه نحوی انگلیسی به فارسی با روش تجزیه چارت "به یک معرفی کلی از روش تجزیه چارت می‌پردازد و سپس مترجمی که آن را LANTنامیده تشریح میکند. ثانی درباره روش تجزیه چارت به نقل از آلن[3](۱۹۹۵) و کرولی[4] (۱۹۹۱)می‌نویسد که این روش یکی از سریعترین روشهای تجزیه نحوی موجود است که همه تجزیه‌های ممکن یک جمله را به طور همزمان به دست می‌آورد و عقبگردنمی‌کند .در این روش برای جلوگیری از تحلیل مجدد سازه‌های مشابه، از یک ساختمان داده‌ای خاص به نام چارت استفاده میشود .تمام سازه‌های مجاز که در بین دو نقطه از یک جمله پیدا میشوند، در چارت نگهداری میشوند تا در صورت لزوم دوباره مورد استفاده قرار بگیرند.

[1] Bresnan
[2] lexical Functional Grammar
[3] J. Allen
[4] G. Krulee

در این مقاله عنوان میشود که در LANT ضمن به‌کارگیری روش چارت، سعی بر این بوده است که دانش ترجمه به گونه‌ای نمایش داده شود که بتوان با معرفی دانش ترجمه زبان‌های مورد نظر، ترجمه از هر زبان دلخواه مبدائی را به یک زبان دلخواه مقصد انجام داد. در این سامانه، دانش ترجمه با زبانی به نام زبان بازنمایی دستور 1 بیان می-شود .به نظر نویسنده این مقاله با استفاده از این زبان می‌توان قواعد دستوری یک زبان را در واحدهایی به نام دستور، واژگان و ساختواژه دقیقا توصیف نمود. LANT با استفاده از قواعد نحوی یک زبان طبیعی که با استفاده از زبان بازنمایی دستور بیان شده-اند، جمله‌های آن زبان را از نظرنحوی تجزیه و تحلیل می‌کند. و برای ترجمه باید علاوه بر قواعد نحوی زبان مبدا، یک طرح ترجمه نیز به کمک زبان بازنمایی دستور تعریف نمود .قواعد ترجمه‌ای این زبان مشابه قواعد گشتاری دستورهای گشتاری است .مترجم قادر است جمله‌های ساده LANT انگلیسی را به فارسی ترجمه کند .نیز این مترجم قادر به ترجمه برخی از الگوهای پیچیده نظیر بند موصولی هم هست.

شمس فرد (۱۳۷۴) نیز در رساله کارشناسی ارشد مهندسی نرم افزار طرحی برای درک متن[1] ارائه کرده است. سیستم دنا قادر به درک متون (محدود شده) فارسی و تبدیل آن ها به بازنمایی داخلی با شبکه وابستگیهای مفهومی بر اساس نظریه شنک می باشد. دنا همچنین قادر به استنتاج نکات ضمنی متن و افزودن این دانش به بازنمایی داخلی است. به

[1] text understanding

این ترتیب می تواند با درک پرسش کاربر پاسخ وی را بر اساس آن چه صراحتا در متن ذکرشده و یا از متن و دانش پیش زمینه سیستم قابل استنتاج است، ارائه نماید.(شمس فرد:۱۳۸۵)

طرح پیشنهادی شمس فرد یک "واژه نامه" دارد که به گفته او "انتظارات" این واژه نامه با "حفره های خالی قاب افعال" پر شده اند. وی با استفاده از "نبشته ها[۱]" سعی کرده است دانش عالم خارج را برای درک متن به کار بندد. او از نظریه وابستگی مفهومی[۲] شنک[۳] برای ارتباط معنایی متن استفاده کرده است.

چنان که مشاهده می شود در رساله مذکور از اصطلاحات رایج زبان شناسی استفاده نشده است. واژه نامه او همان واژگان و حفره های خالی قاب افعال همان چارچوب زیرمقوله ای[۴] و انتظارات همان قواعد زیرمقوله ای[۵] هستند.

دو رساله از محمودی (۱۹۹۰ و ۱۹۹۴) در خارج از کشور ارائه شده است. یکی از این دو رساله در مقطع کارشناسی ارشد و دیگری در مقطع دکتری تهیه شده است. از آن جایی که محتویات رساله کارشناسی ارشد او در رساله دکتری اش وجود دارد فقط رساله دکتری او بررسی خواهد شد. موضوع تحقیق محمودی درباره پردازش گروه اسمی زبان فارسی و نوشتن قاعده های آن بر اساس برنامه پرولوگ[۶] است. او ابتدا به تعیین اجزای

[۱] به جاي نبشته ها لفظ نمايشنامه در اين رساله به كار رفته است.

[2] conceptual dependency
[3] Schank
[4] subcategorization frame
[5] subcategorization rules
[6] Prolog

کلام پرداخته و قائل شدن به طبقای جدید را مطرح کرده است. این طبقات عمدتا باید بر اساس توزیع[1] تعیین شوند. وی سپس به پانزده طبقه واژگانی –نحوی[2] قائل می شود که هر طبقه نیز به طبقاتی فرعی تقسیم می شود. محمودی در نهایت به ساختار گروه اسمی می رسد و الگوریتم های گروه اسمی را ارائه می کند.

شهابی (۱۳۷۶) پس از تحلیل لغوی، ساختواژی و نحوی، جمله‌های فارسی در محدوده معنایی خاص (درخواست اشتغال به کار افراد برای یک سازمان) را با شبکه معنایی بازنمایی می کند. (شمس فرد: ۱۳۸۵)

داوودآبادی و پالهنگ (۱۳۸۴) سیستم دیگری در زمینه پردازش معنایی جملات فارسی ارائه کرده‌اند. این نرم افزار قادر به درک تعداد محدودی از جملات فارسی است. (شمس فرد: ۱۳۸۵)

شمس فرد (۱۳۸۱) یک سیستم یادگیر هستان شناسی از روی متون ساده زبان فارسی، به نام هستی ارائه کرده است. این سیستم پس از تجزیه نحوی جملات، نقش های موضوعی آن ها را استخراج کرده و بر اساس آن ها دانش مفهومی موجد در متن به صورت عناصر هستان شناسی به زبان "کیف[3]" بازنمایی و به یک هسته اولیه هستان شناسی افزوده می-شود. (شمس فرد: ۱۳۸۵)

[1] distribution
[2] Lexico - syntactic
[3] kif

رساله دکتری سیدمهدی سمائی (۱۳۷۷) با عنوان "واژگان در دستورسنج" انگاره نظری برنامه‌ای رایانه‌ای است که با آن می‌توان ترکیب‌های دستوری و غیردستوری زبان فارسی را از هم تمیز داد. به این منظور نگارنده قاعده‌های صوری ساخت جمله‌های خبری ساده مثبت را ارائه کرده است. این تلاش بر اساس فرضیه استقلال نحو چامسکی و الهام گرفتن از نرم‌افزار وردپرفکت انجام شده است. فرض رساله بر این بوده است که این دستگاه باید بتواند ترکیب‌های دستوری و غیردستوری را تمیز دهد و به محدودیت‌های باهم‌آیی ترکیب‌ها توجه نکند. در این کار محدودیت‌ها و قیودی ایجاد می‌شود که برخی از آن‌ها به شرح زیر است: (همان: ۱۳۸)

ناپایداربودن برخی مقوله‌های واژگانی یکی از مشکلات است. این‌که برخی واژه-ها به یک مقوله واژگانی ثابت تعلق ندارند. برای مثال "خوب" متعلق به سه طبقه و"عاقلانه" متعلق به دو طبقه واژگانی مختلف است. این مساله در تحلیل جمله‌هایی از این دست مشکل ایجاد می‌کند:

۱- سیب‌های تمیز و زیبا خوب شسته شدند.

در قواعد آوایی قید شده که اگر پس از صفت‌های مختوم به "-ا" مثل "زیبا" صفت دیگری واقع شود باید یک همخوان میانجی "ی" پس از آن قرار گیرد. در جمله مذکور لفظ خوب پس از زیبا قرار گرفته ولی بین آن‌ها همخوان میانجی نیامده است. ماشین مفروض ممکن است این ترکیب را نادرست بداند و از تشخیص "خوب" در نقش قید

عاجز باشد. فرض کنیم در قواعد حوزه صفت آمده باشد که توالی بیش از دو صفت در گروه اسمی مجاز نیست. طبق این قاعده وقتی دستگاه به "زیبا" برسد این ایراد را مطرح می‌کند که صفت نمی‌تواند خارج از گروه اسمی و به صورت مستقل به‌کار رود و ترکیب غیردستوری است.

یکی از راه‌های مقابله با عناصر مترادف آوردن هم‌نویسه برای این‌گونه عناصر است. دستگاه با دیدن شماره هم‌نویسه در جست‌وجوس عناصر مترادف برمی‌آید. مثلا دو "خوب" در حوزه واژگان هست که یکی صفت و دیگری قید است. با جست‌وجوی رفتار نحوی "خوب" در حوزه قید و انطباق آن با جمله سابق‌الذکر، دستوری بودن کاربرد "خوب" در آن بافت توجیه می‌شود.

یکی دیگر از مشکلات مربوط به تشخیص مرز گروه‌ها از یکدیگر است و در سازه‌یابی[1] جمله زیر چنین مساله‌ای می‌تواند رخ دهد:

۲- پسرهای خوب دانشجویان شهر سیب را خوردند.

مرز گروه اسمی در این جمله پس از "خوب" و "دانشجویان" و "شهر" و "را" می‌تواند باشد:

[[[[پسرهای خوب] ۱ دانشجویان]۲ شهر] ۳سیب را]۴ خوردند

[1] parsing

چنان‌چه مرز گروه اسمی شماره ۱و ۲و ۳ باشد شناسه باید به صورت جمع بیاید ولی اگر شماره ۴ (پس از را) باشد شناسه می‌تواند جمع ئیا مفرد باشدو سازه‌یابی‌هایی از این دست در نهایت می‌تواند تعاملی[1] – یعنی با دخالت انسان- باشد.

نگارنده این رساله در نهایت اظهار داشته است که تلاش این رساله نشان‌دادن انگاره‌ای نظری برای نشان دادن ساز و کار دستورسنج بوده است و نیز نشان دادن مسائلی که در ارتباط با زبان در این کار پیش می‌آید.

"مترجم ماشینی فارسی- انگلیسی : یک نگاه کلی به پروژه شیراز[2]" عنوان مقاله ای است که در آن پروژه ترجمه ماشینی شیراز معرفی شده است. این مقاله در ابتدا یک نگاه کلی به ویژگی‌های زبان‌شناختی زبان فارسی دارد. سپس به شرح مدل شیراز می-پردازد.

نویسندگان این مقاله پروژه شیراز را به عنوان یک پیش‌نمونه برای ترجمه فارسی به انگلیسی معرفی می‌کنند. این پروژه از اکتبر ۱۹۹۷ تا آگست ۱۹۹۹ به طول انجامیده است. این مترجم مجهز به یک فرهنگ لغت الکترونیکی دو زبانه فارسی- انگلیسی شامل حدود ۵۰ ۰۰۰ لفظ و یک تحلیل‌گرصرفی و یک تجزیه‌گرنحوی است. هدف از ساختن

[1] interactive
[2] Persian-English Machin Translation: An Overview of the Shiraz Project

این مترجم بیش‌تر ترجمه اخبار بوده است. این سیستم قادر به واحدسازی[1]، تشخیص کلمات مرکب و فعل سبک [2] است. تجزیه‌گر نحوی قادر به تحلیل گروه اسمی،گروه حرف اضافه‌ای و بند موصولی است.

این مقاله ابتدا اطلاعاتی در مورد زبان فارسی شامل خانواده‌ای که به آن تعلق دارد، نگارش و رسم‌الخط فارسی ارائه می‌دهد. و سپس به طور پراکنده به مسائلی چون ابهام در صرف (به هم‌نویسه‌ها اشاره می‌کند، که البته با جمله "تکواژهای مختلفی که صورت‌های یکسانی دارند" به هم‌نویسه‌ها اشاره می‌کند)، "ساختار فعل‌های سبک[3]" (که در واقع منظور فعل مرکب است) و مسائلی در نحو می‌پردازد. این مسائل عبارت‌اند ازحالت و حالت‌نما، این که در زبان فارسی به استثنای مفعول‌نمای "را" حالت‌نمای مشخصی وجود ندارد و از این‌رو تعیین مرز واژه دشوار است. هم‌چنین این‌که هیچ نشان-گر خاصی که ارتباط بین سازه‌های یک گروه اسمی را مشخص کند وجود ندارد. تنها عنصری که ممکن است چنین نقشی را بازی کند تکواژ اضافه /e/ است که صورت نوشتاری ندارد (مگر این‌که بعد از مصوت‌های /â/ و /u/ بیاید و به صورت واج میانجی

[1] tokenization

[2] Light verb

[3] Light verb construction

/y/ به نگارش دربیاید). ابهام در مرز[1] گروه‌ها نیز مسئله دیگری است که در این قسمت به آن اشاره شده است. تجزیه‌گر نحوی باید مرز گروه‌ها (مثلا گروه اسمی) و نیز نهاد و گزاره را مشخص کند. سپس به چند مورد که می‌تواند به تعیین مرز در گروه اسمی کمک کند اشاره شده است. مثل این که ضمیرها و اسم‌های خاص معمولا آخرین سازه گروه اسمی هستند.

در قسمت چهارم این مقاله ساختار فرهنگ لغت شیراز معرفی شده است. این فرهنگ حدود ۵۰ ۰۰۰ مدخل شامل واژه‌ها، گروه‌ها و اسم‌های خاص است. صورت اسنادی[2] در فرهنگ شیراز برای فعل، مصدر و برای عناصر غیرفعلی، ریشه یا حالت تصریف‌نشده است. هر ورودی به‌طور خودکار به حروف لاتین نوشته می‌شود. معادل سازی برای حروف به نحوی انجام شده است که تا بیشترین حد ممکن همه اطلاعات منتقل شود. طبیعی است که واکه‌های کوتاه که در فارسی نوشته نمی‌شوند در این سیتم هم تبدیل نمی‌شوند.

اجزای کلام موضوع دیگری است که به آن پرداخته شده است. اجزای کلام به دودسته باز و بشته تقسیم می‌شوند. در این فرهنگ لغت اسم‌های خاص، فعل‌ها و فعل‌های

---

[1] boundary

[2] Citation form

سبک به طبقه باز تعلق دارند. و طبقه بسته عناصری چون حروف، لقب‌ها و اعـداد را دربـر می‌گیرد.

فعل‌ها به حالت مصدری در این فرهنگ وارد شده‌اند و از آن جـایی کـه بـه‌دسـت‌آوردن بن‌مضارع از مصدر همیشه ممکن نیست، بن مضارع فعـل‌هـا در یـک دسـته مشـخص وارد شده‌اند. معنی‌واژه‌ها، معانی کم‌بسامد و مترادف‌ها را هم شامل می‌شود. این معانی به ترتیب از پربسامد به کم‌بسامد فهرست می‌شوند. جمع‌های مکسر عربی که در فارسـی وارد شـده-اند، به صورت مستقیم و با ویژگی "جمع" وارد فرهنگ می‌شوند.

قسمت پنجم مقاله با عنوان "نحو فارسی"، با این جمله که "نحوفارسـی در نوشـتار بسیار مبهم است و موجب بروز مشکلات زیادی در تقطیع متن نوشتاری مـی‌شـود"، آغـاز شده است.

ترتیب آزاد کلمات، نبود حالت نمـای مشـخص، اختیـاری بـودن نهـاد و نیـز عـدم نگارش واکه‌های کوتاه به عنوان دلایل این ابهام معرفی شده‌اند.

در ادامه به اجزای برنامه نرم‌افزار پرداخته شده است. فرآیند ترجمه انگلیسی بـه فارسـی در این نرم‌افزار به صورت زیر تشریح شده است:

۱- خوانش و آماده‌سازی متن

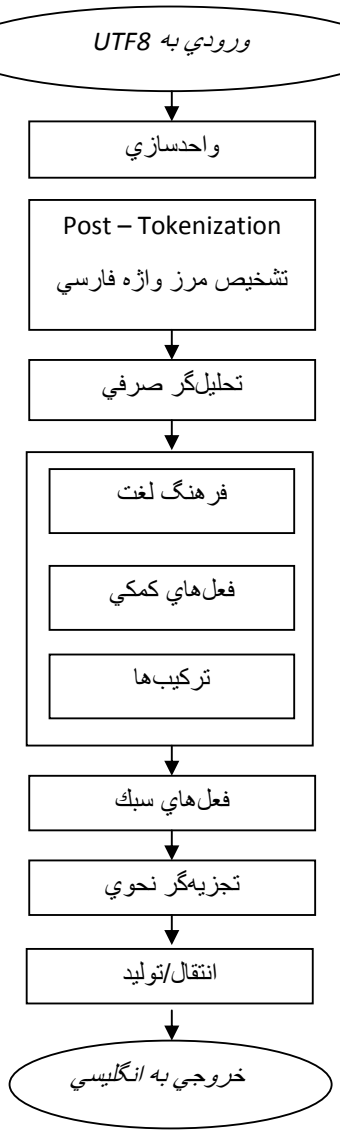
۲- تحلیل صرفی و جستجو در فرهنگ لغت

۳- تجزیه نحوی

۴- انتقال

۵- تولید و آماده‌سازی خروجی به زبان مقصد

نمودار ۱-۱


ورودي به UTF8
واحدسازي
Post – Tokenization
تشخيص مرز واژه فارسي
تحليل‌گر صرفي
فرهنگ لغت
فعل‌هاي كمكي
تركيب‌ها
فعل‌هاي سبك
تجزيه‌گر نحوي
انتقال/توليد
خروجي به انگليسي

برنامه این نرم‌افزار با C++ نوشته شده است. نیز کمی از جاوا بهره گرفته شده است. این نرم‌افزار قادر به ترجمه جمله‌هایی با طول و پیچیدگی متوسط در مدت ۳ تا ۵ ثانیه است.

ترجمه خودکار دوسویه متون فارسی و انگلیسی[۱] مقاله‌ای است که سازوکار مترجم پن‌ترنس[۲] را بررسی می‌کند. پن ترنس یک مترجم متن خودکار دوسویه انگلیسی فارسی است. این مترجم شامل دو بخش اصلی است: پن‌تی۱[۳] وپن تی۲[۴] . پن تی ۱ جمله‌های انگلیسی را به فارسی ترجمه می‌کند و پن تی ۲ مترجم فارسی به انگلیسی است. ابهام‌زدایی از معنای واژه که بخش مهمی از ترجمه است در پن تی یکبا به‌کارگیری ترکیبی از فرهنگ لغت گسترده[۵] و رهیافت‌های پیکره بنیاد، و در پن تی دو با کمک مجموعه‌ای از رهیافت‌های قاعده بنیاد،دانش بنیاد و پیکره بنیاد انجام می‌شود.

این مترجم دارای سه جزء اصلی است. **تحلیل ورودی**[۶] که واژه‌های یک جمله را به همراه اطلاعات مهم درباره این واژه‌ها استخراج می‌کند. **انتقال واژه‌ها**[۷] که هدف آن یافتن بهترین ترجمه برای هر واژه است. و **انتقال ساختاری**[۸] که ترجمه استخراج‌شده را در ساختار زبان مقصد قرار می‌دهد.

---

[1] Automatic Translation between English and Persian Texts
[2] PEnTrans
[3] PEnT1
[4] PEnT۲
[5] Extended dictionary
[6] Input analysis
[7] Lexical transfer
[8] Structural transfer

پن تی ۱ که مترجم انگلیسی به فارسی پن ترنس است، قادر بـه ترجمـه جملـه‌هـای ساده امری، منفی و سئوالی است. ظاهرا منظور از جمله‌های ساده در این نـرم‌افـزار جملـه-هایی است که دارای یک فعل اصلی یا حداکثر دو فعل هستند.

تحلیل ورودی: پن تی۱ در وهله اول از سازه‌یاب سـتنفورد[1] بـرای واحدسـازی، سازه‌یابی و بخش‌هایی از تحلیل صرفی استفاده می‌کند. پن تی ۱ از خروجی ایـن قسـمت که شامل واژه‌ها و اجزای کلام و روابط نحوی سازه‌های مختلف جمله است، استفاده می-کند.

ابهام‌زدایی از معنای واژه زبان مبداء: در زبان مبدا (انگلیسی) از طریق گسترش خوارزمیک (الگوریتم) لسک[2] (لسک:۱۹۸۶) با استفاده از شبکه واژگانی[3] (فـل‌بـاوم[4]:۱۹۹۸؛ بنرجـی و پدرسن[5]:۲۰۰۲) و شبکه واژگانی گسترده[6] (هاراباگیو[7]:۱۹۹۹).

چهارروش برای ابهام‌زدایی در این کار پیشنهاد شده است:

الف: WSD-S : خوارزمیک (الگوریتم) گسترده لسک: بر اسـاس ایـن الگـوریتم واژه-نامه[8] همه معانی واژه مبهم با واژه‌نامه سایر کلمات جمله قیاس می‌شـود. ازمعـانی مختلـف، آن معنی‌ای انتخاب می‌شود که واژه‌نامه‌اش تعداد واژگان مشترک بیشتری با واژه‌نامه سایر

[1] Stanford syntax parser
[2] Lesk algorithm
[3] WordNet
[4] Felbaum
[5] Banerjee and Pedersen
[6] eXtended wordNet
[7] Harabagiu
[8] gloss

کلمات جمله دارد. در مدل به کار رفته در این نرم‌افزار به واژگان موجود در واژه‌نامه وزن‌های مختلف می‌دهند و نهایتا به جای شمارش واژگان مشترک، مجموع وزن این واژگان محاسبه می‌شود.

ب: WSD- E1 : مدل الف، مشکل هم‌نویسه ها و چندمعناها و حتی اجزای کلام مختلف را حل نمی‌کند. از طریق "شبکه واژگانی گسترده" برچسب‌های واژگانی و معنایی استخراج می‌شود. تنها در صورتی کاملا مشابه تلقی می‌شوند که ریشه و مقوله دستوری و معنی آن‌ها مشابه باشد.

ب: WSD -E2 : در این مدل، حرکت بر اساس اولویت صورت می‌گیرد که این ایده از بانرجی و پدرسون (۲۰۰۲) وام گرفته شده است. این کار با هدف فراهم کردن واژه‌های مرتبط و برچسب‌دار بیشتر، صورت می‌گیرد.

ت: WSD -E2 : به منظور رفع مشکل کندی الگوریتم در مورد جمله‌هایی که بیش از ۷ واژه مبهم دارند، از رهیافت معنایی استفاده شده است. در این رهیافت تنها واژه‌هایی که بر اساس خروجی تجزیه‌گر نحوی ستنفورد با یکدیگر رابطه دستوری دارند، با یکدیگر مقایسه می‌شوند.

ابهام‌زدایی از معنای واژه زبان مقصد: برای این قسمت از کار ترجمه یک فرهنگ لغت انگلیسی به فارسی که ترجمه دقیقی از هر معنی ارائه دهد، مورد نیاز است. از آن-جایی‌که نه یک شبکه واژگانی فارسی در ارتباط با انگلیسی و نه ترجمه‌ای از شبکه

واژگانی دایره معنایی[1] موجود بوده است، برای این کار بخشی از یک شبکه واژگانی فارسی طراحی شده است. هم‌چنین از بخش افعال پروژه فارس نت[2] (روحی زاده ۲۰۰۸) برای دستیابی به صورت ماضی و مضارع فعل‌ها بهره برده شده است.

ابتدا معادل فارسی هر واژه از فرهنگ لغتی که نگاشت یک شبکه واژگانی است استخراج می‌شود. چنان‌چه دایره معنایی واژه بیشتر از یک واژه را شامل شود، چند مدخل شامل واژه مبهم و همسایگانش ارائه می‌شود و به گوگل سرچ[3] فرستاده می‌شود تا طبیعی-ترین ترکیب واژه‌ها انتخاب شود.

انتقال ساختاری: در این بخش نمودار درختی جمله به فارسی به دست می‌آید. هر نمودار شامل "گروه[4]"هایی مستقل است که می‌توانند به صورت جداگانه در جمله ترجمه شوند. نمودار درختی فارسی با رهیافت از "پایین به بالا[5]" ساخته می‌شود. به این ترتیب که در مرحله اول ترتیب کلمات بر اساس قواعد دستوری فارسی تعیین می‌شود، سپس جایگاه گروه‌ها نسبت به هم تعیین می‌شود و نهایتا نمودار درختی براساس قواعد نحوی زبان فارسی ساخته می‌شود.

در آخرین مرحله بر اساس نتایج حاصل از WSD ترجمه صحیح هر واژه در جای مناسب (که در بخش نمودار درختی مشخص شده است) قرار می‌گیرد.

---

[1] synset
[2] FarsNet projec
[3] Google search engine
[4] phrase
[5] Bottom-up

به منظور دستیابی به یک ترجمه معقول و معنادار باید اطلاعات صرفی به دست-آمده در فاز پیش‌پردازش، به واژه‌ها اضافه شود. این مرحله که در واقع پس‌پردازش صورت می‌گیرد ، برخی قواعد تصریفی[1] و نگارشی[2] زبان فارسی به‌کارگرفته می‌شود.

پن‌تی ۲: مترجم فارسی به انگلیسی پن تی۲ نیز دارای همان سه بخش اصلی تحلیل داده‌ها، انتقال واژه‌ها و انتقال ساختاری است.

تحلیل ورودی: پن تی ۲ در وهله اول از تجزیه‌گر نحوی کیانی (۲۰۰۹) برای واحدسازی وتحلیل تصریفی استفاده می‌کند. خروجی این مرحله یک سری ریشه واژه‌ها است که بر حسب نقش دستوری(فاعل، مفعول ...) نامگذاری شده‌اند. نیز اطلاعات صرفی این واژه‌ها و برخی برچسب‌ها (مثل اسم خاص) مشخص می‌شود.

انتقال واژه‌ها: هدف این مرحله یافتن بهترین ترجمه برای هر واژه در جمله است. ابهام بزرگترین این مرحله است.

ابهام‌زدایی از معنای واژه: در این مترجم از رهیافتی دانش بنیاد برای حل مسئله ابهام استفاده می‌شود. این رهیافت شامل دو مرحله است:

۱- استفاده از قواعد بینشی[3]: مبنای این قواعد دو فاکتور اصلی است: خود واژه و واژه‌های همسایه و میز سه ویژگی این فاکتورها یعنی نقش دستوری، مقوله دستوری و واژگانی که با آن‌ها باهم‌آیی دارند.

[1] inflectional
[2] orthographic
[3] heuristic

بر این اساس برای قواعد این مرحله ۸ تعریف شده است : خود واژه، واژه‌های همسایه، مقوله دستوری و نقش دستوری واژه و واژه‌های همسایه، جایگاه[1] واژه و نیز مقوله مفهومی واژه‌های همسایه (۲۰ مقوله مفهومی تعریف شده است). برای حل مشکل ابهام، پن تی 2 قاعده مناسب با شرایط واژه مبهم را پیدا می‌کند.

۲- جستجو در اینترنت: چنان‌چه پس از مرحله اول واژه مبهمی باقی مانده باشد، آن-گاه پن تی۲ از اینترنت به عنوان یک پیکره زبانی بزرگ استفاده می‌نماید. این مرحله پس از انتقال ساختاری (که در آن جمله کامل انگلیسی تقریبا استخراج می‌شود) انجام می‌شود. در این مرحله جمله کامل انگلیسی یا بخشی از آن در گوگل مورد جستجو قرار می‌گیرد. هر بار یکی از معانی مختلف واژه مبهم جستجو می‌شود و نهایتا آن معنی که تعداد مثال بیشتری برایش یافت می‌شود، انتخاب می‌شود.

انتقال ساختاری: واژه‌های ترجمه شده باید بر اساس قواعد آرایش واژه‌های انگلیسی مرتب شوند. ابتدا باید سازه[2]های دستوری در جای مناسب قرارداده شوند. علاوه بر این به یک سری فرآیندهای دیگری هم نیاز هست تا ترجمه مطلوب‌تری حاصل شود. به این منظور بیش از ۵۰ قاعده که در ۱۳ مقوله اصلی(بر اساس ۱۳ نقش دستوری اصلی) و

[1] location
[2] constituent

۶ دسته جانبی جای می‌گیرند، تعریف شده است. قواعد اصلی، ترجمه هر نقش دستوری را در جایگاه خود قرار می‌دهند و قواعد جانبی یا برای آماده کردن گروه‌های فارسی برای ترجمه به انگلیسی به کار می‌روند، یا برای بهبود جمله انگلیسی تولیدشده.

مقاله "ابزارهای اساسی برای پردازش متون فارسی[1] " این مقاله ابزارهایی کلیدی برای پردازش متون فارسی را معرفی می‌کند. سامانه معرفی شده در این مقاله STeP-1 نام دارد و قادر است متن فارسی دریافتی را از نظر صرفی تحلیل، و اجزای کلام را نشان-گذاری کند.

مقاله "پردازش متون فارسی: دستاوردهای گذشته، چالش‌های پیش‌رو". مهرنوش شمس فرد. **(۱۳۸۵).** این مقاله به بررسی پردازش متون فارسی و مسائل مطرح در آن می‌پردازد. در این راستا ابتدا به بررسی و تحلیل فعالیت‌های انجام شده در رابطه با پردازش متون فارسی در ایران و سایر کشورهای جهان در سال‌های اخیر می‌پردازد و سپس با طرح مسائل و چالش‌های این امر نگاهی به فعالیت‌های موردنیاز آتی برای رساندن تحقیقات گذشته به کاربردهای آینده دارد.

در مقاله "صرف زبان فارسی از دیدگاه رایانه[2]" یک بررسی از صرف زبان فارسی با دیدگاه کاربرد در رایانه ارائه می‌شود. اجزای کلام و تکواژها نیز در این مقاله بررسی

[1] A Set of Fundamental Tools for Persian Text Processing
[2] Persian Computational Morphology

می‌شوند. هم‌چنین به پارادایم فعلی نیز پرداخته می‌شود. در این گزارش تحلیل‌گر صرفی پروژه شیراز هم بررسی می‌شود و نمونه‌هایی از قواعد آن نیز ارائه می‌شود.

هشام فیلی (۱۳۸۳) در مقاله‌ای با عنوان" ترجمه ماشینی انگلیسی به فارسی بر مبنای نظریه احتمالات و دستور کلمه"نمونه‌ای از یک سامانه مترجم ماشینی انگلیسی به فارسی

مبتنی بر دستور درخت افزایشی هم زمان ارائه میدهد .در نمونه معرفی شده ترجمه صرفا ازانگلیسی به فارسی است، ولی نویسنده معتقد است با وجود ویژگی‌های این مدل امکان به‌کارگیری آن در جهت عکس نیز وجود دارد . اولین مرحله از ترجمه در این روش" فازتجزیه‌گر "نامیده شده که در آن جمله در زبان مبدا مورد تجزیه قرار می-گیرد .مرحله بعدی "فاز انتقال" است که خود شامل سه بخش است .در این فاز یک درخت اشتقاق زبان مبدا را به عنوان ورودی گرفته و یک درخت اشتقاق زبان مقصد را خروجی می‌دهد و در واقع هسته اصلی جمله ترجمه میشود .سه بخش این فاز عبارت اند از : انتقال درختی،انتقال واژگانی و انتقال ویژگی . سومین تحلیل ساختواژی است . در ادامه ترجمه چند نمونه ارائه شده است .در این مقاله دستور درخت افزایشی هم زمان به عنوان روشی کارآمد معرفی شده که نقایص خود را هم دارد .از جمله این که میزان استفاده از اطلاعات معنایی و مفهومی در این مدل ناکامل است، و در نتیجه گاهی قادر به رفع ابهامات واژگانی نیست

۱-۴-۵- جمع‌بندی

تعداد زیادی مقاله وجود دارد که می‌توان هریک را به نحوی با موضوع ترجمه ماشینی مرتبط دانست. از جمله:

- ترجمه ماشینی انگلیسی به فارسی بر مبنای نظریه احتمالات و گرامر لغوی (فیلی:۱۳۸۳)
- پیکره متنی زبان فارسی (بیجن‌خان و مرادزاده.:۱۳۸۳)
- طراحی یک واژگان جامع برای پردازشگر زبان فارسی(نوربالا:۱۳۸۰)

هم چنین فهرستی از پایان‌نامه‌های ارائه‌شده در زمینه "زبان فارسی و رایانه" از ۱۳۵۵-۱۳۸۴ در مقاله‌ای با همین عنوان (ناصح:۱۳۸۵) در مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه ارائه شده است.

بدیهی است ارائه تمام مقالات و پژوهش‌هایی که ممکن است به نحوی با موضوع ترجمه ماشینی مرتبط باشند غیرضروری به‌نظر می‌رسد. چرا که ممکن است به جای کمک به روشن‌ترشدن موضوع موجب سردرگمی شود. تکرار مفاهیم نیز در بسیاری از منابع، دلیل دیگری برای این امر است.

به هر روی تلاش نگارنده براین بوده است که ضمن بررسی منابعی که دسترسی به آن‌ها ممکن بوده، بتواند تصویر مناسبی از روند پیدایش و پیشرفت ترجمه ماشینی در ایران و جهان ارائه دهد. در مواردی که توضیح‌های بیشتر ضروری به نظر می‌رسید

مقاله‌ها به تفصیل مورد بررسی قرار گرفتند، در غیر این‌صورت به ارائه توضیح مختصری بسنده شد.

فصل دوم: ترجمه ماشینی در زبان فارسی

۲-۱- مقدمه

چنان‌که در بخش اول هم توضیح داده شد ترجمه ماشینی یک پژوهش میان رشته-ای است که در آن حداقل به دو تخصص نیاز است. در اغلب منابع مرتبط به این حوزه بر جنبه برنامه نویسی ترجمه ماشینی تاکید شده و غالبا تحقیقات زبان‌شناختی لازم برای ساختن مترجم ماشینی را مفروض انگاشته‌اند.

در فصل حاضر قصد داریم به ارائه طرح های مرتبط با تحقق ترجمه ماشینی در زبان فارسی بپردازیم. واقعیت این است که مواد لازم برای ساختن مترجم ماشینی از زبان مورد تحلیل به دست می‌آید. چنان‌چه زبان مفروض مورد تحلیل دقیق و متقن قرار نگیرد نمی‌تواند براساس آن برنامه رایانه‌ای منسجم و کاملی را عرضه نمود. ساختار زبان مورد تحلیل باید به دقت استخراج و به خصوص در حوزه‌های واژگان و نحو آماده برنامه نویسی شود. یعنی واژگان لازم بسته به سبک مورد تحلیل از پیکره زبان فارسی انتخاب و با ویژگی‌های صرفی و نحوی خود در واژگان قرار می‌گیرد. این واژگان کاربردی نظیر واژگان ذهنی اهل زبان خواهد داشت. در مرحله بعدی باید سازه‌های زبان موردتحلیل، و نحو آن به دقت و صراحت مشخص شود و به شکل زبانی صوری و آماده برنامه نویسی درآید.

۲-۲- زبان فارسی

زبان فارسی به خانواده زبان‌های هندواروپایی تعلق دارد. تعداد سخنگویان زبان فارسی از سخنگویان دیگر زبان‌های جدید ایرانی بیشتر است (ماهوتیان: ۱۳۷۸).

۲-۲-۱-خط فارسی

زبان فارسی با تغییراتی درخط عربی و از راست به چپ نوشته می‌شود و سی‌دو حرف دارد.

از جمله مشکلات خط فارسی می‌توان به موارد زیر اشاره کرد:

- وجود چند نویسه برای یک آوا (س، ث، ص برای /s/، و ت، ط برای /t/)،
- وجود یک نویسه برای چند آوا (ـُ، و برای /o/)،
- چند شکلی بودن حروف با توجه جای قرارگرفتن در کلمه (مثل صـ(در ابتدای کلمه)، ـصـ(در میان کلمه)، ـص چسبیده به انتهای کلمه، ص در انتهای کلمه و جدا) زیرا خط فارسی "پیوسته نویس " است و در نگارش آن حروف به هم می-چسبند.
- عدم نگارش مصوت‌های کوتاه (ـّ، ـِ، ـُ)

از دیدگاه ترجمه ماشینی:

- دشواری در تعیین مرز کلمه (بویژه در مورد کلمات پیچیده مثل کلمات مرکب و اشتقاقی)

۲-۳- ترجمه متن فارسی

به طور کلی برای ترجمه متن، ابتدا باید آن متن در زبان مبدا مورد تحلیل قرار بگیرد. برای ساختن یک نرم‌افزار ترجمه به سه بخش متفاوت نیاز است: یک بخش واژگان، یک تحلیل‌گرصرفی و یک تجزیه‌گرنحوی. انتخاب واژه مناسب و ترجمه جمله با ساختاری صحیح در زبان مقصد، ترجمه نهایی جمله را به‌دست می‌دهد. در واقع پس ازتحلیل متن در زبان مبدا، امکان انتقال آن به زبان مقصد میسر می‌شود. این در حالی است که چنان تحلیلی در زبان مقصد هم امکان‌پذیر است (به این معنی که زبان مقصد هم، به همان میزان مورد تحلیل و مطالعه قرارگرفته و قواعد آن استخراج شده است[۱]).

۲-۳-۱- استخراج واژه‌ها

قبل از جستجوی واژه در فرهنگ، نرم‌افزار باید قادر به تشخیص آن باشد. یعنی باید بتواند مرز کلمه را تعیین کند.

[۱] تجزیه و تحلیل صرفی و نحوی باید در هر دو زبان مبداء و مقصد انجام شود. در اینجا متناسب با موضوع طرح، مثال‌ها در مورد زبان فارسی بررسی می‌شود.

یکی از راهکارهای ابهام‌زدایی در تعیین مرز کلمه، استفاده صحیح از "فاصله" و "نیم‌فاصله" در نوشتن متن است. اگر بین واژه‌ها از "فاصله" کامل، و بین اجزای تشکیل دهنده یک کلمه پیچیده (در صورتی که جدانوشته شود، مثل "می‌روم") از "نیم‌فاصله" استفاده شود (مثل "نرم‌افزار" به جای "نرم افزار") ، نرم‌افزار قادر به تشخیص مرز کلمه خواهد بود.

به هرحال تحلیل‌گر صرفی نرم‌فزار نیز در تعیین مرز کلمه نقش مهمی بازی خواهد کرد. با توضیح بخش‌های ضروری این تحلیل‌گر چگونگی این تاثیرگذاری آشکار خواهد شد.

۲-۳-۲- واژگان[1]

در واژگان اطلاعات آوایی، صرفی، نحوی و معنایی یک کلمه باید گنجانده شود. اطلاعات آوایی شامل نحوه تلفظ واژه است با استفاده از علایم آوانگار بین‌المللی باید در واژگان نگاشته شود. اطلاعات صرفی و نحوی به مقولات دستوری مانند اسم و فعل و ... اشاره دارد. معانی متعدد یک واژه باید بر اساس بسامد کاربرد آن، در واژگان گنجانده شود.

[1] lexicon

۲-۳-۳- تحلیل صرفی و نحوی

تحلیلی نظام‌مند از ساختار زبان را دستور می‌نامند. دستور زبان را از یک دیدگاه می‌توان به دستور تجویزی و توصیفی تقسیم کرد. دستور توصیفی دستوری است که در آن به توصیف زبان آن‌چنان که هست، دست می‌زنند. دستور تجویزی شامل قواعدی برای کاربرد درست زبان در جامعه است (کریستال :۱۹۹۲).

آن‌چه که در تحقیقات زبانی در حوزه پردازش زبان کاربرد دارد باید توصیفی باشد. یعنی توصیف صوری زبان آن‌چنان که هست، نه آنچنان که بایدباشد.

بر اساس تعریف سنتی دستور در زبان‌شناسی، این واژه به مطالعه ساختاری زبان فارغ از آواشناسی[۱] و معنی‌شناسی[۲] دلالت می‌کند و به دو بخش "صرف[۳]" و "نحو[۴]" تقسیم می‌شود(کریستال :۱۹۹۲).

به طور کلی می‌توان گفت که صرف به بررسی اجزای کلمه می‌پردازد و نحو به بررسی ساختار گروه و جمله.

۲-۳-۳-۱- صرف

ذخیره همه واژه‌های یک زبان در حافظه نرم‌افزار اقتصادی به نظر نمی‌رسد. با کمک صرف می‌توان با ادغام اجزا و عناصر موجود در زبان کلمات جدید را تولید کرد.

---

[1] phonology
[2] semantics
[3] morphology
[4] syntax

صرف شامل دو بخش "تصریف" و "واژه سازی" است. تصریف افزودن وند به کلمه است به نحوی که مقوله نحوی واژه نهایی با واژه پایه یکی است. مثل افزودن "ها" جمع به "اسم".

واژه سازی شامل اشتقاق[1] و ترکیب[2] است. کلمات مشتق از یک پایه[3] به اضافه وند ساخته می‌شوند. مانند دانشمند، بی باک. کلمات مرکب از اجتماع اسم و صفت و قید و بن فعل با هم ساخته می‌شوند، نظیر کتابخانه، عقب گرد، خانه نشین.

با توجه به آن‌چه گفته شد یک تحلیل‌گر صرفی بخش مهمی از نرم‌افزار مترجم را تشکیل می‌دهد. برای تهیه چنین تحلیل‌گری به یک مجموعه قواعد مدون نیاز است که با دیدگاه کاربرد در ترجمه ماشینی تدوین شده باشد. تحلیل‌گر مفروض باید قادر به شناسایی کلمات و اجزای تشکیل‌دهنده آن‌ها باشد. به این ترتیب برچسب‌گذاری اجزای کلام نیز میسر می‌شود.

برای توضیح بیشتر آن‌چه که در تدوین قواعد مورد نیاز است، بهتر دیدیم به معرفی دو نرم‌افزار واژه‌شکن و فعل‌شکن بپردازیم که نگارنده با هدف به‌کارگیری در ترجمه ماشینی اقدام به تهیه آن نموده است.

۲-۳-۳-۱-۱- واژه‌شکن و فعل‌شکن: نرم افزارهایی برای تحلیل صرفی

[1] derivetion
[2] componding
[3] base

۲-۳-۳-۱-۱-۱- واژه‌شکن

واژه‌شکن فارسی برای تحلیل کلمات پیچیده ساخته شده است. این نرم افزار کلمات مشتق و مرکب را تجزیه و مقوله آن ها را مشخص می کند.

واژه شکن فارسی اهداف زیر را تامین می‌کند:

الف. تجزیه کلمات مشتق به ریشه و وند (پسوند و پیشوند)

ب. تجزیه کلمات مرکب به عناصر سازنده آن

ج. تعیین مقوله دستوری کلمات مشتق و کلمات پیچیده

ساختمان واژه شکن

واژه شکن در بخش عمده دارد: بخش الفاظ و بخش قواعد. بخش الفاظ شامل انواع مقولات دستوری (اسم، صفت، عدد، قید، ضمیر، بن فعل) است. وندها (پیشوندها و پسوندها) را هم باید جزء این بخش دانست.

الف. بخش اسم ها

در این بخش سی و پنج نوع اسم وجود دارد: اسم، اسم جاندار، اسم معنی، اسم عمل، اسم زمان، اسم خویش، اسم مصدر، اسم مرکب، اسم جاندار مرکب، اسم معنی اشتقاقی، اسم بن ساختی، اسم جاندار بن ساختی، اسم خویش مرکب، اسم جاندار اشتقاقی، اسم عمل اشتقاقی، اسم اشتقاقی، اسم خویش اشتقاقی، اسم ظرف اشتقاقی، اسم مکان اشتقاقی، اسم

مجموعه، اسم زمان اشتقاقی، اسم خویش مرکب بن ساختی، اسم معنی مرکب اشتقاقی، اسم معنی بن ساختی اشتقاقی، اسم مجموعه مرکب، اسم مجموعه بن ساختی، اسم خویش مرکب اشتقاقی، اسم مرکب بن ساختی، اسم مکان مرکب اشتقاقی، اسم مکان بن ساختی اشتقاقی، اسم مرکب اشتقاقی، اسم بن ساختی اشتقاقی، اسم جاندار مرکب اشتقاقی، اسم جاندار بنساختی اشتقاقی، اسم عمل مرکب اشتقاقی، اسم عمل بن ساختی اشتقاقی.

ب. بخش صفت ها

در فهرست صفت ها دوازده نوع صفت تشخیص داده شده است: صفت، صفت مفعولی، صفتمفعولی منفی، صفت مبالغه، صفت مبالغه منفی، صفت مرکب، صفت بن ساختی، صفت ترتیبی، صفت اشتقاقی، صفت مرکب اشتقاقی، صفت بن ساختی اشتقاقی، صفت مرکب بن ساختی.

ج. بخش بن ها

شامل آخرین بن مضارع و بن ماضی افعال. آن دسته ازبن های مضارع و ماضی که تبدیل به اسم شده اند و یا ابتدا اسم بوده اند و به صورت مصدر جعلی درآمده اند در فهرست بن ها واسامی مشترک اند و در هر دو جا درج شده اند.

د. بخش قیود و ضمایر

در این فهرست دونوع قید (قید، قید جهت) و یک نوع ضمیر وجود دارد.

ه. بخش اعداد

در این طبقه اعداد یک تا بیست و هم چنین سی تا صد (به صورت متناوب ده تایی) وجود دارد

بخش وندها

شامل شش پیشوند (مبالغه، دارندگی، منفی ساز۱، منفی ساز۲، اشتراک.)

سی و هشت پسوند نیز در فهرست وندها آمده است: شباهت، آغشتگی، اسم معنی مرکب ساز، مبالغه، تقلید، محافظت، ظرف، شباهت ۱، شباهت ۲، شباهت ۳، شباهت۴، شباهت۵، شباهت۶، تصغیر، تصغیر۱، نسبت، نسبت۱، مکان، مکان۱، مکان۲، مکان۳،مجموعه، کنندگی، کنندگی ۱، عمل، عمل۱، عمل۲، دارندگی، دارندگی۱، دارندگی۲، دارندگی۳، ترتیب، ترتیب۱، جنس، فاعلی، فاعلی ۱، مصدر ساز، مفعولی.

بخش قواعد

در بخش فهرست قواعد انواع امکانات هم نشینی تکواژها و وندها با یکدیگر به صورت قاعده آمده است. در این جا شیوه استخراج قواعد به اختصار ذکر می شود.

نمونه‌ای از قواعدواژه‌شکن:

کلمات اشتقاقی
اسم + وند شباهت = صفت اشتقاقی
اسم مرکب + وند شباهت = صفت مرکب اشتقاقی
کلمات مرکب

اسم + صفت = اسم جاندار مرکب
اسم اشتقاقی + صفت = اسم جاندار
کمات بن ساختی
اسم + بن مضارع = اسم جاندار بن ساختی
اسم اشتقاقی + بن مضارع = اسم بن ساختی اشتقاقی

شیوه استخراج قواعد

واژه شکن همانند مغز انسان عمل می کند. یعنی به جای تکیه کردن بر محفوظات از قواعدی حتی المقدور حشوزدایی شده و تعمیم پذیر استفاده می کند. هر تکواژ (چه تکواژ مستقل و چه مقید) برای قرار گرفتن در کنار تکواژهای دیگر محدودیت هایی دارد. بدین معنی که قبل و بعد از بعضی تکواژها می آید و در کنار برخی دیگر نمی آید. کلمات پیچیده نیز به نوبه خود تابع محدودیت هایی هستند. برای دست یافتن به امکانات و محدودیت های هم نشینی تکواژها و کلمات پیچیده ابتداء کلمات پیچیده شامل ترکیبات اشتقاقی و غیر اشتقاقی و کلمات مشتق مرکب و ترکیبات بن ساختی از فرهنگ معین استخراج شد. علت انتخاب فرهنگ معین برای استخراج داده‌ها این بود که ساخت‌های قدیمی و و مهجور نیز در این فرهنگ وجود دارد. در مرحله بعد فهرستی از اجزای تشکیل دهنده ترکیبات استخراج شده تهیه شد. یعنی مقوله های متعارف دستوری نظیر اسم و صفت و قید و ضمیر و غیره و هم چنین وندها (پیشوندها و پسوندها) که در ساختمان ترکیبات اشتقاقی و غیر اشتقاقی و بن ساختی و مشتق مرکب به کار رفته بودند جدا و طبقه

بندی شدند. داده‌های استخراج شده اولیه در طی کاردچار جرح و تعدیل های فراوانی شدند که به مقتضای کار بود.

اجزاء استخراج شده از ترکیبات پیچیده یک به یک در بافتی که به کار رفته بودند و بافت‌های بالقوه‌ای احتمال بکار رفتن آن ها بود بررسی شدند و به هریک برچسبی داده شد . در این مرحله مشخص شد که مقوله‌های متعارفی که محققان تا کنون در آثارشان تشخیص داده و نامگذاری کرده اند کفایت نمی کند و باید از مولفه های مکمل برای نامیدن اجزاء آن استفاده کرد. در تحقیقاتی که تا کنون درباره مقوله های دستوری شده است از مقوله های اسم و صفت و ضمیر و قید و حرف اضافه و حرف ربط و اصوات نام برد اند. محمودی (۱۹۹۴) از جمله محققانی است که به پانزده مقوله دستوری در زبان فارسی قایل شده است. آن چه در کار واژه شکن اهمیت داشت تعداد کل مقوله ها نبود بلکه مقوله های فرعی هر مقوله بود. بدین معنی که مقوله اسم بسته به نقشی که در هر کلمه پیچیده داشت باید مولفه ای جداگانه می گرفت و برچسب جدیدی به آن نسبت داده می شد. محمودی عقیده دارد که مقوله های دستوری باید بر اساس توزیع تعیین شوند (همانجا). استخراج مولفه های هر مقوله بر اساس همین نظر محمودی انجام نشد. بدین ترتیب برای هر مقوله به مقوله هایی معنایی در چارچوب مقوله اصلی قائل شدیم و بر اساس آن ها مقوله ها به چند مقوله جدید تقسیم شدند. سی و پنج نوع اسم و دوازده نوع صفت حاصل همین کار بودند.

برای قاعده نویسی دو کار انجام شد:

الف. قاعده هم نشینی مقوله ها به زبان طبیعی نوشته شد.

ب. قاعده ها به صورت حروف قراردادی لاتین درآمدند تا مناسب برنامه نویسی رایانه ای شوند. مثلا از حروف s و n و a وغیره و اندیس های عددی برای نشان دادن قواعد استفاده شد.

نمونه:

Formula	
Formula	Small Formula
A01I01	N12
A01I02	A07
A01N01	N09
A01N02	N09
A01N03	N09
A01N04	N09
A01N05	N09
A01N07	N09

Formula Descriptions	
Code	Description
A01	صفت
A02	صفت مفعولی
A03	صفت مفعولی منفی
A04	صفت مبالغه
A05	صفت مبالفه منفی
A06	صفت مرکب
A07	صفت بن ساختی

ویژگی های واژه شکن

یکی از مشکلات نظری در تهیه واژه شکن مسئله دستوری و غیر دستوری بودن ساخت ها بود. غرض از ساختن واژه شکن آن بود که نرم افزار مفروض بتواند ساخت های دستوری را از ساخت های غیردستوری را تمیز دهد و سپس الگوی ساخت آن را عرضه کند.

بنابراین مشکل اصلی این بود که واژه شکن چه محدوده ای را مجاز و چه محدوده ای را غیر مجاز بداند.

سئوال این بود که اگر قرار باشد ساختی را غیردستوری بدانیم معیار ما چه خواهد بود . چامسکی در ساخت های نحوی آن جا که درباره استقلال دستور زبان بحث می کند به این نکته می پردازد که چگونه جمله دستوری را از جمله غیردستوری تمیز دهیم (سمیعی، ۱۳۶۲، ص ۱۴ تا ۲۰) : در رشته های دستوری و غیر دستوری را در عمل بر چه پایه هایی از یکدیگر باز می شناسیم؟ .... " پاسخ های چندی که بلافاصله به ذهن خطور می کنند نمی توانند درست باشند ... دستور زبان پرتو رفتار گوینده است که بر پایه تجربه محدود و تصادفی خود از زبان می تواند تعداد نامحدودی جمله تازه تولید کند.... در حقیقت هر توضیحی درباره مفهوم دستوری بودن در زبان L را می توان بدست دادن توضیح این جنبه اساسی از رفتار زبانی شمرد. " (همانجا ص۱۵). به نظر چامسکی " مفهوم دستوری بودن را با مفهوم بهره ور بودن یا معنی دار بودن.....نمی توان همسان دانست." (همانجا، ص ۱۵)

او جستجو " برای یافتن تعریف دستوری بودن بر پایه معنی شناسی" را کاری عبث می داند (همانجا، ص ۱۵). از طرفی دستوری بودن را تابع بسامد و " از لحاظ تخمین آماری در حد بالا بودن" نیز نمی داند (همانجا، ص ۱۶)[1].

[1] نگارنده این گزارش در رساله دکتری خود در پی طراحی ماشین مفروضی بوده که جمله های دستوری را بدون توسل به معنی ازجمله های غیردستور تمییز دهد. یعنی بر مبنای نظریه چامسکی در کتاب *ساخت های نحوی*.

سه نکته مهم در نقل قول های مذکور وجود دارد:

الف. رفتار زبانی یکی از ملاک های تمییز جمله دستوری از غیر دستوری است.

ب. دستوری بودن را می توان فارغ از معنی بررسی کرد.

ج. بسامد وقوع تاثیری در دستوری یا غیردستوری بودن ندارد.

این بحث گرچه در حوزه نحو صوت گرفته است تعمیم آن به حوزه صرف هم امکان پذیر است ومی توان نتیجه گرفت که:

می شود در مورد معنی ترکیبات پیچیده تا حدی با تسامح برخورد کرد و غیرمتعارف بودن معنی آن ها را دلیل غیردستوری بودن ترکیبات ندانست.

واژه شکن با این دیدگاه ساخته شده است. یکی از تبعات برگزیدن چنین دیدگاهی این است که بسیاری از ساخت های بالقوه را که خوش ساخت[1] هستند ولی چه بسا تا کنون کسی به کار نبرده است می توان دستوری دانست. دستوری دانستن ترکیباتی نظیر چوب نشین و گرفتارانه که به قیاس با خانه نشین و دانشمندانه ساخته شده اند به همین علت است.

[1] well- formed

بی نظمی موجود در ساخت های پیچیده (به خصوص بی نظمی ترکیبات اشتقاقی) نیز دلیل موجهی برای دستوری دانستن ترکیبات پیچیده بالقوه‌ای است که معنی غیرمتعارف دارند مثلا پسوند " –ا " را می توان به صفت های " پهن" و "دراز" اضافه کرد و "پهنا" و " درازا" را ساخت. چنانچه این قاعده را تعمیم ندهیم و ساخت هایی نظیر "نازکا" و "سختا" و " نرما" را دستوری ندانیم ناچار خواهیم شد به استثناءهای زیادی قائل شویم که خلاف نظریه های یادگیری زبان است. یعنی مجبور خواهیم شد همه این قاعده ها را تعمیم ناپذیر فرض کنیم و ترکیبات موجودی را که با این قاعده ساخته شده اند مستقیما در حافظه رایانه قراردهیم. هرچند در مواردی نیز چاره ای جز قراردادن مستقیم ترکیبات پیچیده در حافظه وجود نداشته است. مثلا ترکیباتی که با بن ماضی و پسوند " –ار" ساخته می شوند چنان از لحاظ معنایی ناهمگون اند که نمی توان برای آن ها قاعده عام داد و باید آن ها را مستقیم در حافظه گذاشت.

یکی دیگر از مشکلات نظری این بود که آیا قواعد باید دوری[1] باشد یا خیر.

قواعدی را که می توانند تکرار شوند و جمله بسازند قواعد دوری یا مکرر[2] می‌نامند (کریستال، ۱۹۹۲ ، ص۳۲۸). مثلا قاعده قرارگرفتن صفت بعد از اسم در زبان فارسی قاعده ای دوری است زیرا منطقا می توان گفت: این کتاب قطور خواندنی جذب سبز.....

[1] recursive
[2] iterative

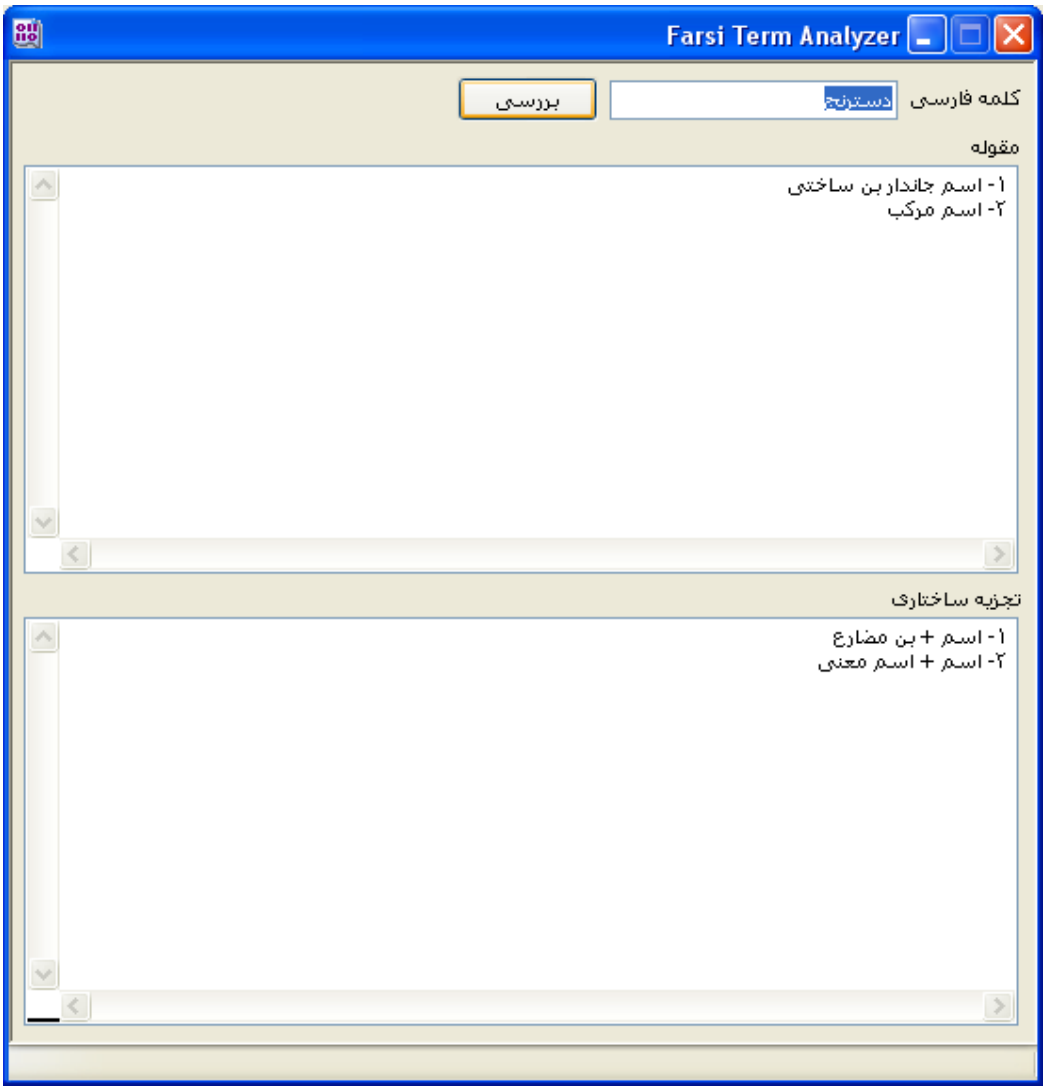
این مسئله در کار واژه شکن که در حوزه صرف عمل می کند نیز صدق می کند. مثلا اسم + وند تصغیر (کتاب + ک) ←کتابک، که کتابک خود اسم است و منطقا بار دیگر با پسوند " – ک" ترکیب می شود و کتابکک را می سازد و الی آخر. یا اسم جاندار + وند مبالغه (کبوترباز) اسم جاندار که " کبوتر باز" خود اسم جاندار است و منطقا بار دیگر با پسوند " – باز " ترکیب می شود و بار دیگر "کبوتربازباز" را می سازد وقس‌الهذا. در زبان عادی کاربرد مانع از دوری شدن ساخت‌ها می شود. در کار واژه‌شکن مشکل به این صورت حل شد که برای کتابک و کبوتر باز نامی به جز اسم و اسم جاندار داده شد (اسم تصغیر اشتقاقی و اسم جاندار مرکب) و قاعده دیگری برای ترکیب این ساختهای جدید با پسوند "ـ ک" و "ـ باز" داده نشد تا مانع از دوری شدن قواعد شود.

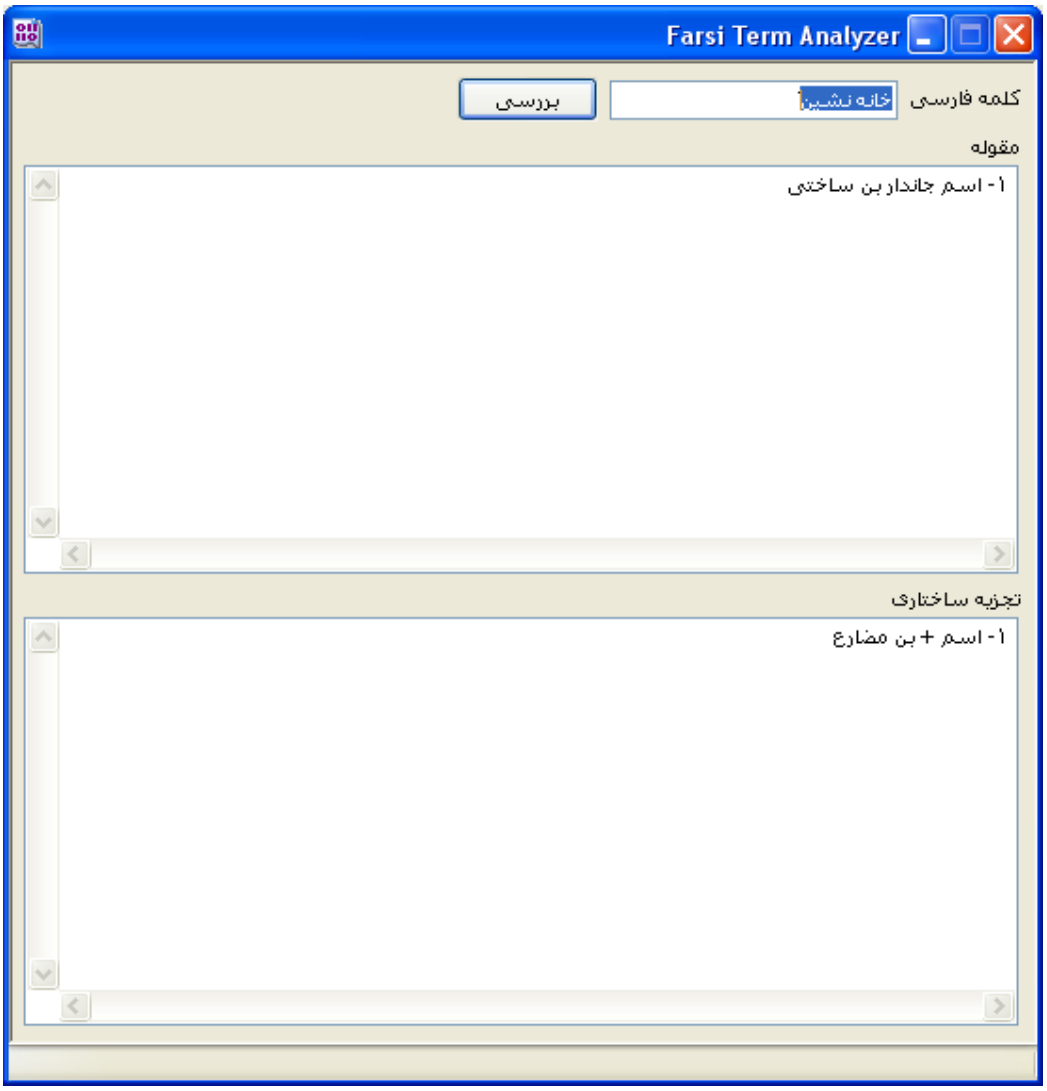
قائل شدن به دوری بودن قواعد در کار واژه شکن باعث ساخته شدن ترکیباتی دستوری می شود. اما از آنجایی‌که این ترکیبات انتهایی ندارند و در کارائی واقعی واژه شکن اخلال به وجود می آورند، این ویژگی مهار شده است. بدین معنی که قواعد تکرارپذیر نیستند و در مرحله ای متوقف می شوند.

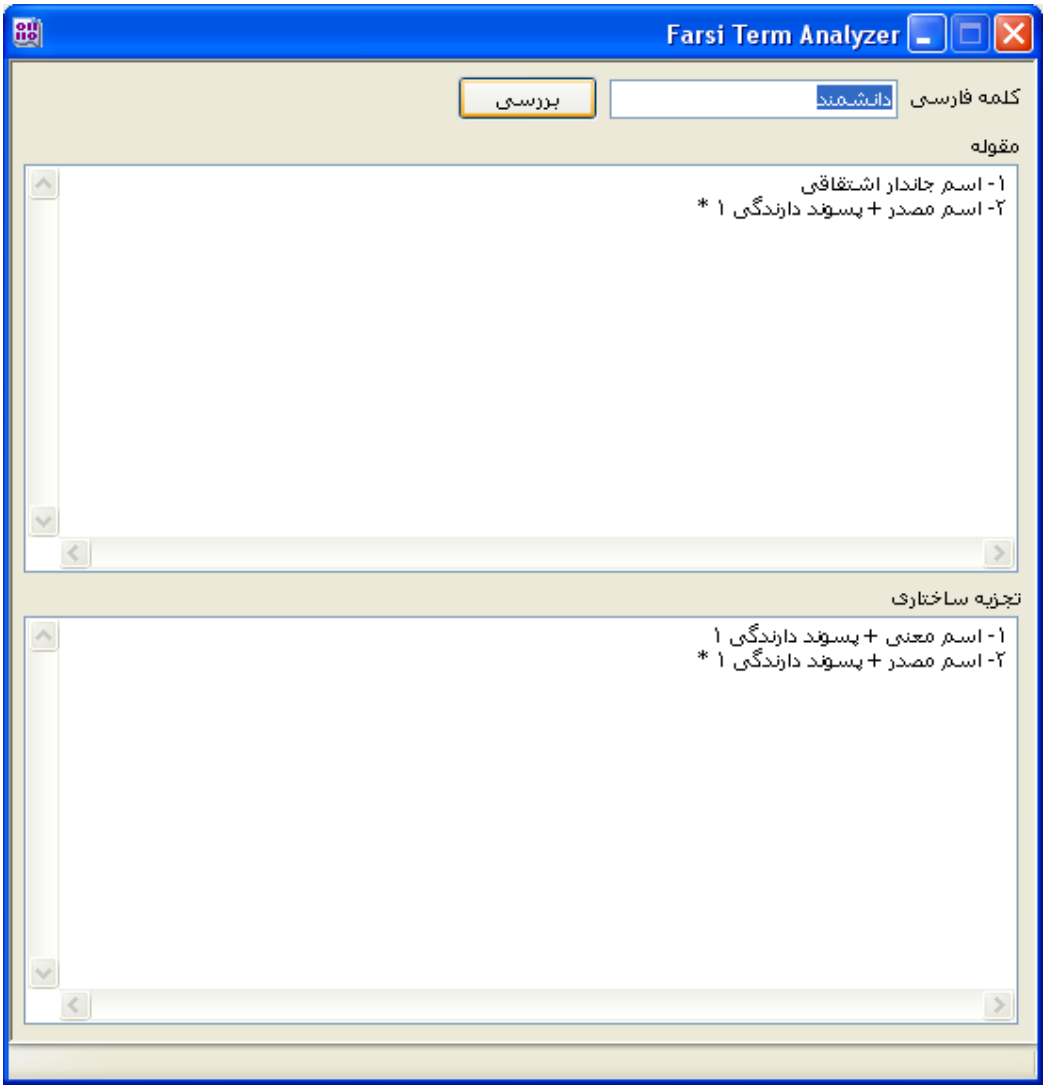
شیوه تجزیه ترکیبات و تعیین مقوله دستوری

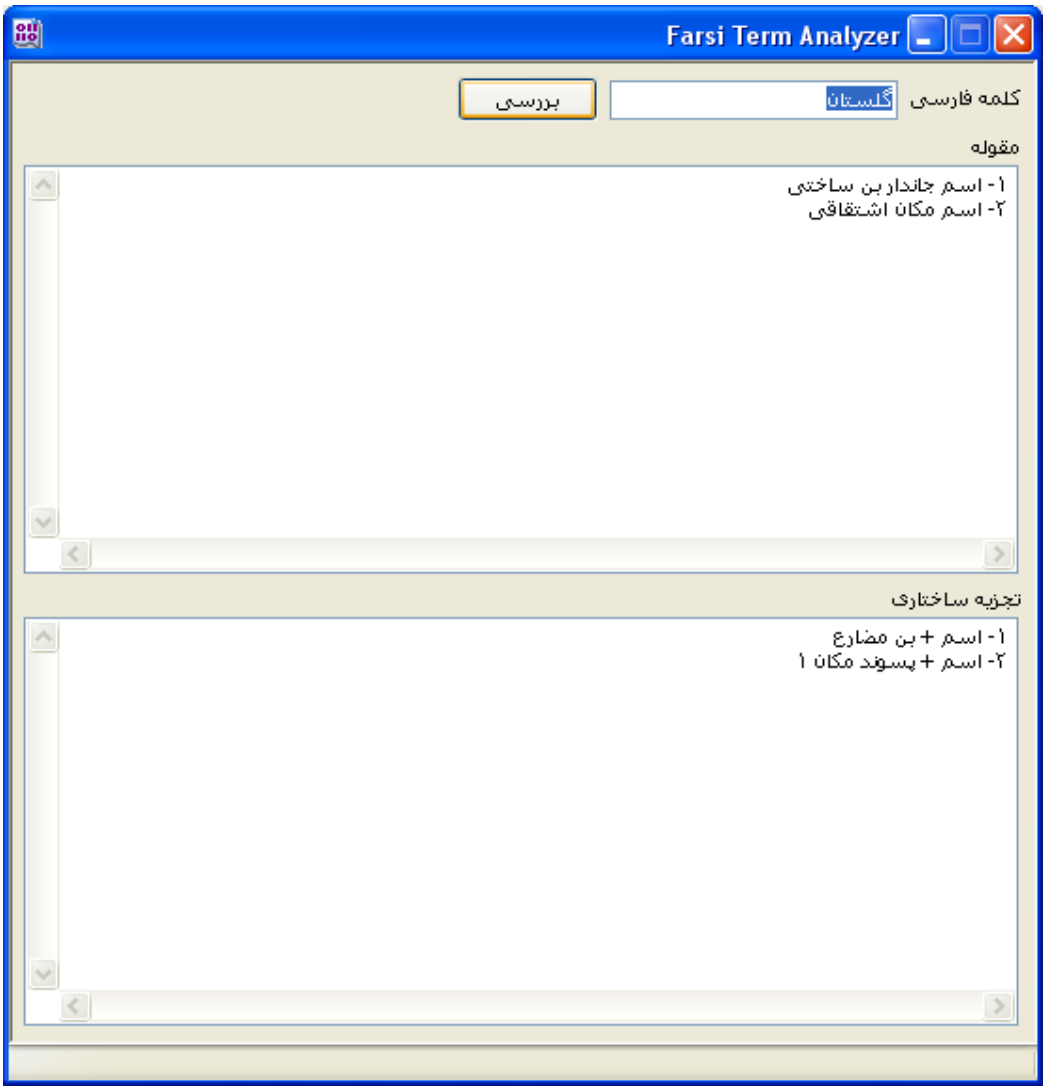
واژه‌شکن ابتدا مقوله کلمه پیشنهادی را مشخص و سپس آن را به اجزاء سازنده‌اش تجزیه می کند. درصورتی که قاعده ساخت کلمه پیشنهادی در حافظه رایانه موجود نباشد فقط به

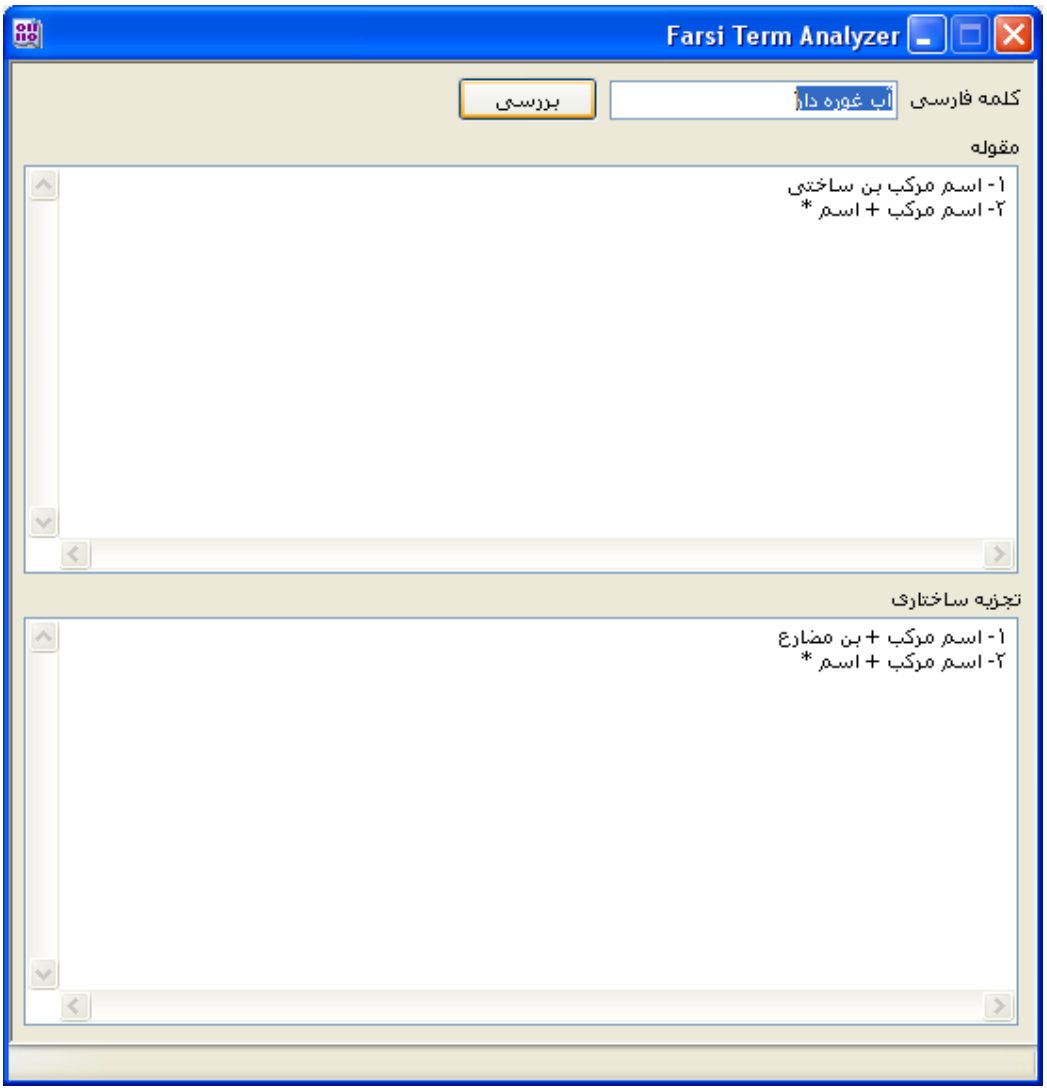
تجزیه کلمه مذکور اکتفا می شود. همه این موارد مشروط به این معنی است که عناصر سازنده کلمه مفروظ در حافظه موجود باشد. در صورت موجود نبودن کلمات پیشنهادی در حافظه واژه شکن می توان آن کلمه را به فهرست افزود و سپس آن را برای تحلیل به واژه شکن سپرد. مثال:

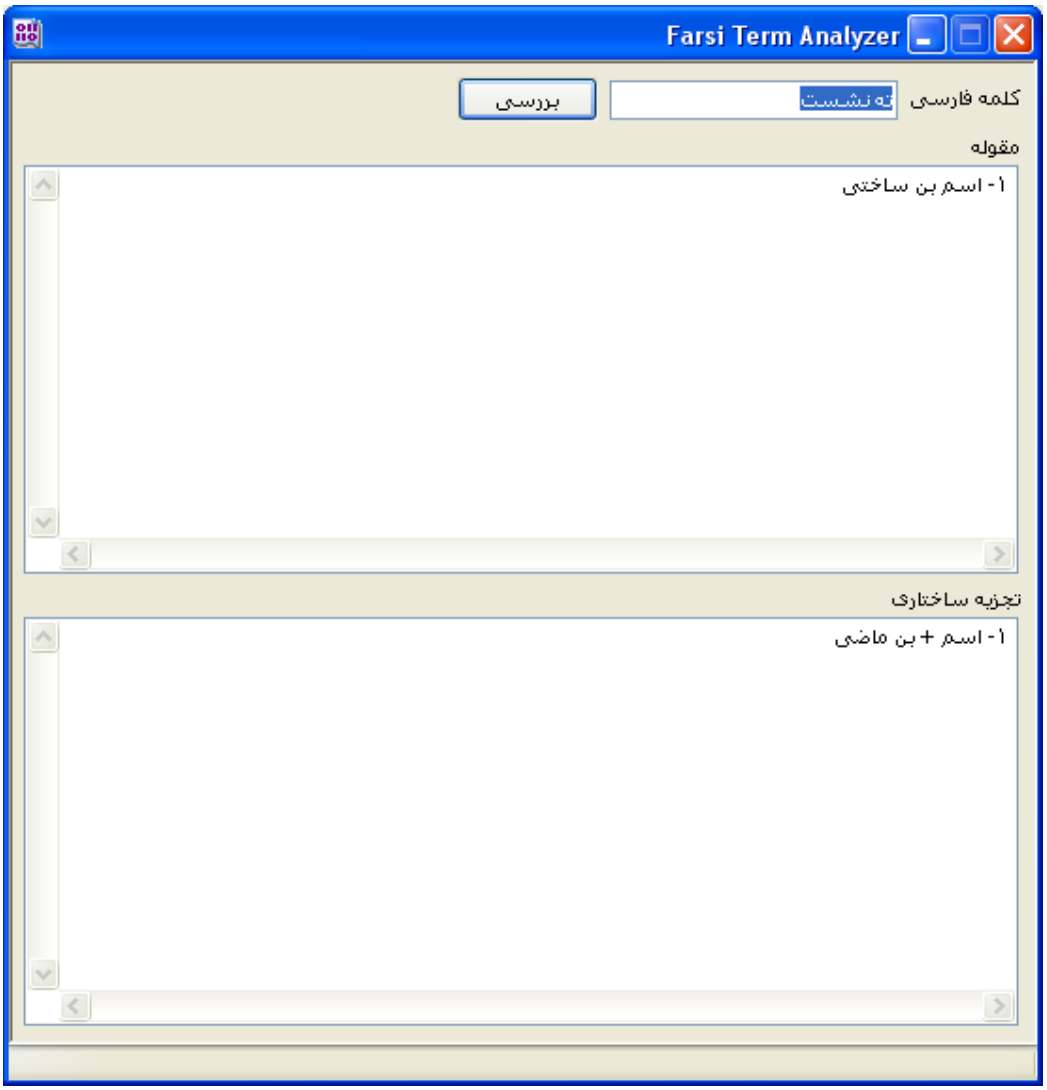


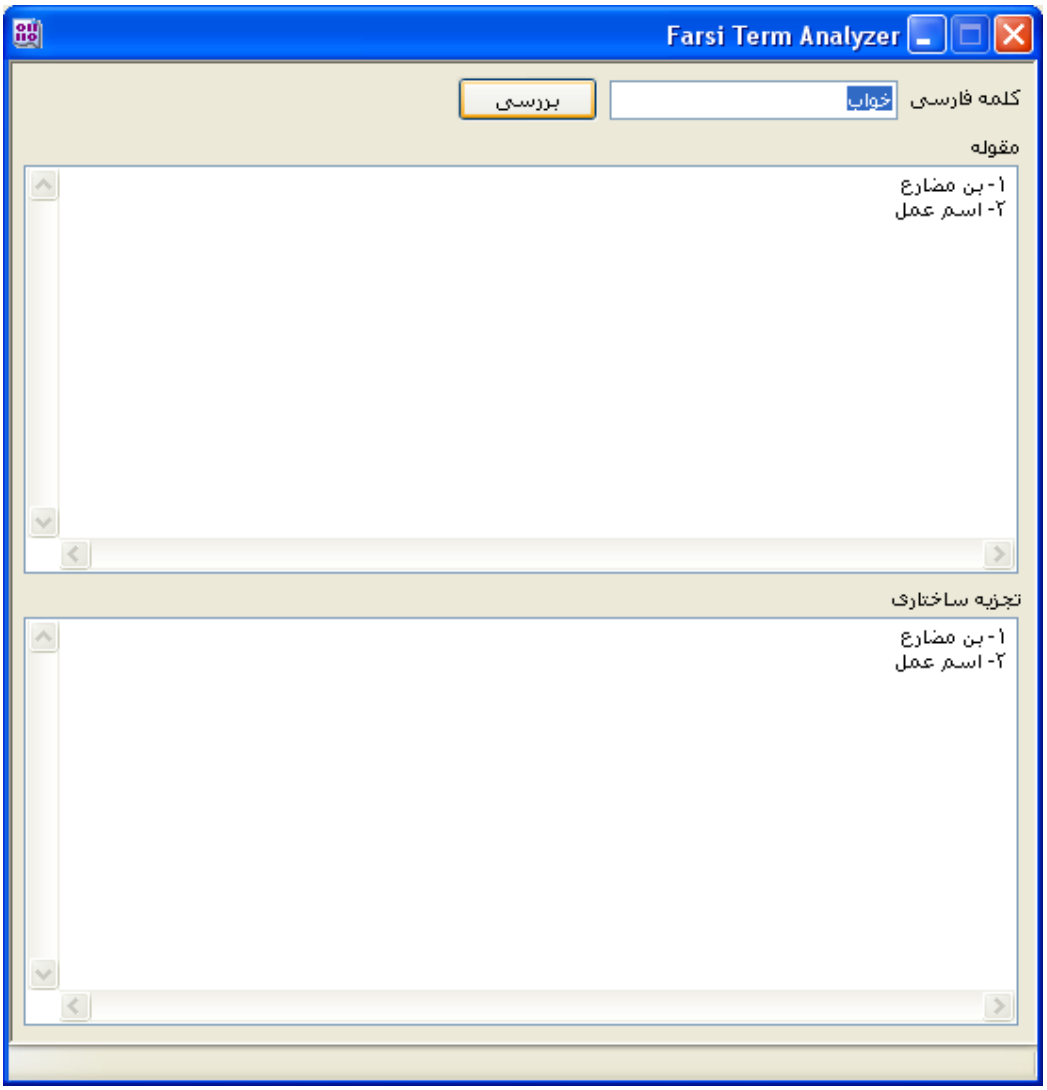

Farsi Term Analyzer
کلمه فارسی
خانه نشین
بررسی
مقوله
۱- اسم جاندار بن ساختی
تجزیه ساختاری
۱- اسم + بن مضارع

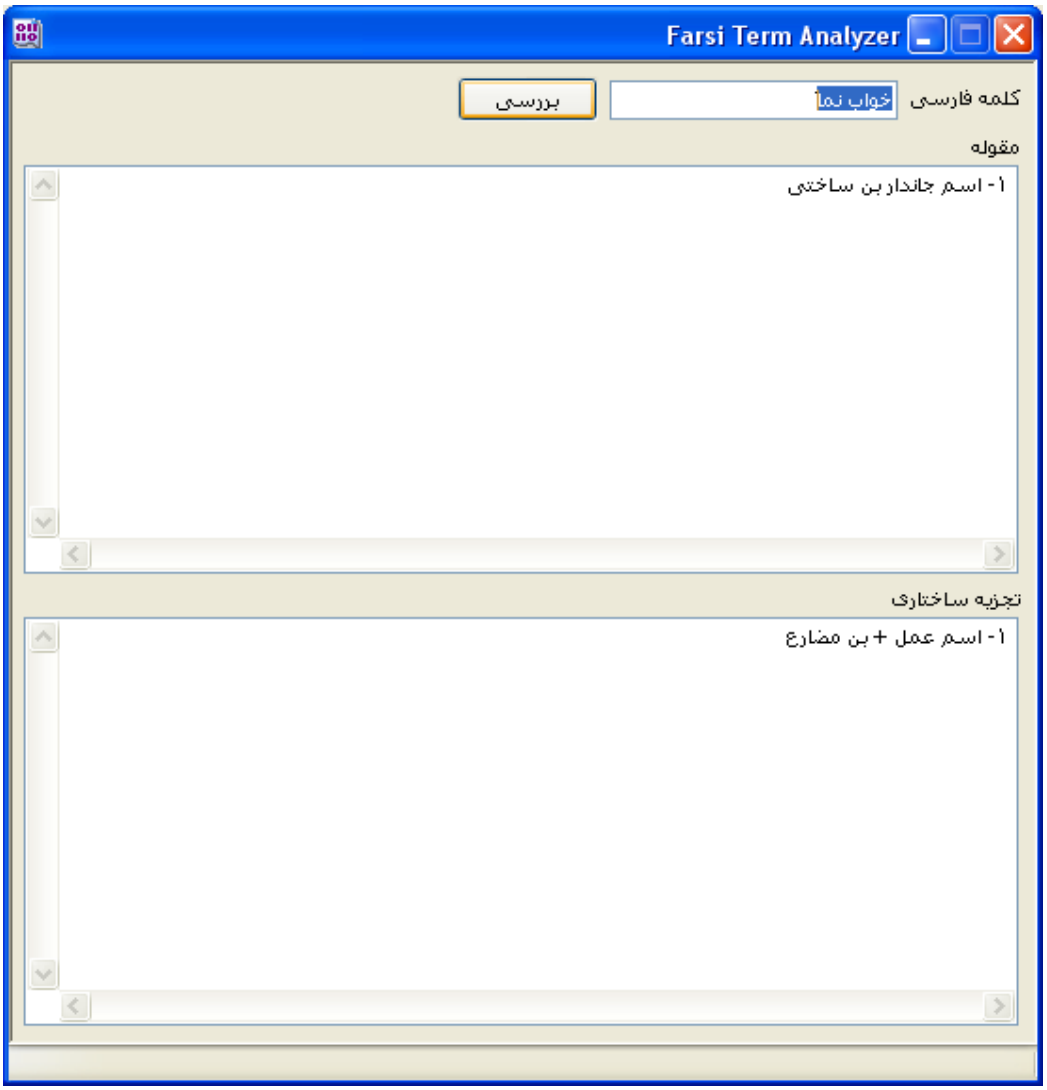

Farsi Term Analyzer
کلمه فارسی
دانشمند
بررسی
مقوله
۱- اسم جاندار اشتقاقی
۲- اسم مصدر + پسوند دارندگی ۱ *
تجزیه ساختاری
۱- اسم معنی + پسوند دارندگی ۱
۲- اسم مصدر + پسوند دارندگی ۱ *

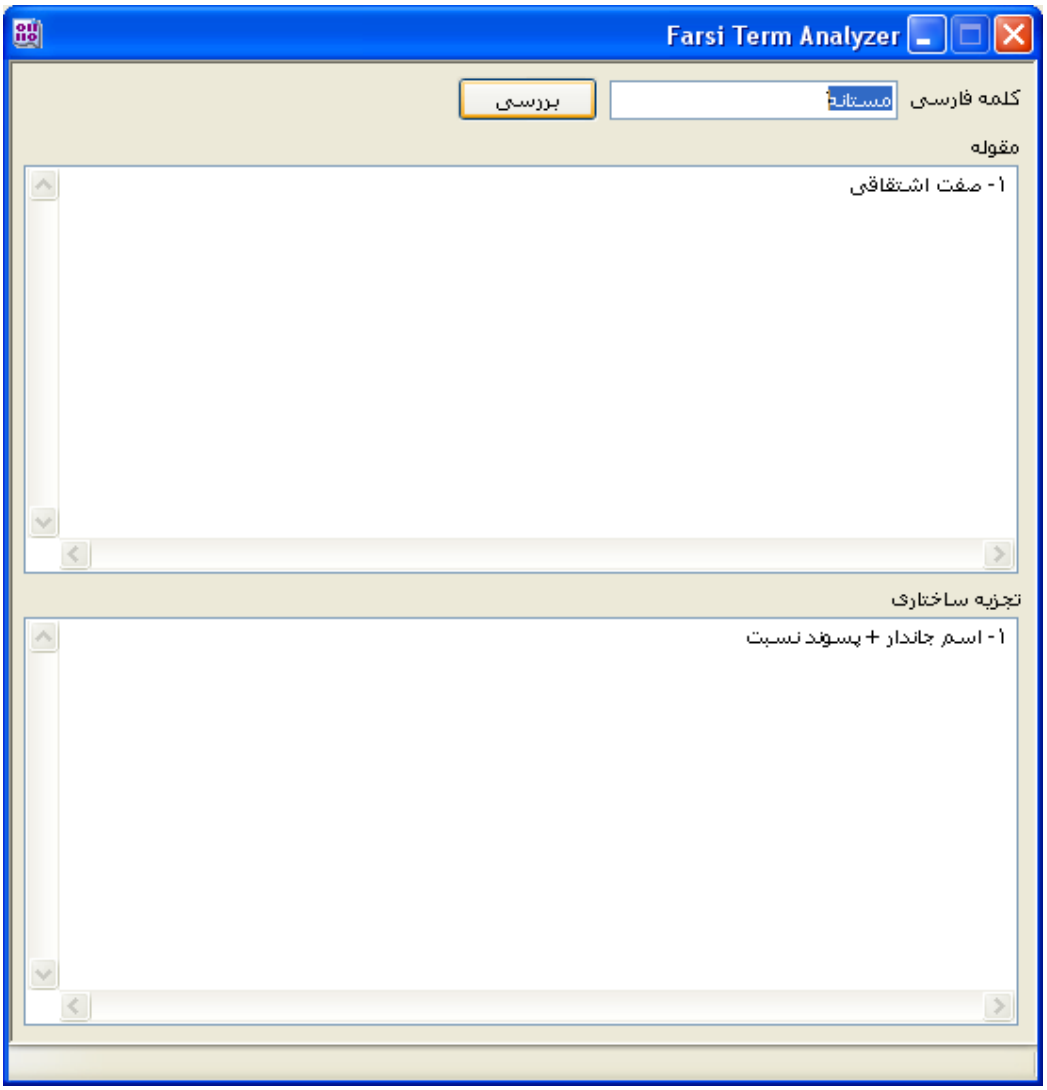

Farsi Term Analyzer
کلمه فارسی
گلستان
بررسی
مقوله
۱- اسم جاندار بن ساختی
۲- اسم مکان اشتقاقی
تجزیه ساختاری
۱- اسم + بن مضارع
۲- اسم + پسوند مکان ۱


Farsi Term Analyzer
کلمه فارسی
آب غوره دار
بررسی
مقوله
۱- اسم مرکب بن ساختی
۲- اسم مرکب + اسم *
تجزیه ساختاری
۱- اسم مرکب + بن مضارع
۲- اسم مرکب + اسم *


Farsi Term Analyzer
کلمه فارسی
نه نشست
بررسی
مقوله
۱- اسم بن ساختی
تجزیه ساختاری
۱- اسم + بن ماضی


Farsi Term Analyzer
کلمه فارسی
خواب
بررسی
مقوله
۱- بن مضارع
۲- اسم عمل
تجزیه ساختاری
۱- بن مضارع
۲- اسم عمل


Farsi Term Analyzer
کلمه فارسی
خواب نما
بررسی
مقوله
۱- اسم جاندار بن ساختی
تجزیه ساختاری
۱- اسم عمل + بن مضارع



یادآوری: واژه شکن هر کلمه پیچیده را در دو خانه تحلیل می کند (خانه بالا و خانه پایین) در ردیف اول خانه بالا مقوله دستوری و در ردیف اول خانه پایین شکل تجزیه شده آن را می دهد. واژه شکن البته در اغلب داده ها به این اکتفا نمی کند و تحلیل آن را ارائه می دهد. در این حالت فقط ردیف اول خانه بالا و ردیف اول خانه پایین معتبرند و باقی

تحلیل ها مورد نظر نیستند. گاهی کلمه ای دومقوله دستوری دارد که هر دو معتبرند بدیهی است در چنین حالتی اطلاعات هر دو ردیف مربوط به کلمه خانه بالا و پایین معتبرند . البته تحلیل های اضافی واژه شکن نیز بی فایده نیستند و قابلیت نرم افزار را ثابت می کند. تحلیل های اضافی با علامت ستاره مشخص شده اند.

برای هر کلمه مقوله دستوری و صورت تجزیه شده آن می آید. مثلا خانه‌نشین اسم جاندار بن ساختی است یعنی از اسم و بن مضارع ساخته شده است که در بالا مقوله دستوری (اسم جاندار بن ساختی) و در پایین شکل تجزیه شده (اسم + بن مضارع) آمده است.

واژه شکن کلمه دسترنج را اسم جاندار بن ساختی می داند زیرا این ترکیب شامل اسم (دست) و بن مضارع (رنج)است. یعنی طبق قاعده و به قیاس با زورگیر و خانه نشین می توان دسترنج را اسم جاندار دانست. واژه شکن در مرحله بعد دسترنج را شامل قاعده اسم + اسم معنی می داند و آن را اسم مرکب می نامد.

صورت هایی که هم اسم و هم بن فعل اند با هر دو ویژگی در فهرست قواعد آمده اند ولی چون قاعده بن مضارع + بن مضارع معتبر نیست واژه شکن خواب نما را شامل اسم عمل + بن مضارع می داند و آن را اسم جاندار می گیرد.

خلاصه

واژه‌شکن کلمات پیچیده فارسی را تجزیه و مقوله آن را مشخص می کند. واژه‌شکن ساخت های بالقوه خوش‌ساخت را دستوری می داند و مقوله آن را به دست می‌دهد هرچند که ممکن است این ساخت ها از لحاظ معنایی غیر متعارف باشد. واژه شکن بدین ترتیب ساخت هایی نظیر چوب نشین و معلمستان را دستوری می داند و مقوله دستوری آن را مشخص و آن را به اجزاء سازنده اش تجزیه می کند.

قواعد دوری در واژه شکن عمل نمی کند. هر قاعده فقط یک بار عمل می کند و تکرار نمی شود. از ترکیب کتاب + پسوند تصغیر لفظی نظیر کتابک و از ترکیب اسم جاندار+پسوند مبالغه لفظ کبوترباز ساخته می شود ولی کتابک که خود اسم و کبوترباز که خود اسم جاندار است بار دیگر با پسوند تصغیر "- ک" و پسوند مبالغه "- باز" ترکیب نمی شوند تا ترکیبی دوری ساخته نشود.

۲-۳-۳-۱-۲- فعل شکن

فعل شکن برنامه‌ای رایانه ای است، که می‌تواند فعل های ساده را تشخیص داده و تجزیه کند. در این طرح فعل‌های مرکب و پیشوندی بررسی نشده‌اند. زیرا باحذف همکرد در فعل‌های مرکب و عنصرغیرفعلی در فعل‌های پیشوندی، آن‌چه باقی می‌ماند در واقع یک فعل ساده است. چگونگی بررسی گروه فعلی در بخش نحو قابل بررسی است.

قواعد ساخت صیغگان افعال

قاعده ۱، قاعده ساخت صفت مفعولی است که در ساخت ماضی نقلی و ماضی بعید و ماضی التزامی به کار می‌رود.

۱	بن ماضی + ه = صفت مفعولی

قواعد شماره ۲ تا ۸ و نیز ۱۱ قواعد کلی ساخت زمان‌های پرکاربرد زبان فارسی یعنی حال ساده، حال التزامی، ماضی نقلی، گذشته ساده، گذشته استمراری، ماضی بعید، ماضی التزامی و آینده را نشان می‌دهد. همان‌طور که ملاحظه می‌شود این قواعد کلی‌اند و مثلا در قاعده ۲ : *می+ بن مضارع+ شناسه مضارع*[1] *= حال ساده* با جایگزینی هر یک از شناسه-ها این صیغه برای تمامی شخص‌ها(اول، دوم، سوم) و شمارها(مفرد، جمع) صرف می-شود.

از آن‌جایی که این قواعد برای کاربرد در نرم‌افزار نوشته شده است، نیازی به تعریف چگونگی اشتقاق بن‌ماضی و مضارع از یکدیگر نداریم زیرا فهرست این بن‌ها در پایگاه داده‌های نرم افزار وجود دارد که هریک از آن‌ها به صورت مستقل در فرآیند ساخت فعل

۱ نوشتن قواعد کلی با رویکرد کاربرد در ترجمه ماشینی باعث شد به سه دسته شناسه قائل شویم: شناسه‌های مضارع، شناسه‌های ماضی و شناسه‌های ماضی نقلی.

شرکت می‌کنند. نرم‌افزار مفروض قرار است قادر به تجزیه ساختار فعل باشد. که در این صورت هم نیازی به تعریف چگونگی اشتقاق بن‌ها از یکدیگر نیست.

۲	می+ بن مضارع+ شناسه مضارع = حال ساده
۳	ب + بن مضارع + شناسه مضارع = حال التزامی
۴	صفت مفعولی + شناسه نقلی = ماضی نقلی
۵	بن ماضی+ شناسه ماضی= گذشته ساده
۶	می + بن ماضی + شناسه ماضی = گذشته استمراری
۷	صفت مفعولی + بود+ شناسه ماضی = ماضی بعید
۸	صفت مفعولی + باش+ شناسه مضارع = ماضی التزامی

صیغه امر تنها برای دوم شخص مفرد صرف می‌شود که به صورت قاعده ۹ نوشته‌شده است. ۱۰ در واقع قاعده ساخت امر منفی یا همان نهی است .

۹	ب+ بن مضارع = امر دوم شخص مفرد
۱۰	ن+ بن مضارع = نهی دوم شخص مفرد
۱۱	خواه + شناسه مضارع + بن ماضی = آینده

همه صیغه‌ها می‌توانند صورت منفی داشته باشند. در یک قاعده کلی می‌توان گفت با اضافه کردن پیشوند منفی‌ساز"ن" در ابتدای هر صیغه صورت منفی آن به دست می‌آید

(قاعده ۱۲). قواعد ۱۲-۱ تا ۱۲-۷ صورت تفصیلی قاعده ۱۲ اند. تنها ساخت حال التزامی منفی از این قاعده مستثنی است که به صورت قپاعده ۱۳ نوشته شد.

۱۲	۱- ن+ فعل = فعل منفی
۱-۱۲	ن + حال ساده = حال ساده منفی
۲-۱۲	ن +صفت مفعولی + شناسه نقلی = ماضی نقلی منفی
۳-۱۲	ن + بن ماضی+ شناسه ماضی= گذشته ساده منفی
۴-۱۲	ن + می + بن ماضی + شناسه ماضی = گذشته استمراری منفی
۵-۱۲	ن + صفت مفعولی + بود+ شناسه ماضی = ماضی بعید منفی
۶-۱۲	ن + صفت مفعولی + باش+ شناسه مضارع = ماضی التزامی منفی
۷-۱۲	ن + خواه + شناسه مضارع + بن ماضی = آینده منفی
۱۳	۲- ن + بن مضارع + شناسه مضارع = حال التزامی منفی

برای اینکه بتوان قواعد را در برنامه نویسی کامپیوتری به‌کار برد، لازم است که با یک زبان میانجی آن‌ها را کدگذاری و سپس قواعد را با استفاده از آن سیستم کدگذاری بازنویسی کرد. در جدول‌های زیر این مراحل انجام شده است.

بن ماضی	S 1

بن مضارع	S 2

x در H_x، G_x و N_x متغیری است که می تواند از ۱ تا ۶ متغیر باشد. و به ترتیبی که در جدول دیده می‌شود بر شخص و شمارهای مختلف دلالت کند. پس وقتی در قاعده‌ای می‌نویسیم H_x یعنی هر یک از صورت‌های آن (از H_1 تا H_6) می‌تواند به کار رود.

شناسه های مضارع H_x					
م	اول شخص مفرد	H_1	یم	اول شخص جمع	H_4
ی	دوم شخص مفرد	H_2	ید	دوم شخص جمع	H_5
د	سوم شخص مفرد	H_3	ند	سوم شخص جمع	H_6

شناسه های ماضی G_X					
م	اول شخص مفرد	G_1	یم	اول شخص جمع	G_4
ی	دوم شخص مفرد	G_2	ید	دوم شخص جمع	G_5
Ø	سوم شخص مفرد	G_3	ند	سوم شخص جمع	G_6

شناسه های ماضی نقلی N_X					
ام	اول شخص مفرد	N_1	ایم	اول شخص جمع	N_4
ای	دوم شخص مفرد	N_2	اید	دوم شخص جمع	N_5
است	سوم شخص مفرد	N_3	اند	سوم شخص جمع	N_6

متغیر X در شکل نهایی فعل همان شخص و شمار را نشان می‌دهد و H_X، G_X و N_X تعیین کننده آن هستند. به عبارت دیگر X در $V_{Y\,X}$، با X در H_X، G_X یا N_X سازنده آن برابر است. برای مثال در قاعده $P1 + S2 + H_X = V_{A\,X}$ اگر $H_X = H_1$ باشد (یعنی شناسه اول شخص مفرد) آن‌گاه $V_{A\,X} = V_{A\,1}$ (یعنی فعل حال ساده اول شخص مفرد).

متغیر Y در $V_{Y\,X}$، بر زمان فعل دلالت دارد. همان‌طور که در جدول زیر پیداست Y می‌تواند A باشد و بر حال ساده دلالت کند، B باشد و بر ماضی نقلی دلالت کند، C باشد و بر گذشته ساده دلالت کند، D باشد و بر گذشته استمراری دلالت کند، E باشد و بر ماضی بعید دلالت کند، F باشد و بر ماضی التزامی دلالت کند، G باشد و بر آینده دلالت کند. در حال التزامی ($V_{K\,X}$)،امر($V_{L\,X}$) و نهی هم متغیر اول در واقع همان Y است، اما در جدولی جداگانه قرار گرفته‌اند زیرا ساخت منفی آن‌ها با بقیه زمان‌های ذکر شده تفاوت دارد (قاعده حال التزامی منفی *ن + بن مضارع + شناسه مضارع = حال التزامی*

منفی است و نهی صورت نفی امر است که تنها برای دوم شخص مفرد صرف مـی شـود و قاعده آن *ن+ بن مضارع = نهی دوم شخص مفرد* است) و قاعده کلی *ن+ فعل = فعل منفی* شامل آن نمی‌شود.

زمان V Y X			
حال ساده	V AX	ماضی بعید	V E X
ماضی نقلی	V B X	ماضی التزامی	V F X
گذشته ساده	V C X	آینده	V G X
گذشته استمراری	V D X		

حال التزامی	V K X
امر	V L X
نهی	V J X

می	P1
ب	P2
ن	P3
ه	P4

با در نظر داشتن آن‌چه پیش‌تر توضیح داده شد، فعل منفی لفظی است که بر کل صورت-های منفی دلالت می‌کند. اگرچه حال التزامی منفی هم نوعی فعل منفی است، چون قاعده

ساخت آن با سایر ساخت‌های منفی متفاوت است لازم است با لفظی متفاوت نـام گـذاری شود.

فعل منفی	$VN_{Y\ X}$
حال التزامی منفی	$VN_{K\ x}$

صفت مفعولی	A

بود	F1
باش	F2
خواه	F3

در جدول‌های زیر قواعد و صورت کدنویسی شده آن‌ها نشان داده شده‌است.

۱	بن ماضی + ه = صفت مفعولی	S1 + P4 = A

۲	می+ بن مضارع+ شناسه مضارع = حال ساده	$P1 + S2 + H_X = V_{A\ X}$
۳	ب + بن مضارع + شناسه مضارع = حال التزامی	$P2 + S2 + H\ x = V_{K\ X}$
۳-۱ استثناء: ب+ ی+ بن مضارعی که با آ آغاز می شود+ شناسه مضارع = فعل حال التزامی		
۴	صفت مفعولی + شناسه نقلی = ماضی نقلی	$A + N_X = V_{B\ X}$
۵	بن ماضی+ شناسه ماضی= گذشته ساده	$S1 + G_X = V_{C\ X}$

۶	می + بن ماضی + شناسه ماضی = گذشته استمراری	$P1 + S1+ G_X = V_{D X}$
۷	صفت مفعولی + بود+ شناسه ماضی = ماضی بعید	$A + F1 + G_X = V_{E X}$
۸	صفت مفعولی + باش+ شناسه مضارع = ماضی التزامی	$A + F2 + H_X = V_{F X}$

۹	ب+ بن مضارع = امر دوم شخص مفرد	$P2 + S2 = V_{L 2}$
۱۰	ن+ بن مضارع = نهی دوم شخص مفرد	$P3+ S2 = V_{J 2}$
۱۱	خواه + شناسه مضارع + بن ماضی = آینده	$F3 + H_X + S1 = V_{G X}$

۱۲	ن+ فعل = فعل منفی	$P3 + Vy\ x= VN_{Y x}$
۱-۱۲	*ن + حال ساده = حال ساده منفی*	$P3 + Vax = VN_{AX}$
۲-۱۲	*ن +صفت مفعولی + شناسه نقلی = ماضی نقلی منفی*	$P3 + A + Nx = VN_{BX}$
۳-۱۲	*ن + بن ماضی+ شناسه ماضی= گذشته ساده منفی*	$P3 + S1 + Gx = VN_{CX}$
۴-۱۲	*ن + می + بن ماضی + شناسه ماضی = گذشته استمراری منفی*	$P3 + P1 + S1 = VN_{DX}$
۵-۱۲	*ن + صفت مفعولی + بود+ شناسه ماضی = ماضی بعید منفی*	$P3 + A + F1 + Gx = VN_{EX}$
۶-۱۲	*ن + صفت مفعولی + باش+ شناسه مضارع = ماضی التزامی منفی*	$P3 + A + F2 + Hx = VN_{FX}$

۷-۱۲	*ن + خواه + شناسه مضارع + بن ماضی = آینده منفی*	$P3 + F3 + Hx + S1 = VN_{GX}$
۱۳	ن + بن مضارع + شناسه مضارع = حال التزامی منفی	$P3 + S2 + Hx = VN_{KX}$

۲-۳-۳-۲-نحو

قواعد صوری دستور زبان مقصد یکی از ملزومات ترجمه ماشینی است. در ترجمه باید برای تمامی حوزه نحو زبان مقصد قواعدی صوری استخراج کرد. این قواعد که بسیاری از آن‌ها مستقل‌اند و می توان آن‌ها را بدون ارتباط با یکدیگر مشخص کرد، در نهایت به هم می‌پیوندند وقاعده‌ای منسجم برای نحو زبان مقصد می سازند.

سرفصل مواردی که باید در حوزه نحو مورد مطالعه دقیق قراربگیرند عبارت‌اند از:

- انواع جمله
- ساختار درونی جمله
- گروه صفتی
- گروه قیدی
- گروه حرف اضافه‌ای
- گروه اسمی

- ضمایر
- همپایگی
- پیروسازی

هر یک از سرفصل‌های بالا که موضوع کلی یک طرح پژوهشی است، به قسمت‌های مختلفی تقسیم می‌شود، برای مثال:

- انواع جمله: در این قسمت باید به جملات خبری، پرسشی، امری و نیز نقل‌قول‌ها پرداخته شود. هریک از این عنوان‌ها نیز، خود دارای عناوین فرعی است. برای مثال در مورد جملات پرسشی، انواع جمله پرسشی، عناصر پرسشی‌ساز و ...
- ساختار درونی جمله: جمله ربطی، فعلی و انواع قیدگونه‌ها

تمام موضوعات معرفی‌شده باید از دیدگاه کاربرد در ترجمه ماشینی مورد مطالعه قرار بگیرد. یعنی خروجی هر طرح باید داده‌های مورد نیاز تجزیه‌گرنحوی نرم‌افزار ترجمه را تامین کند. دقیقا مانند واژه‌شکن و فعل‌شکن که خروجی آن‌ها برای کاربرد در تحلیل-گر صرفی نرم‌افزار ترجمه طراحی شده است. در ادامه برای نشان دادن چگونگی مطالعه مباحث نام‌برده، به معرفی پژوهشی درباره وابسته‌های پیشین گروه اسمی، که توسط نگارنده انجام شده است، می‌پردازیم.

۲-۳-۳-۲-۱- وابسته‌های پیشین گروه اسمی

گروه اسمی یکی از مهم ترین سازه های نحوی زبان فارسی و اغلب زبان های دنیا است که نقش مهمی در قواعد نحوی دارد.هر گروه اسمی یک هسته و یک مجموعه وابسته دارد. هسته گروه اسمی معمولا اسم یا ضمیر است. وجود هسته اجباری و وابسته‌ها اختیاری است. اگر گروه اسمی فقط از یک کلمه ساخته شده باشد همان کلمه هسته گروه اسمی خواهد بود.

وابسته های پیش ازهسته را وابسته های پیشین، و وابسته های پس از هسته را وابسته های پسین می نامند. وابسته‌های پیشین گروه اسمی به طبقات مشخصی نظیر صفت‌های اشاره و اعداد و صفت مبهم و صفت عالی و صفت های پرسشی و تعجبی و شاخص‌ها تقسیم می‌شوند. این طبقات هریک دارای اعضایی هستند. اعضای طبقات مذکور با قواعدی روی محور هم نشینی در کنار هم قرار می‌گیرند و قاعده‌های باهم آیی خاصی دارند. یعنی هم از لحاظ تقدم و تاخر و هم از نظر نوع وابسته‌ها تابع قاعده هستند. از طرفی برخی از این عناصر مانعه الجمع‌اند.

وابسته‌های پیشین زبان فارسی، در منابع مختلف به طور اجمالی، به صورت زیر ارائه شـده-اند:

صادقی و ارژنگ (۱۳۵۶)	مشکوة الدینی (۱۳۷۳)	حق شناس و دیگران (۱۳۸۵)	سمائی (۱۳۸۱)
شاخص ها	وابسته وصفی	لقب ها (شاخص ها)	شاخص
اعداداصلی *(و ممیز)*	وابسته عدد و ممیز	عددهای اصلی *(و ممیز)*	عدد *(و ممیز) (شامل اصلی و ترتیبی)*
عددهای ترتیبی و صفت های عالی	وابسته عدد ترتیبی	عددهای ترتیبی و صفت های عالی	
	وابسته صفت برترین		صفت عالی
صفت های اشاره	وابسته های اشاره	صفت های اشاره	صفت اشاره
صفت های پرسشی	وابسته های پرسشی	صفت های پرسشی	صفت پرسشی
صفت های تعجبی	وابسته های تعجبی	صفت های تعجبی	صفت تعجبی
صفت های مبهم	وابسته نامشخص	صفت های مبهم	صفت مبهم
یک نکره		یک نشانه نکره	یک نکره
		صفت ها بیانی	

در طرحی که به بررسی وابسـته‌هـای پیشـین در زبـان فارسـی، بـا نگـاه کـاربرد در ترجمـه ماشینی می‌پردازد، این وابسته‌ها به ترتیب جدول[1] زیر دسته‌بندی شده‌اند:

[1] در رابطه با جدول شماره ۱ توجه به چند نکته ضروري است:
۱- صفت‌هاي تعجبي بيشتر در سبك محاوره كاربرد دارد و در اين طرح برررسي نشده‌اند.
۲- بسا، بسي، فلان و بعضي ديگر از اعضاي طبقات به دلايل سبكي حذف شده‌اند.
۳- " يك نكره" به دلايل سبكي و نيز اين كه بايد با وابسته پسين "ي" بررسي شود، حذف شده‌است.
٤- شاخص به دليل هم مقوله بودن با اسامي و لزوم اختصاص مولفه هاي معنايي انتزاعي به آن‌ها از فهرست قواعد حذف شده اند.
٥- دو نوع عدد ترتيبي وجود دارد كه در طرح واژه شكن و عددشكن از آن ها به عنوان عددترتيبي نوع اول و عدد ترتيبي نوع دوم نام برده شده است. قاعده اعداد ترتيبي نوع اول و دوم در واژه شكن به ترتيب به صورت عدد *اصلي + وند ترتيب۱ (پنجم)* و عدد *اصلي + وند ترتيب* ۲ (پنجمين) است. عدد ترتيبي نوع اول وابسته پسين و عدد ترتيبي نوع دوم وابسته پيشين است. در اين طرح منظور ازعدد ترتيبي، عدد ترتيبي نوع دوم است.

جدول شماره ۱:

صفت اشاره		صفت مبهم		صفت پرسشی		اعداد	ممیز	صفت عالی
بسیط	این	هر	بسیط	چند	بسیط	اعداد اصلی	عدد	صفت + ترین
	آن	هیچ		چندمین		اعداد ترتیبی		یگانه
	همین	همه		چه				تنها
	همان	بعضی		چگونه				
مرکب	چنین	برخی		چه نوع				
	چنان	چند		چقدر				
	یک چنین	چندین		کدام				
	اینهمه	بعضی از	مرکب	کدامیک از				
		برخی از	مرکب					
		هریک از						
		هرکدام از						
		هیچیک از						
		هیچکدام از						

وابسته های پیشین برای قرارگرفتن در کنار هم محدودیت هایی دارنـد. یعنـی هـر وابسـته ای ممکن است قبل و یا بعد از بعضی وابسته ها بیاید و در کنار بعضی دیگر قرار نگیرد. هم‌نشینی وابسته های پیشین ابتدا به صورت دو به دو بررسی شد که حاصل آن در جدول شماره ۲ ارائه شده است:

جدول شماره۲:

صفت اشاره + صفت مبهم*	صفت اشاره + صفت پرسشی*	صفت اشاره بسیط + اعداد اصلی
*(صفت اشاره بسیط + صفت مبهم بسیط *)*	*(صفت اشاره بسیط + صفت پرسشی بسیط *)*	صفت اشاره مرکب + اعداد اصلی *
*(صفت اشاره بسیط + صفت مبهم مرکب *)*	*(صفت اشاره بسیط + صفت پرسشی مرکب *)*	
*(صفت اشاره مرکب+ صفت مبهم بسیط *)*	*(صفت اشاره مرکب+ صفت پرسشی بسیط *)*	
*(صفت اشاره مرکب+ صفت مبهم مرکب *)*	*(صفت اشاره مرکب + صفت پرسشی مرکب *)*	
صفت اشاره بسیط + همه/ چند	صفت پرسشی بسیط + صفت اشاره*	صفت اشاره + اعداد ترتیبی نوع اول*
	*(صفت پرسشی بسیط + صفت اشاره بسیط *)*	*(صفت اشاره بسیط + اعداد ترتیبی نوع اول *)*
	(صفت پرسشی بسیط + صفت اشاره مرکب)*	*(صفت اشاره مرکب + اعداد ترتیبی نوع اول *)*
صفت مبهم بسیط + صفت اشاره*	صفت پرسشی مرکب + صفت اشاره بسیط	اعداد + صفت اشاره*
(صفت مبهم بسیط + صفت اشاره بسیط)*		*(اعداد اصلی + صفت اشاره بسیط *)*
(صفت مبهم بسیط + صفت اشاره مرکب)*		*(اعداد اصلی + صفت اشاره مرکب *)*
		*(اعداد ترتیبی نوع اول + صفت اشاره بسیط *)*
		(اعداد ترتیبی نوع اول + صفت اشاره

		مرکب*)
	صفت پرسشی مرکب + صفت اشاره مرکب*	
صفت مبهم مرکب + صفت اشاره بسیط		
صفت مبهم مرکب + صفت اشاره مرکب*		
صفت اشاره + صفت مبهم* (صفت اشاره بسیط + صفت مبهم بسیط*) (صفت اشاره بسیط + صفت مبهم مرکب*) (صفت اشاره مرکب+ صفت مبهم بسیط*) (صفت اشاره مرکب+ صفت مبهم مرکب*)	صفت اشاره + صفت پرسشی* (صفت اشاره بسیط + صفت پرسشی بسیط*) (صفت اشاره بسیط + صفت پرسشی مرکب*) (صفت اشاره مرکب+ صفت پرسشی بسیط*) (صفت اشاره مرکب + صفت پرسشی مرکب*)	صفت اشاره بسیط + اعداد اصلی
		صفت اشاره مرکب + اعداد اصلی *
صفت اشاره بسیط + همه/ چند	صفت پرسشی بسیط + صفت اشاره* (صفت پرسشی بسیط + صفت اشاره بسیط*) (صفت پرسشی بسیط + صفت اشاره مرکب*)	صفت اشاره + اعداد ترتیبی نوع اول* (صفت اشاره بسیط + اعداد ترتیبی نوع اول *) (صفت اشاره مرکب + اعداد ترتیبی نوع اول *)
صفت مبهم بسیط + صفت اشاره* (صفت مبهم بسیط + صفت اشاره بسیط*)	صفت پرسشی مرکب + صفت اشاره بسیط	اعداد + صفت اشاره* (اعداد اصلی + صفت اشاره بسیط*)

(صفت مبهم بسیط + صفت اشاره مرکب*)		(اعداد اصلی + صفت اشاره مرکب*) (اعداد ترتیبی نوع اول + صفت اشاره بسیط*) (اعداد ترتیبی نوع اول + صفت اشاره مرکب*)
	صفت پرسشی مرکب + صفت اشاره مرکب*	
صفت مبهم مرکب + صفت اشاره بسیط		
صفت مبهم مرکب + صفت اشاره مرکب*		
صفت اشاره + صفت مبهم* (صفت اشاره بسیط + صفت مبهم بسیط*) (صفت اشاره بسیط + صفت مبهم مرکب*) (صفت اشاره مرکب+ صفت مبهم بسیط*) (صفت اشاره مرکب+ صفت مبهم مرکب*)	صفت اشاره + صفت پرسشی* (صفت اشاره بسیط + صفت پرسشی بسیط*) (صفت اشاره بسیط + صفت پرسشی مرکب*) (صفت اشاره مرکب+ صفت پرسشی بسیط*) (صفت اشاره مرکب + صفت پرسشی مرکب*)	صفت اشاره بسیط + اعداد اصلی
		صفت اشاره مرکب + اعداد اصلی *
صفت اشاره بسیط + همه/ چند	صفت پرسشی بسیط + صفت اشاره* (صفت پرسشی بسیط + صفت اشاره بسیط*) (صفت پرسشی بسیط + صفت اشاره مرکب*)	صفت اشاره + اعداد ترتیبی نوع اول* (صفت اشاره بسیط + اعداد ترتیبی نوع اول *)

		(صفت اشاره مرکب + اعداد ترتیبی نوع اول *)
صفت مبهم بسیط + صفت اشاره* (صفت مبهم بسیط + صفت اشاره بسیط*) (صفت مبهم بسیط + صفت اشاره مرکب*)	صفت پرسشی مرکب + صفت اشاره بسیط	اعداد + صفت اشاره* (اعداد اصلی + صفت اشاره بسیط*) (اعداد اصلی + صفت اشاره مرکب*) (اعداد ترتیبی نوع اول + صفت اشاره بسیط*) (اعداد ترتیبی نوع اول + صفت اشاره مرکب*)
	صفت پرسشی مرکب + صفت اشاره مرکب*	
صفت مبهم مرکب + صفت اشاره بسیط		
صفت مبهم مرکب + صفت اشاره مرکب*		
صفت اشاره بسیط + صفت عالی	صفت مبهم + صفت پرسشی* (صفت مبهم بسیط + صفت پرسشی بسیط*) (صفت مبهم بسیط + صفت پرسشی مرکب*) (صفت مبهم مرکب+ صفت پرسشی بسیط*) (صفت مبهم مرکب + صفت پرسشی مرکب*)	صفت پرسشی بسیط + اعداد* (صفت پرسشی بسیط + اعداد اصلی*) (صفت پرسشی بسیط + اعداد ترتیبی*) کدام +اعداد

		(کدام + اعداد اصلی) (کدام + اعداد ترتیبی)
صفت اشاره مرکب + صفت عالی*	صفت پرسشی + صفت مبهم* (صفت پرسشی بسیط + صفت مبهم بسیط*) (صفت پرسشی بسیط + صفت مبهم مرکب*) (صفت پرسشی مرکب+ صفت مبهم بسیط*) (صفت پرسشی مرکب + صفت مبهم مرکب*)	صفت پرسشی مرکب + اعداد اصلی
		صفت پرسشی مرکب + اعداد ترتیبی*
صفت عالی + صفت اشاره* (صفت عالی + صفت اشاره بسیط *) (صفت عالی + صفت اشاره مرکب*)	صفت مبهم + اعداد* (صفت مبهم بسیط + اعداد اصلی*) (صفت مبهم بسیط + اعداد ترتیبی نوع اول*) (صفت مبهم مرکب+ اعداد اصلی*) (صفت مبهم مرکب + اعداد ترتیبی نوع اول*)	
	صفت مبهم + صفت عالی* (صفت مبهم بسیط +صفت عالی*) (صفت مبهم مرکب+ صفت عالی*)	صفت پرسشی بسیط +صفت عالی* صفت پرسشی مرکب + صفت عالی صفت پرسشی مرکب + یگانه/ تنها *

		کدام + صفت عالی چند/ چندمین + ممیز
		اعداد + صفت پرسشی* (اعداد اصلی + صفت پرسشی بسیط*) (اعداد ترتیبی نوع اول + صفت پرسشی مرکب*) (اعداد اصلی + صفت پرسشی بسیط*) (اعداد ترتیبی نوع اول + صفت پرسشی مرکب*)
		صفت عالی+ صفت پرسشی* (صفت عالی + صفت پرسشی بسیط*) (صفت عالی + صفت پرسشی مرکب*)
صفت اشاره بسیط + صفت عالی	صفت مبهم + صفت پرسشی* (صفت مبهم بسیط + صفت پرسشی بسیط*) (صفت مبهم بسیط + صفت پرسشی مرکب*) (صفت مبهم مرکب+ صفت پرسشی بسیط*) (صفت مبهم مرکب + صفت پرسشی مرکب*)	صفت پرسشی بسیط + اعداد* (صفت پرسشی بسیط + اعداد اصلی*) (صفت پرسشی بسیط + اعداد ترتیبی*) کدام +اعداد

		(کدام + اعداد اصلی) (کدام + اعداد ترتیبی)
صفت اشاره مرکب + صفت عالی*	صفت پرسشی + صفت مبهم* (صفت پرسشی بسیط + صفت مبهم بسیط*) (صفت پرسشی بسیط + صفت مبهم مرکب*) (صفت پرسشی مرکب+ صفت مبهم بسیط*) (صفت پرسشی مرکب + صفت مبهم مرکب*)	صفت پرسشی مرکب + اعداد اصلی
		صفت پرسشی مرکب + اعداد ترتیبی*
صفت عالی + صفت اشاره* (صفت عالی + صفت اشاره بسیط *) (صفت عالی + صفت اشاره مرکب*)	صفت مبهم + اعداد* (صفت مبهم بسیط + اعداد اصلی*) (صفت مبهم بسیط + اعداد ترتیبی نوع اول*) (صفت مبهم مرکب+ اعداد اصلی*) (صفت مبهم مرکب + اعداد ترتیبی نوع اول*)	
	صفت مبهم + صفت عالی* (صفت مبهم بسیط +صفت عالی*) (صفت مبهم مرکب+ صفت عالی*)	صفت پرسشی بسیط +صفت عالی* صفت پرسشی مرکب + صفت عالی صفت پرسشی مرکب + یگانه/ تنها *

		کدام + صفت عالی چند/ چندمین + ممیز
		اعداد + صفت پرسشی* (اعداد اصلی + صفت پرسشی بسیط*) (اعداد ترتیبی نوع اول + صفت پرسشی مرکب*) (اعداد اصلی + صفت پرسشی بسیط*) (اعداد ترتیبی نوع اول + صفت پرسشی مرکب*)
		صفت عالی+ صفت پرسشی* (صفت عالی + صفت پرسشی بسیط*) (صفت عالی + صفت پرسشی مرکب*)

سپس امکان هم نشینی همه وابسته های پیشین با یکدیگر مورد بررسی قرار گرفت .

جدول شماره ۳ شامل قواعد همه هم‌نشینی‌های مجاز است:

جدول شماره ۳:

صفت پرسشی مرکب + صفت اشاره بسیط	صفت پرسشی مرکب + صفت اشاره بسیط + همه/چند صفت پرسشی مرکب + صفت اشاره بسیط + اعداد اصلی صفت پرسشی مرکب + صفت اشاره بسیط + اعداد اصلی + ممیز صفت پرسشی مرکب + صفت اشاره بسیط + صفت عالی
صفت اشاره بسیط + اعداد اصلی	
صفت اشاره بسیط + همه/ چند	
صفت اشاره بسیط + صفت عالی	
صفت پرسشی مرکب + اعداد اصلی	صفت پرسشی مرکب + اعداد اصلی+ ممیز
صفت پرسشی مرکب + صفت عالی	
صفت مبهم مرکب + صفت اشاره بسیط	
اعداد + ممیز (اعداد اصلی+ ممیز) (اعداد ترتیبی + ممیز)	کدام +اعداد+ممیز (کدام + اعداد اصلی+ ممیز) (کدام + اعداد ترتیبی+ ممیز)
کدام +اعداد (کدام + اعداد اصلی) (کدام + اعداد ترتیبی)	
کدام + صفت عالی	
چند/ چندمین + ممیز	

برای استفاده این قواعد در نرم افـزار نخسـت بایـد وابسـته هـای پیشـین کـد گـذاری شـوند. (جدول شماره ۵):

جدول شماره۵:

صفت اشاره	A10
صفت اشاره بسیط	A10a
صفت اشاره مرکب	A10b
صفت مبهم	A11
صفت مبهم بسیط	A11a
صفت مبهم مرکب	A11b
همه/چند	A11a1
صفت پرسشی	A12
صفت پرسشی بسیط	A12a
صفت پرسشی مرکب	A12b
چند/چندمین	A12a1
کدام	A12a2
صفت عالی	A13
صفت + ترین	A01S39
اعداد	M
اعداد اصلی	M01
اعداد ترتیبی	A08 (M01S25)
ممیز	Q

در رابطه با جدول شماره ۵ به نکات زیر توجه کنید:

۱- صفت عالی با همان کدگذاری مورد استفاده در واژه شکن کد گـذاری شـده اسـت A13. قاعده ساخت صفت عالی در واژه شکن A01S39 اسـت. A01 "صـفت" و S39 پسوند "- ترین" است.

۲- اعداد هم با کدگذاری مورد استفاده در عدد شکن[۱] در این جا کدگذاری شده‌اند.

[۱] عددشکن فارسی نرم‌افزاری است که قادر است گروه اعداد را پردازش کند و قواعد آن را معین کند و صورت دستوری و غیردستوری آن را هم تمییز دهد. این نرم‌افزار نیز یکی از طرح‌های پژوهشی نگارنده است.

با کمک این زبان میانجی، قواعد برای استفاده در نرم افزار بازنویسی شد:

جدول شماره ۶:

صفت پرسشی مرکب + صفت اشاره بسیط A12b+ A10a	صفت پرسشی مرکب + صفت اشاره بسیط + همه/چند A12b+ A10a+A11a1 صفت پرسشی مرکب + صفت اشاره بسیط + اعداد اصلی A12b+ A10a+M01 صفت پرسشی مرکب + صفت اشاره بسیط + اعداد اصلی + ممیز A12b+ A10a+M01+Q صفت پرسشی مرکب + صفت اشاره بسیط + صفت عالی A12b+ A10a+A13
صفت اشاره بسیط + اعداد اصلی A10a+M01	
صفت اشاره بسیط + همه/ چند A10a+ A11a1	
صفت اشاره بسیط + صفت عالی A10a+A13	
صفت پرسشی مرکب + اعداد اصلی A12b+M01	صفت پرسشی مرکب + اعداد اصلی+ ممیز A12b+ M01+Q
صفت پرسشی مرکب + صفت عالی A12b+ A13	
صفت مبهم مرکب + صفت اشاره بسیط A11b+ A10a	
اعداد + ممیز M+Q (اعداد اصلی+ ممیز) (M01+Q) (اعداد ترتیبی + ممیز) (A08+Q)	کدام +اعداد+ممیز A12a 2+M+Q (کدام + اعداد اصلی+ ممیز) A12a 2+M01+Q)((کدام + اعداد ترتیبی+ ممیز) (A12a 2+M01+A08)

کدام +اعداد A12a 2+M (کدام + اعداد اصلی) (A12a 2+M01) (کدام + اعداد ترتیبی) (A12a 2+A08)	
کدام + صفت عالی A12a 2+A13	
چند/ چندمین + ممیز A12a1+Q	

جدول شماره ۷:

A12b+ A10a
A10a+M01
A10a+ A11a1
A10a+A13
A12b+M01
A12b+ A13
A11b+ A10a
M+Q (M01+Q) (A08+Q)
A12a 2+M (A12a 2+M01) (A12a 2+A08)
A12a 2+A13
A12a1+Q

A12b+ A10a+A11a1 A12b+ A10a+M01 A12b+ A10a+M01+Q A12b+ A10a+A13
A12b+ M01+Q
A12a 2+M+Q (A12a 2+M01+Q) (A12a 2+M01+A08)

جمع‌بندی

فصل اول به مرور کلی سابقه پژوهش‌های انجام‌شده در زمینه مترجم ماشینی و معرفی نرم-افزارها اختصاص داده شد . در این بخش ابتدا به تعریف زبان‌شناسی رایانه‌ای و ترجمه ماشینی به عنوان یکی از فعالیت‌های این حوزه پرداخته شد.

در این فصل تاریخ ترجمه ماشینی بعد از جنگ جهانی دوم تا امروز از نظر گذرانده شد و منابع، روش‌ها، نظریه‌ها و فناوری‌های موجود و پیشرفت‌های صورت‌گرفته در این زمینه معرفی و نقد شد.

در فصل دوم به طرح‌های مرتبط با تحقق مترجم ماشینی پرداخته شد. در این فصل به تفصیل توضیح داه شد که برای داشتن یک مترجم ماشینی، زبان مبدا و مقصد باید به طور دقیق بررسی و قواعد آن استخراج شود. متناسب با موضوع طرح حاضر("شناسایی طرح‌های مرتبط با تحقق مترجم ماشینی در زبان فارسی")، به بررسی زبان‌فارسی از دیدگاه ترجمه ماشینی پرداختیم. برای نشان‌دادن چگونگی استخراج قواعد، چند طرح که پیش‌تر توسط نگارنده انجام شده بود، معرفی شد. شایان ذکر است که با همکاری متخصصان رایانه، برای دو طرح از طرح‌های نام‌برده، نرم‌افزار تهیه شده است.

به طور کلی طرح‌های مرتبط با تحقق مترجم ماشینی در زبان فارسی عبارتند از:

- تدوین واژگان (با مشخصات تصریح‌شده در متن)

- قواعد ساخت انواع جمله پرسشی
- قواعد ساخت انواع جمله امری
- قواعد ساخت انواع جمله خبری
- قواعد ساخت جملات همپایه
- قواعد ساخت گروه اسمی
- قواعد ساخت گروه صفتی
- قواعد ساخت گروه قیدی
- قواعد ساخت گروه حرف‌اضافه‌ای
- قواعد مرجع‌گزینی
- قواعد پیروسازی با بند اسمی
- قواعد پیروسازی با بند صفتی
- قواعد پیروسازی با بند قیدی
- قواعد تطابق زمان‌ها

واژه‌نامه فارسی به انگلیسی

آواشناسی	phonology
ابهام ساختاری	lexical ambiguity
ابهام واژگانی	lexical ambiguity
ابهام‌زدایی از معنای واژه	word-sense disambiguation
اشتقاق	derivation
انتقال ساختاری	structural transfer
انتقال واژه‌ها	lexical transfer
بافت	context
بینشی	heuristic
پایانه کلمات	ending
پایه	base
پایین به بالا	bottom-up
پس ویرایش	post-editing
پس‌ویرایش آماری قواعد	rules post-processed by statistics
پیش ویرایش	pre-editing
پیکره	corpus
تحلیل ورودی	input analysis
ترتیب کلمات	word Order
ترجمه تعاملی	interactive translation
ترجمه رایانه‌ای	computer-aided translation
ترجمه ماشینی انتقال بنیان	transfer-based machine translation

hybrid machine translation	ترجمه ماشینی پیوندی
shallow-transfer machine translation	ترجمه ماشینی سطحی
dictionary-based machine translation	ترجمه ماشینی فرهنگ بنیان
example-based machine translation	ترجمه ماشینی مثال بنیان
interlingua machine translation	ترجمه ماشینی میان‌زبانی
compounding	ترکیب
inflectional	تصریفی
interactive	تعاملی
tokenization	واحدسازی
distribution	توزیع
location	جایگاه
substitution	جایگزین‌سازی
subcategorization frame	چارچوب زیرمقوله ای
polysemy	چند معنایی
linear	خطی
well-formed	خوش‌ساخت
universal encyclopedia	دائره‌المعارف جهانی
linguistics knowledge	دانش زبانی
synset	دایره معنایی
text understanding	درک متن
lexical Functional Grammar	دستور واژگانی نقشمند

recursive	دوری
controlled language	زبان مهارشده
light verb construction	ساختار فعل‌های سبک
constituent	سازه
parsing	سازه‌یابی
shallow	سطحی
motphology	صرف
citation form	صورت اسنادی
subworld	عالم فرعی
punctuation	علائم سجاوندی
deep	عمیق
extended dictionary	فرهنگ لغت گسترده
Light verb	فعل سبک
open source	متن‌باز
scenario	فیلمنامه
rule-based	قاعده بنیان
domain	قلمرو
subcategorization rules	قواعد زیرمقوله ای
analogy	قیاس
Key words	کلیدواژه ها
phrase	گروه
unrestricted	محدود نشده
boundary	مرز

semantics	معنی‌شناسی
Iterative	مکرر
formulaic	منضبط
intermediary	میانجی
syntax	نحو
orthographic	نگارشی
scripts	نمایشنامه
conceptual dependency	وابستگی مفهومی
lexicon	واژگان
lexico - syntactic	واژگانی – نحوی
gloss	واژه‌نامه
statistics guided by rules	ویرایش آمار با قواعد
homonymy	هم معنایی
homophone	هماوا

واژه‌نامه انگلیسی به فارسی

analogy	قیاس
base	پایه
bottom-up	پایین به بالا
boundary	مرز
citation form	صورت اسنادی
compounding	ترکیب
computer-aided translation	ترجمه رایانه‌ای
conceptual dependency	وابستگی مفهومی
constituent	سازه
context	بافت
controlled language	زبان مهارشده
corpus	پیکره
deep	عمیق
derivation	اشتقاق
dictionary-based machine translation	ترجمه ماشینی فرهنگ بنیان
distribution	توزیع
domain	قلمرو
ending	پایانه کلمات
example-based machine translation	ترجمه ماشینی مثال بنیان
extended dictionary	فرهنگ لغت گسترده
formulaic	منضبط

gloss	واژه‌نامه
heuristic	بینشی
homonymy	هم معنایی
homophone	هماوا
hybrid machine translation	ترجمه ماشینی پیوندی
inflectional	تصریفی
input analysis	تحلیل ورودی
interactive	تعاملی
interactive translation	ترجمه تعاملی
interlingua machine translation	ترجمه ماشینی میان‌زبانی
intermediary	میانجی
Iterative	مکرر
Key words	کلیدواژه ها
lexical ambiguity	ابهام ساختاری
lexical ambiguity	ابهام واژگانی
lexical Functional Grammar	دستور واژگانی نقشمند
lexical transfer	انتقال واژه‌ها
lexicon	واژگان
lexico - syntactic	واژگانی – نحوی
Light verb	فعل سبک
light verb construction	ساختار فعل‌های سبک
linear	خطی

linguistics knowledge	دانش زبانی
location	جایگاه
motphology	صرف
open source	متن‌باز
orthographic	نگارشی
parsing	سازه‌یابی
phonology	آواشناسی
phrase	گروه
polysemy	چند معنایی
post-editing	پس ویرایش
pre-editing	پیش ویرایش
punctuation	علائم سجاوندی
recursive	دوری
rule-based	قاعده بنیان
rules post-processed by statistics	پس‌ویرایش آماری قواعد
scenario	فیلمنامه
scripts	نمایشنامه
semantics	معنی‌شناسی
shallow	سطحی
shallow-transfer machine translation	ترجمه ماشینی سطحی
statistics guided by rules	ویرایش آمار با قواعد
structural transfer	انتقال ساختاری

subcategorization frame	چارچوب زیرمقوله ای
subcategorization rules	قواعد زیرمقوله ای
substitution	جایگزین‌سازی
subworld	عالم فرعی
synset	دایره معنایی
syntax	نحو
text understanding	درک متن
tokenization	واحدسازی
transfer-based machine translation	ترجمه ماشینی انتقال بنیان
universal encyclopedia	دائره‌المعارف جهانی
unrestricted	محدود نشده
well-formed	خوش‌ساخت
word Order	ترتیب کلمات
word-sense disambiguation	ابهام‌زدایی از معنای واژه

فهرست منابع

بیجن‌خان، م و مرادزاده، ش.، (۱۳۸۳)، "پیکره متنی زبان فارسی" ، مجموعه سخنرانی‌های اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران.

ثانی، غ و سلیمانی، ع، (۱۳۷۸)،" ترجمه نحوی انگلیسی به فارسی به روش تجزیه چارت"، زبان فارسی ورایانه، تهران.

حق شناس، ع، (۱۳۸۵). دستور زبان فارسی، وزارت آموزش و پرورش.

داودآبادی، م. و پالهنگ،م.، (۱۳۸۴)، "دانا: عاملی با قابلیت درک و نوشتار فارسی"، یازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران.

سمائی، م.، (۱۳۷۷)، " واژگان در دستور سنج: انگاره نظری"، رساله دکتری، دانشکده ادبیات و علوم انسانی، دانشگاه تهران.

-------، (۱۳۸۱)، پردازش گروه اسمی، فصلنامه اطلاع رسانی، دوره ۱۸، شماره ۱ و ۲.

-------، (۱۳۸۴). گروه اعداد در ترجمه ماشینی. *فصلنامه علوم و فناوری اطلاعات،۲۱ (۱)*، ۳۳-۲۱.

سمیعی، ا. (مترجم)، (۱۳۶۲)، **ساخت های نحوی**، نوشته نعام چامسکی، انتشارات خوارزمی.

شمس‌فرد، م.، (۱۳۷۴)، "درک متن فارسی"، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف.

شمس‌فرد، م.، (۱۳۸۵)، "پردازش متون فارسی: دستاوردهای گذشته، چالش‌های پیش‌رو"، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه.

صادقی، ع، (۱۳۵۶). دستور زبان فارسی (سال دوم فرهنگ و ادب)، انتشارات آموزش و پرورش.

صامتی، ح و بیجن‌خان، م.، (۱۳۸۹) **زبان فارسی و رایانه**، سمت.

فیلی، ه.، (۱۳۸۳)، " ترجمه ماشینی انگلیسی به فارسی بر مبنای نظریه احتمالات و گرامر لغوی"، مجموعه سخنرانی‌های اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران.

فیلی، ه و ثانی، غ، (۱۳۸۳) ،" ترجمه ماشینی انگلیسی به فارسی بر مبنای نظریه احتمالات و دستور کلمه"، زبان فارسی و رایانه.

مشکوةالدینی، م.، (۱۳۷۳). دستور زبان فارسی (بر پایه نظریه گشتاری)، انتشارات دانشگاه فردوسی.

ناصح، م.، (۱۳۸۵)، "فهرست پایان‌نامه‌های ازائه شده در زمینه زبان فارسی و رایانه، ۱۳۵۵- ۱۳۸۴"، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه.

نوربالا، م.، (۱۳۸۰)، "طراحی یک واژگان جامع برای پردازشگر زبان فارسی"، پایان‌نامه کارشناسی ارشد، دانشگاه علم و صنعت.

Amtrup, J., Mansouri Rad, H., Megerdoomian, K., Zajac .(2000). *Persian – English Machine Translation: An Overview of the Shiraz Project*, Memoranda in Computer and Cognitive Science.
between English and Persian Texts, Third International Workshop on Computational Approaches to Arabic-Script based languages (CAASL3), Ottawa, Canada.

Booth, K. H. V. (1967). Machine aided translation with a post-editor. In A. D. Booth (Ed.), *Machine translation* (pp. 53-76). Amesterdam: North-Holland Publishing Company.

Crystal, D, 1992, an encyclopedic dictionary of language and languages, Blackwell.

Delvaney, Emile (1972). *La machine a traduire.* Paris: PUF.

Dreyfus, H. L. & Dreyfus, S. E. (1989). *Mind over machine.* Oxford , Uk: B. Blackwell.

Fellbaum, C. 1998. *WordNet An Electronic Lexical Database*, MIT Press. Banerjee, S. and T. Pederson 2002. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*. Lecture Notes in Computer Science, Berlin/Heidelberg, Springer.

Harabagiu, S. M., G. A. Miller, rt al. 1999. Wordnet 2 – *a morphologically and semantically enhanced resource*, University of Maryland, SIGLEX-99.

Katz, J., & Fodor, J. (1963). The structure of a semantic theory. *Language*, 39, 170 – 210.

Megredoomian. M., 2000, *Persian Computational Morphology: AUnification-Based Approach,* Memoranda in Computer and Cognitive Science. New Mexico State University.

Katz, J., & Postal, P. M. (1964). *An integrated theory of linguistic descriptions*. Cambridge: MIT press.

Kiani. S. 2009, *Persian text tokenization and chunking*, 14th international CSI Computer Conference. Tehran. Iran.

Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone frome an ice craem cone.* SIGDOC 86:the 5th annual international conference on Systems documentation, New York, USA, ACM Press.

LISA (Libray and Information Science Abstracts). (1997).

Mahmoudi, S. M. (1994). *Contribution au traitment automatique de la langue persan.*These de doctorat, Universite de Lyon II .

Motazedi, Y., Saedi, C., Shamsfard, M., (2009). "Automatic Translation

Rosseta, M.T. (1994). *Rosseta Compositional Translation,* Cluwer Academic Publishers.

Rouhizadeh. M., M. Shamsfard and M. A. Yarmohamadi, 2008, Buiding a Wordnet for Persian Verbs, 4th Global WordNet conference (GWC 2008), Szeged, Hungary.

Shamsfard. M., H. Sadat Jafari and M. Ilbeigy, 2010, *A Set of Fundamental Tools for Persian Text Processing.* Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta.

http://en.wikipedia.org/wiki/Machine_translation (accssed at 02.03.2011)

سایر منابع و ماخذ:

[i] Nagao, M. 1981. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, in Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds.) North- Holland, pp. 173-180, 1984.

[ii] "the Association for Computational Linguistics - 2003 ACL Lifetime Achievement Award". Association for Computational Linguistics. http://www.aclweb.org/index.php?option=com_content&task=view&id=36&Itemid=30. Retrieved 2010-03-10.

[iii] Boretz, Adam, "AppTek Launches Hybrid Machine Translation Software" SpeechTechMag.com (posted 2 MAR 2009)

[iv] Milestones in machine translation - No.6: Bar-Hillel and the nonfeasibility of FAHQT by John Hutchins

[v] Bar-Hillel (1960), "Automatic Translation of Languages". Available online at http://www.mt-archive.info/Bar-Hillel-1960.pdf

[vi] Melby, Alan. The Possibility of Language (Amsterdam:Benjamins, 1995, 27-41)

[vii] Wooten, Adam. "A Simple Model Outlining Translation Technology" T&I Business (February 14, 2006)

[viii] Appendix III of 'The present status of automatic translation of languages', Advances in Computers, vol.1 (1960), p.158-163. Reprinted in Y.Bar-Hillel: Language and information (Reading, Mass.: Addison-Wesley, 1964), p.174-179.

[ix] Human quality machine translation solution by Ta with you