



مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

زبان‌شناسی رایانشی: مدل‌ها، منابع و کاربردها

نوشته: ایگور بلشاکف و الکساندر گلیخ

ترجمه:

محمد رضا فلاحتی قدیمی فومنی

دانشیار مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

فاطمه احمدی نسب

استادیار دانشگاه شهید چمران اهواز

چاپ دوم

شهریور ۱۴۰۱

- سرشناسه : بلشاکف، ایگور ا. Bolshakov, Igor A.
- عنوان و نام پدیدآور : زبان‌شناسی رایانشی: مدل‌ها، منابع و کاربردها/ نوشته ایگور بلشاکف و الکساندر گلیخ؛ ترجمه محمدرضا فلاحتی‌قدیمی‌فومنی، فاطمه احمدی‌نسب.
- مشخصات نشر : شیراز: وزارت علوم، تحقیقات و فناوری، معاونت پژوهشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، ۱۴۰۱.
- مشخصات ظاهری : ۲۰۶ ص.؛ مصور، نمودار.
- شابک : 978-622-9514-43-6
- وضعیت فهرست نویسی : فیبا
- یادداشت : عنوان اصلی: Computational Linguistics: Models, Resources, Applications, 2004.
- موضوع : زبان‌شناسی کامپیوتری
- موضوع : Computational linguistics
- موضوع : پردازش زبان طبیعی
- موضوع : (Natural language processing (Computer science
- شناسه افزوده : گلیخ، الکساندر، ۱۹۶۲- م.
- شناسه افزوده : -Gelbukh, Alexander, 1962
- شناسه افزوده : فلاحتی‌قدیمی‌فومنی، محمدرضا، ۱۳۵۲-، مترجم
- شناسه افزوده : Falahati Qadimi Fumani, Mohammad Reza
- شناسه افزوده : احمدی‌نسب، فاطمه، ۱۳۵۸-، مترجم
- شناسه افزوده : ایران. وزارت علوم، تحقیقات و فناوری. مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری
- شناسه افزوده : Regional Information Center for Science and Technology
- رده بندی کنگره : P۹۸/۸۲ب ۱۳۹۷
- رده بندی دیویی : ۲۸۵/۴۱۰
- شماره کتابشناسی ملی : ۵۵۳۰۱۸۴



مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

زبان‌شناسی رایانشی: مدل‌ها، منابع و کاربردها

نوشته: ایگور بلشاکف و الکساندر گلیخ

ترجمه: محمدرضا فلاحتی‌قدیمی‌فومنی / فاطمه احمدی‌نسب

ناشر: اداره انتشارات مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری (RICEST)

شابک: ۹۷۸-۶۰۰-۹۸۲۰۵-۳-۵

شمارگان: ۲۰۰ نسخه | چاپ اول: زمستان ۱۳۹۷ | چاپ دوم: تابستان ۱۴۰۱

معاونت: ۰۷۱-۳۶۴۶۷۰۸۴ | انتشارات: ۰۷۱-۳۶۴۶۸۴۵۲ | نمابر: ۰۷۱-۳۶۴۶۸۳۵۲

صفحه خانگی: www.ricest.ac.ir | پست الکترونیک: publication@ricest.ac.ir

مقدمه مترجمان

کتاب «زبان‌شناسی رایانه‌ای: مدل‌ها، منابع و کاربردها» یکی از آثاری است که به زبان ساده به توضیح مفاهیم اصلی و زبان‌شناختی بنیادی در حوزه زبان‌شناسی رایانه‌ای پرداخته است و مسائل و دغدغه‌های زبان‌شناسی رایانه‌ای را با مثال‌هایی از زبان اسپانیولی مطرح می‌کند. شاید برخی از مخاطبین خرده بگیرند که اولاً تاریخ انتشار نسخه انگلیسی کتاب ۲۰۰۴ است و برای حوزه‌ای پویا و مدام در حال پیشرفت مانند زبان‌شناسی رایانه‌ای اثری قدیمی به شمار می‌رود و دیگر آن که خواننده فارسی زبان چه نیازی به مطالعه و بررسی زبان اسپانیولی دارد. در مورد اول باید گفت که خوشبختانه این کتاب مفاهیم اساسی و مسائل زبان‌شناختی زبان‌شناسی رایانه‌ای را به زبانی ساده بیان کرده است و حتی با گذشت بیش از ۱۵ سال از تاریخ انتشار نسخه اصلی کتاب مطالب هنوز از اهمیت برخوردار بوده چرا که به ویژگی‌های زبانی توجه داشته و هر خواننده علاقمند به زبان‌شناسی رایانه‌ای باید نحوه مطالعه و توجه به ویژگی‌های زبانی را بیاموزد. بنابراین کتاب حاضر می‌تواند دیدگاه بهتری را از مسائل زبانی در اختیار مخاطبین علاقمند به هوش مصنوعی، زبان‌شناسی رایانه‌ای و حتی علاقمندان به مسائل زبان‌شناختی محض قرار دهد. در مورد دوم، مترجمین این اثر معتقدند که اتفاقاً آشنایی با دغدغه‌های زبان اسپانیولی در فضای رایانه‌ای حتی برای مخاطب فارسی زبان نیز یک مزیت است. چرا که به علت سیطره علمی زبان انگلیسی بیشتر مخاطبین این حوزه‌ها با مسائل زبان انگلیسی آشنا شده‌اند و از قضا آشنایی با مسائل یک زبان دیگر مانند اسپانیولی دید بهتری از مسائل کلی زبان‌ها و تنوع آنها و لزوم توجه به این تفاوت‌ها و شباهت‌ها را در اختیار مخاطب قرار می‌دهد و اتفاقاً می‌تواند چراغ راه برای مطالعه زبان فارسی و مناسب‌سازی آن با فضای رایانه و حل دغدغه‌های زبان‌شناختی رایانه‌ای این زبان باشد. در ترجمه این اثر تا حد ممکن سعی شده است که از اصطلاحات زبان‌شناختی مصوب فرهنگستان زبان و ادب فارسی و اصطلاحات رایج در حوزه زبان‌شناختی و مورد توافق غالب زبان‌شناسان استفاده شود. در برخی از موارد مترجمان توضیحاتی را لازم دانستند که به صورت پاورقی ارائه کرده‌اند. علاوه بر این در برخی از موارد، مسائل و پرسش‌هایی که به لهجه‌های محلی اسپانیولی مربوط می‌شد و یا باعث سردرگمی خواننده می‌گردید حذف شده و به جای آن از مثال‌های فارسی استفاده شد. در ضمن ویرایش این اثر تا حد ممکن بر اساس مصوبات فرهنگستان زبان و ادب فارسی صورت گرفته است.

ب زبان‌شناسی رایانشی: مدل‌ها، منابع و کاربردها

امید است که کتاب فوق مورد توجه مخاطبین فارسی‌زبان قرار گرفته و پاسخگوی بخشی از دغدغه‌های علمی آنها در حوزه زبان‌شناسی رایانه‌ای باشد. مترجمین این اثر آماده دریافت نظرات و انتقادات مخاطبین عزیز خود به رایانامه مترجم اول اثر به نشانی mrfalahat@yahoo.com هستند.

با آرزوی سربلندی همگان

فهرست مندرجات

پیشگفتار.....	۱
کتابی جدید دربارهٔ زبان‌شناسی رایانه‌ای.....	۱
اهداف و مخاطبان کتاب حاضر.....	۶
ارتباط با علوم رایانه.....	۸
ارتباط با هوش مصنوعی.....	۹
انتخاب موضوعات.....	۱۰
منابع وبی برای کتاب حاضر.....	۱۱
سپاس‌گزاری.....	۱۱
۱. مقدمه.....	۱۲
نقش پردازش زبان طبیعی.....	۱۲
زبان‌شناسی و ساختار آن.....	۱۴
زبان‌شناسی رایانه‌ای چیست؟.....	۲۴
واژه چیست؟.....	۲۵
نقش مهم دانش زیربنایی.....	۲۸
نتیجه‌گیری.....	۳۱
۲. نگاهی تاریخی به موضوع.....	۳۳
رویکرد ساخت‌گرایان.....	۳۳
ظهور چامسکی.....	۳۴
دستور بافت-آزاد ساده.....	۳۵
دستورهای گشتاری.....	۳۸
پژوهش‌های زبان‌شناختی پس از چامسکی: ظرفیت‌ها و تفسیر.....	۳۹
پژوهش‌های زبان‌شناسی بعد از چامسکی: محدودیت‌ها.....	۴۲
دستور ساخت‌گروهی هسته‌بنیان.....	۴۴
رویکرد وحدت.....	۴۵
نظریهٔ معنا ⇔ متن: مبدل چندمرحله‌ای و الگوهای حاکمیت.....	۴۷
نظریهٔ معنا ⇔ متن: نمودار درختی وابستگی.....	۴۹
نظریهٔ معنا ⇔ متن: پیوندهای معنایی.....	۵۱
نتیجه‌گیری.....	۵۳

۵۴ محصولات زبان‌شناسی رایانه‌ای: حال و آینده
۵۴ طبقه‌بندی سامانه‌های زبانی کاربردی
۵۵ تیره‌گذاری خودکار
۵۷ غلط‌یاب املایی
۵۹ غلط‌یاب دستوری
۶۱ غلط‌یاب سبکی
۶۳ ارجاع به واژه‌ها و ترکیب‌های واژگانی
۶۶ بازیابی اطلاعات
۷۰ خلاصه‌نویسی موضوعی
۷۴ ترجمه خودکار
۷۸ رابط زبان طبیعی
۷۹ استخراج داده‌های واقعی از متون
۸۰ تولید متن
۸۱ نظام‌های درک زبان
۸۳ نظام‌های مرتبط
۸۶ نتیجه‌گیری
۸۷ ۴. زبان به عنوان مبدل معنا \Leftrightarrow متن
۸۷ رویکردهای مختلف در خصوص زبان طبیعی
۸۹ زبان به عنوان یک مبدل دوسویه
۹۴ متن چیست؟
۹۹ معنا چیست؟
۱۰۲ دو راه برای نمایش معنا
۱۰۵ تجزیه و تحلیل معنا
۱۰۷ غیر منحصراً به فرد بودن انطباق معنا \leftarrow متن: هم‌معنایی
۱۰۹ غیر منحصراً بودن انطباق متن \leftarrow معنا: هم‌نامی
۱۱۲ جستاری بیشتر در مبحث هم‌نامی
۱۱۷ ویژگی چندسطحی گشتار معنا \Leftrightarrow متن
۱۲۱ ترجمه به عنوان یک گشتار چندمرحله‌ای
۱۲۳ دو روی یک نشانه
۱۲۴ نشانه زبانی

زبان‌شناسی رایانشی: مدل‌ها، منابع و کاربردها ۵

- نشانه‌زبانی در MMT ۱۲۴
- نشانه‌زبانی در HPSG (دستور ساخت‌گروهی هسته‌بنیان) ۱۲۵
- آیا دال‌ها طبیعی یا قراردادی هستند؟ ۱۲۶
- مقایسه دیدگاه‌های زایشی، MMT و محدودیت‌بنیان ۱۲۷
- نتیجه‌گیری ۱۳۴
۵. مدل‌های زبان‌شناختی ۱۳۵
- مدل‌سازی به چه معناست؟ ۱۳۵
- مدل‌های عصب‌شناختی زبان ۱۳۶
- مدل‌های روان‌شناختی زبان ۱۳۷
- مدل‌های نقش‌گرای زبان ۱۳۸
- مدل‌های پژوهشی زبان‌شناختی ۱۴۰
- ویژگی‌های مشترک مدل‌های جدید زبانی ۱۴۰
- ویژگی‌های منحصر به فرد مدل معنا \leftrightarrow متن ۱۴۳
- مدل‌های کاهش‌یافته ۱۴۷
- آیا واقعاً به مدل‌های زبان‌شناختی نیازمندیم؟ ۱۴۹
- قیاس در زبان‌های طبیعی ۱۵۰
- رویکردهای تجربی در برابر رویکردهای عقلی ۱۵۲
- قلمرو محدود نظریه‌های زبان‌شناختی نوین ۱۵۴
- نتیجه‌گیری ۱۵۷
- تمرین‌ها ۱۵۸
- پرسش‌های تشریحی ۱۵۸
- پرسش‌های چهارگزینه‌ای ۱۶۱
- فهرست منابع ۱۷۳
- پیوست‌ها ۱۷۶
- واژه‌نامه فارسی به انگلیسی ۱۷۹
- واژه‌نامه انگلیسی به فارسی ۱۸۹

پیشگفتار

در آغاز شاید این سؤال مطرح شود که علی رغم وجود کتاب‌های متعدد در خصوص زبان‌شناسی رایانشی، علت پیشنهاد و نگارش کتابی جدید در این خصوص از سوی نویسندگان حاضر چه بوده است؟ اهداف عمده‌ای که در کتاب حاضر دنبال می‌شود، چیست؟ مخاطبان کتاب چه کسانی هستند؟ ویژگی‌های اصلی این کتاب چیست؟ و سرانجام، بین محتوای کتاب حاضر و شاخه‌های مختلف علوم رایانه‌ای چه ارتباطی وجود دارد؟ در پیشگفتار حاضر سعی خواهد شد برای پرسش‌های بالا پاسخ‌هایی جامع و مانع ارائه گردد.

کتابی جدید دربارهٔ زبان‌شناسی رایانه‌ای

در سال‌های اخیر، موفقیتی که نرم‌افزارهای جدید پردازش زبان طبیعی کسب نموده‌اند همگان را به تحسین واداشته است. این روزها، برنامه‌ها و نرم‌افزارهای غلطیاب املائی و دستوری^۱، بازیابی اطلاعات از دادگان‌های اسناد^۲، ترجمهٔ بین‌زبانی و مواردی از این قبیل در نسخه‌های میلیونی در سطح جهان تکثیر و فروخته می‌شود.

با این همه، باید اذعان کرد که چنین برنامه‌هایی هنوز فاقد قوهٔ هوش و ادراک واقعی هستند. در حقیقت، ایجاد نرم‌افزارهای درک و تولید گفتار با چنان کیفیتی که بتوانند عملیات ترجمهٔ خودکار و نیز ارتباط انسان-ماشین را به گونه‌ای مطلوب در زبان طبیعی با همه پیچیدگی‌ها و ظرایفش به انجام برسانند، هنوز در حد یک بلندپروازی باقی مانده و در عمل محقق نگردیده است. البته، باید اذعان نمود که برای نیل به این هدف و رفع مشکلات تحقق آن، در طی نیم قرن گذشته تلاش‌هایی فراوان و ارزنده به عمل آمده است.

با توجه به سابقهٔ طولانی تلاش‌های به عمل آمده، چنین برمی‌آید که تولیدکنندگان

نرم‌افزارهای جدید باید به جای راه‌حل‌های موردی و مقطعی^۱ از روش‌ها و نتایج موجود از حوزه زیربنایی دیگر زبان‌شناسی استفاده نمایند. در حقیقت، هیچ یک از پیشرفت‌های جدید مانند افزایش سرعت رایانه، پالایش و بهینه‌سازی ابزارهای برنامه‌نویسی و توسعه بیشتر نظام‌های محدود درک زبانی^۲ در حیطه‌های خاص نتوانسته‌اند در خصوص یکی از بحث‌برانگیزترین مسائل علوم جدید به نام درک خودکار متن^۳ توفیق لازم را کسب نمایند.

نویسندگان اثر حاضر معتقدند مسئله درک خودکار متن که در قرن گذشته لاینحل باقی ماند، در اوایل قرن حاضر و توسط افرادی مرتفع خواهد گردید که هم اکنون به عنوان دانشجویانی جوان بر روی نیمکت دانشگاه‌ها نشسته‌اند. در حقیقت، کتاب حاضر به بررسی مدل‌های زبان‌شناسی رایانه‌ای و کاربرد آنها می‌پردازد و مخاطبان آن را همین قشر از دانشجویان تشکیل می‌دهد، دانشجویانی که در حال حاضر در دانشگاه‌های سراسر جهان و در حوزه علوم و فناوری رایانه^۴ مشغول به تحصیل هستند و مایلند در زمینه تولید نرم‌افزارهای پردازش زبان طبیعی^۵ فعالیت نمایند. بنابراین، انتظار ما این است که مخاطبان کتاب حاضر تا حدی با علوم رایانه آشنایی داشته باشند؛ هر چند آشنایی خاص و تخصصی با مفاهیم علوم انسانی (به طور عام) و زبان‌شناسی (به طور خاص) پیش فرض مطالعه کتاب حاضر نیست.

به تازگی، منابع و کتاب‌های گوناگونی در خصوص پردازش زبان طبیعی وارد بازار گردیده که از آن جمله می‌توان به [۱، ۲، ۷، ۹] اشاره نمود. این منابع، گزینه‌هایی بسیار مناسب هستند و می‌توان آنها را در برنامه آموزشی دانشجویان حوزه رایانه و علاقه‌مند به حوزه پردازش زبان طبیعی گنجانند. با وجود این، برای افرادی که دانش زبان‌شناسی اندکی دارند (و دانشجویان علوم رایانه نیز از جمله این افراد به شمار می‌روند)، درک محتوای کتاب‌های فوق تا حدی سنگین بوده و به منابع و جزوات جانبی و کمکی نیاز دارند. در حقیقت، کتاب‌های فوق برای این دسته از دانشجویان به دلایل زیر نامناسب است:

۱- بسیاری از این کتاب‌ها برای زبان انگلیسی نوشته شده است. نکته قابل تأمل این است که

1 ad hoc

2 language understanding system

3 automatic text understanding

4 computer science and technology

5 Natural Language Processing (NLP)

زبان انگلیسی و اسپانیولی^۱ علی‌رغم شباهت‌های واژگانی فراوان، از نظر ساختار دستوری با یکدیگر تفاوت فاحشی دارند. زبان انگلیسی بر خلاف اسپانیولی، از قواعد ترتیب واژگانی^۲ بسیار خشکی برخوردار بوده و دارای اصول ساختواژه^۳ بسیار ساده‌ای است. از این رو، تعمیم کورکورانه تحلیل‌ها و قواعد نحوی^۴ و ساختوازی انگلیسی به اسپانیولی نادرست و خطا خواهد بود.

۲- تنها تعداد اندکی از این گونه کتاب‌ها به بررسی جامع و مقابله‌ای نظریه‌های پردازش متن^۵ پرداخته‌اند. ضمناً، همان اندک آثاری نیز که در این حوزه وارد شده‌اند به توفیق چندانی دست نیافته‌اند، چرا که در رویکردشان از الگوهای روش‌شناختی بسیار پراکنده و نامتجانسی استفاده کرده‌اند. این پراکندگی تا بدانجاست که الگوهای ارائه شده گاه از نظر تعاریف، مفاهیم و نشانه‌گذاری‌ها^۶ در تضاد با یکدیگرند.

۳- در عمده این کتاب‌ها بیشتر به قواعد نحو^۷ توجه شده است. در برخی از این کتاب‌ها، مبحث زبان‌شناسی رایانه‌ای حتی به دانشی دربارهٔ نحو زبان انگلیسی تنزل پیدا کرده و محدود شده است. با وجود این، زبان‌شناسی در مفهوم کلی خود، به مطالعه سطوح مختلف زبانی از جمله واج‌شناسی^۸، ساخت‌واژه، نحو و معناشناسی^۹ می‌پردازد. در کتاب‌های مورد اشاره هر یک از سطوح زبانی مذکور با نگاه صرف زبان‌شناسانه بررسی شده است و حال آنکه در نوشتار حاضر این سطوح زبانی در ارتباط با زبان‌شناسی رایانشی مورد بررسی و تحلیل قرار می‌گیرد.

با توجه به تمام پیچیدگی‌ها و اختلاف نظرهای موجود در این رشته، که اتفاقاً با آهنگی پرشتاب در حال گسترش است، برای کتاب حاضر مشخصه‌های روش‌شناختی زیر ترسیم گردید:

۱- تقریباً تمامی مثال‌های به کار رفته در این اثر از اسپانیولی گرفته شده است. تنها در دو مورد از زبان‌های دیگر شاهد و مثال آورده شده است: (۱) زمانی که مقایسه دو یا چند زبان

۱. مؤلفین کتاب حاضر در دانشگاه‌های اسپانیولی زبان مشغول به کار هستند، در نتیجه در اثر حاضر به زبان اسپانیولی توجه داشته‌اند [مترجمان].

2 word order

3 morphology

4 syntactic rules

5 text processing

6 notation

7 formalisms of syntax

8 phonology

9 semantics

ضروری بود و (۲) زمانی که مشخصه مورد بحث در اسپانیولی وجود نداشت.

۲- سعی مؤلفان بر این بوده تا حقایق و نکاتی بسیار متنوع و البته مرتبط با پردازش زبان طبیعی را از حوزه زبان‌شناسی همگانی استخراج و در کتاب حاضر منعکس نمایند. از این رو، در تهیه مطالب این مجموعه، صرفاً به کلی‌گویی و ذکر مفاهیم تکراری موجود در کتاب‌های زبان انگلیسی مختلف بسنده نشده است.

۳- در سرفصل کتاب حاضر، رویکردهای مختلف زبان‌شناسی رایانه‌ای در هم تلفیق گردیده و هر جا که مقدور بوده ارتباط بین این رویکردها به تصویر کشیده شده است.

۴- مطالب در کتاب حاضر به شکلی تدریجی و با آهنگی کند ارائه می‌گردد تا برای خوانندگان فاقد پیشینه زبان‌شناسی نیز قابل درک باشد.

در حقیقت، نظر نویسندگان حاضر این بوده است که گردآوری رویکردهای نامتجانس آن هم در یک مجلد و ذیل یک عنوان نامناسب است و از این رو تمام سعی خود را نموده‌اند که کتاب حاضر به یک کتاب مرجع تبدیل نگردد و برای هر چیز و هر نکته‌ای به منابع بسیار و آثار پژوهشگران برجسته ارجاع داده نشود. بلکه برعکس، در کتاب حاضر به پیوستگی، یکپارچگی و استقلال اثر، اهمیت بسزایی داده شده است.

در ادامه، دو رویکردی که بیش از هر رویکرد دیگری بر محتوای کتاب حاضر تأثیرگذار بوده است، معرفی خواهد شد:

۱- رویکرد اول به نظریه معنا \Leftrightarrow متن^۱ (MTT) مربوط است که در نیمه دهه شصت توسط ایگور ملچوک، الکساندر ژلکوفسکی و یوری آپرسیان معرفی شد. این نظریه، توصیف عناصر، سطوح و ساختار زبان‌های طبیعی را آسان‌تر نموده است. نظریه معنا- متن را می‌توان برای توصیف تمامی زبان‌ها به کار گرفت به خصوص برای زبان‌های ترتیب- آزاد^۲ مانند اسپانیولی^۳. علاوه بر این، نظریه MTT امکان اعتباربخشی و گسترش اصطلاحات تخصصی^۴ و روش‌شناسی سنتی زبان‌شناسی را فراهم نموده است.

1 Meaning Text Theory (MTT)

2 free word order

۳ زبان فارسی نیز یک زبان ترتیب آزاد است [مترجمان].

4 terminology

۲- دستور ساخت گروهی هسته‌بنیان^۱ (HPSG) در دهه گذشته و توسط کارل پلارد و ایوان ساگ مطرح گردید. این دستور، در سنت جدید دستور زایشی^۲ چامسکی، به عنوان پیشرفته‌ترین صورت‌گرایی کاربردی^۳ در توصیف و پردازش زبان طبیعی شناخته شده است. دستور ساخت گروهی هسته‌بنیان همانند نظریه معنا- متن از تمامی حقایق شناخته شده برای توصیف زبان‌های طبیعی بهره می‌جوید و حقایق جدیدی را به آن می‌افزاید. این نظریه مانند بیشتر نظریه‌های صورت‌گرایی موجود، عمدتاً بر روی زبان انگلیسی محک زده شده است. البته، در سال‌های اخیر این نظریه به شدت مورد توجه پژوهشگران زبان‌های مختلف از جمله اسپانیولی قرار گرفته است. از آن جا که تاکنون بخش عمده تحقیقات مربوط به پردازش زبان طبیعی در چارچوب نظریات چامسکی به انجام رسیده، لازم است متخصص زبان‌شناسی رایانه‌ای در خصوص رویکرد دستور زایشی اطلاعات بیشتری را کسب نماید.

در انتخاب مطالب کتاب حاضر، از مشاهدات و تجارب شخصی خود و همکاران‌مان و مشکلاتی که در دوران فعالیت حرفه‌ای خود در حوزه زبان‌شناسی رایانه‌ای با آن مواجه بوده‌ایم استفاده شده است. مطالب عنوان شده در کتاب حاضر به مشکلاتی می‌پردازد که حتی در زمان حاضر نیز برای برنامه‌نویسان فاقد پیشینه زبان‌شناسی، دغدغه‌ای اساسی محسوب می‌گردد. خوانندگان نوشتار حاضر پس از آشنایی با محتوا و مفاهیم مطرح شده با اعتماد به نفس و اطمینان بیشتری می‌توانند به مطالعه شاخه‌های مختلف زبان‌شناسی رایانه‌ای بپردازند. این حوزه از دانش دارای زیربخش‌های مختلفی است که از آن جمله می‌توان به موارد زیر اشاره نمود:

- ابزارها و ساخت‌های ریاضی زبان‌شناسی رایانه‌ای
- واج‌شناسی
- ساخت‌واژه
- نحو در هر دو سطح روساخت و ژرف‌ساخت
- معناشناسی

محتوای کتاب حاضر از مطالبی برگرفته شده است که نویسندگان آن از سال ۱۹۹۷ تاکنون

1 head-driven phrase structure grammar

2 generative grammar

3 practical formalism

در قالب درس زبان‌شناسی رایانه‌ای در مرکز پژوهش‌های رایانشی واقع در مؤسسه ملی پلی‌تکنیک مکزیکو سیتی تدریس نموده‌اند. دوره درس مذکور به ارائه ایده‌ها و مفاهیم ضروری برای ساخت نرم‌افزارهای هوشمند پردازش زبانی می‌پردازد. البته سعی بر آن بود که از پرداختن به جزئیات و پیچیدگی‌های الگوریتم‌های خاص نظام‌های بازی اجتناب شود. در حقیقت، نظر نگارندگان این مجموعه آن است که پرداختن عملی به مباحثی چون الگوریتم‌ها، معماری و نیز نگهداری نظام‌های زبانی کاربردی واقعی از حوصله کتاب حاضر خارج بوده و باید در اثری جداگانه مورد مطالعه و بحث قرار گیرد.

از آنجایی که بیشتر تحقیقات به عمل آمده پیرامون موضوع حاضر (بدون توجه به محل انجام)، به انگلیسی به چاپ رسیده، شایسته است دانشجویان مقدمه‌ای را در خصوص این رشته و به زبان انگلیسی مطالعه نمایند.

برای تماس با نویسندگان کتاب حاضر می‌توانید از طریق رایانامه‌های igor@cic.ipn.mx، gelbukh@gelbukh.com و gelbukh@cic.ipn.mx و همچنین آدرس‌های اینترنتی www.gelbukh.com و www.cic.ipn.mx/~gelbukh اقدام نمایید. ضمناً، کتاب حاضر از طریق آدرس اینترنتی www.gelbukh.com/clbook قابل دسترسی است.

اهداف و مخاطبان کتاب حاضر

از جمله مهم‌ترین اهداف کتاب حاضر عبارت است از: (۱) ارائه اطلاعاتی مختصر اما زیربنایی به دانشجویان در مورد زبان‌شناسی همگانی، (۲) توصیف مدل‌های زبانی جدیدی که عملکرد زبان‌های طبیعی را توصیف می‌کنند و (۳) تشریح این نکته که در چارچوب یک زبان طبیعی خاص چگونه می‌توان داده‌های مورد نیاز برای ساخت نظام‌های پردازش زبان طبیعی نظیر جداول زبانی^۱ و واژه‌نامه‌های ماشینی^۲ را کامپایل کرد. بنابراین، هدف نهایی کتاب حاضر آن است که به خوانندگان بیاموزد چگونه می‌توان ابزار لازم را برای ساخت برنامه‌ها و نظام‌های پردازش زبان طبیعی خودکار تهیه کرد. شایان ذکر است که مخاطبان کتاب حاضر عمدتاً دانشجویان علوم

1 linguistic tables

2 machine dictionaries

رایانه (شامل نرم‌افزار، سخت‌افزار، مدیریت پایگاه داده‌ها، بازیابی اطلاعات، هوش مصنوعی^۱) هستند.

در تمامی فصول کتاب حاضر، همواره به کاربرد مطالب عنوان شده در زبان اسپانیولی توجهی خاص شده است. با این همه، این تصور که مجموعه حاضر صرفاً به اسپانیولی پرداخته، تصویری باطل است. در حقیقت، برای درک بسیاری از مفاهیم اساسی، از زبان‌های دیگر مانند انگلیسی، فرانسه، پرتغالی، روسی نیز شاهد و مثال آورده شده است. از این رو، کتاب حاضر در یک سطح عام برای مخاطبینی از زبان‌های دیگر نیز مفید و قابل استفاده است. بسیاری از منابعی که در بخش پیشینه ذکر گردیده است اطلاعات و مثال‌های بسیار جالبی را در اختیار خواننده قرار می‌دهد. با این همه، همان‌طور که پیش‌تر نیز ذکر شد، هر جا که مقدر بوده و به منظور مقایسه، در کنار مثال‌های اسپانیولی از زبان‌های دیگر نیز شاهد و مثال آورده شده است.

از جمله مشکلات قابل توجه در این حوزه، عدم تسلط دانشجویان مؤسسات فنی آمریکای لاتین بر دانش و اطلاعات زبان‌شناسی است. در حقیقت، این دسته از دانشجویان صرفاً اطلاعاتی مختصر و به جا مانده از دوران مدرسه با خود به همراه دارند. از این رو، سعی شده کتاب حاضر به گونه‌ای نوشته شود که برای مخاطبان فاقد هرگونه اطلاعات دستوری مقدماتی نیز قابل درک باشد. ضمناً، نباید فراموش کرد که زبان مادری در مدارس، بیشتر به صورت تجویزی^۲ (تجویز صورت‌های ارجح گفتاری و نوشتاری) تدریس می‌شود و نه توصیفی^۳ (توصیف این که ساختار واقعی زبان چیست و به چه شکلی در جامعه به کار می‌رود). این در حالی است که تنها با انجام توصیف‌هایی علمی و کامل است که می‌توان ساختارهای زبانی درست و پذیرفتنی را از ساختارهای نادرست تشخیص داد و موارد غیرمرتبط به زبان مورد بررسی را مشخص نمود. پر واضح است که بدون ارائه توصیفی کامل و صحیح از زبان، رایانه خطاهای نامعمول و غیرمنطقی بسیاری را از دید انسان مرتکب خواهد شد. در حقیقت، عدم ارائه چنین توصیف‌هایی است که سبب گردیده است مشکلات پردازش متن با موفقیت مرتفع نگردد.

1 artificial intelligence

2 prescriptive

3 descriptive

ارتباط با علوم رایانه

هر چند در کتاب حاضر بر روی موضوعات نظری زبان تأکید بسیار شده است، نباید چنین برداشت کرد که تمرکز اصلی این مجموعه بر روی مفاهیم نظری بوده و بین علوم رایانه و زبان‌شناسی رایانه‌ای ارتباط لازم وجود ندارد. علوم رایانه و برنامه‌نویسی ابزار مهم و اصلی تمام حوزه‌های پردازش اطلاعات به شمار می‌روند. از این رو، آشنایی خوانندگان اثر حاضر با علوم رایانه و برنامه‌نویسی مقدماتی یکی دیگر از پیش‌فرض‌های نویسندگان بوده است.

هدف کتاب حاضر یاری و کمک به دانشجویان در تولید و گسترش نظام‌های نرم‌افزاری کاربردی و انتخاب مدل‌ها و ساختار داده مورد نیاز در این نظام‌هاست. البته، این تصور که ابزار و زیرساخت‌های رایانه‌ای عرضه شده در چند دهه اخیر، برای انجام پژوهش‌های نظری و زیربنایی در حیطه زبان‌شناسی رایانه‌ای کافی است، تصوری باطل است. در حقیقت، ساختاربندی مناسب برنامه‌های زبانی، فناوری هدف‌بنیاد و زبان‌های تخصصی هوش مصنوعی مانند لیسپ^۱ یا پرولاگ^۲ هیچ‌کدام نمی‌توانند به تنهایی مسائل زبان‌شناسی رایانه‌ای را مرتفع سازند. از این رو است که به تمامی این گونه تکنیک‌ها صرفاً به عنوان ابزاری مفید – و نه راه‌حلی نهایی برای رفع مسائل رایانه‌ای – نگریسته می‌شود.

همان گونه که در کتاب حاضر عنوان گردید، هدف غایی بسیاری از نظام‌های زبان‌شناسی کاربردی آن است که یک متن زبان طبیعی فرمت نشده^۳ و آماده نشده را به نوعی بازنمود معنایی^۴ تبدیل نمایند و برعکس، بازنمود معنایی یک متن را به یک متن زبان طبیعی فرمت نشده و آماده نشده تبدیل کنند. این هدف، هدف غایی هر نوع سیستم کاربردی^۵، محسوب می‌گردد.

با این همه، تولید نرم‌افزار صرفاً بخش کوچکی از کل فرایند ساخت یک سیستم پردازش زبان طبیعی را تشکیل می‌دهد. به بیان ساده‌تر، برنامه‌نویس پس از انتخاب نوع برنامه، وظایف متعددی را نیز بر عهده دارد که از آن جمله می‌توان به استخراج داده^۶ از متون مختلف و یا واژه‌نامه

1 Lisp

2 Prolog

3 unformatted

4 semantic representation

5 applied system

6 extraction of data

مکتوب و تبدیل آنها به جداول دستوری و واژه‌نامه‌های رایانه‌ای اشاره نمود. نسخه اولیه این گونه متن‌ها برای اهدافی خاص می‌تواند کاملاً فرمت نشده، نیمه فرمت شده و یا صوری شده باشند. به عنوان نمونه، اگر یک واژه‌نامه غیرماشینی داشته باشیم (خواه در قالب یک فایل متنی یا یک پایگاه داده) وظیفه برنامه‌نویس بدین شکل توصیف می‌شود که باید ابتدا هر مدخل از مدخل‌های واژه‌نامه را جستجو کند و سپس تمامی داده‌ها و اطلاعات ضروری برای انجام وظیفه فوق را از آن مدخل استخراج نماید.

کتاب حاضر حاوی اطلاعاتی است که به ما می‌آموزد چگونه وظایفی از این دست را به شکلی عادی و روزمره به انجام برسانیم. از این رو، مجدداً تأکید می‌کنیم که خوانندگان کتاب حاضر باید با برنامه‌نویسی آشنایی داشته باشند چرا که هدف نهایی کتاب حاضر نیز تهیه برنامه‌های پردازش زبانی است.

ارتباط با هوش مصنوعی

ارتباطی که بین زبان‌شناسی رایانه‌ای و هوش مصنوعی ترسیم شده، ارتباطی نسبتاً پیچیده است. آن دسته از نظام‌های هوش مصنوعی که برای پردازش زبان طبیعی در ساختار خود از زیرنظام‌ها استفاده می‌کنند، به طور مستقیم بر مفاهیم و روش‌های موجود در حوزه زبان‌شناسی رایانه‌ای بنا شده‌اند. در ضمن، برخی از روش‌های دیگری که عموماً خاص حوزه هوش مصنوعی محسوب می‌گردند مانند الگوریتم‌های جستجو برای تصمیم‌گیری در نمودارهای درختی^۱، ماتریکس‌ها و سایر ساختارهای پیچیده، اغلب در عمل در نظام‌های نرم‌افزاری زبانی نیز کاربرد دارند.

درست به همین خاطر نیز هست که بسیاری از متخصصان هوش مصنوعی [۲، ۴۰، ۵۴]، زبان‌شناسی رایانه‌ای را بخشی از این حوزه می‌دانند. اگر چه چنین تصویری تهدیدی اساسی برای زبان‌شناسی رایانشی محسوب نمی‌گردد، نگارندگان حاضر بر این باورند که قائل بودن به این نوع رابطه چندان دارای پایه و اساس نیست. دلیل این مدعا آن است که زبان‌شناسی رایانه‌ای دارای

شالوده نظری، خویشاوندان علمی، روش‌های مدیریت دانش^۱ و جستجوی تصمیم، دارای استقلال و ویژگی‌های منحصر به فرد خود است. در حقیقت، جالب آنجاست که اتفاقاً نظام‌های پردازش زبان طبیعی که در بسیاری از منابع حوزه هوش مصنوعی مورد بحث و مطالعه قرار گرفته‌اند غالباً بی‌اهمیت بوده و منسوخ شده‌اند.

هرچند بین دو حوزه زبان‌شناسی رایانشی و هوش مصنوعی تفاوت‌های زیادی وجود دارد، افرادی که با هر دو حیطه آشنا باشند از کارایی و قابلیت‌های به مراتب بیشتر برخوردارند. در حقیقت، احاطه بر هر دو حوزه، پژوهشگران را قادر می‌سازد تا در مواقع لزوم به جای دوباره‌کاری، از امکانات و آموخته‌های خود در حوزه دیگر، برای نیل به اهداف پژوهشی خود استفاده نمایند. از این رو، به خوانندگان کتاب حاضر توصیه می‌شود به شکلی شایسته خود را با حیطه هوش مصنوعی آشنا سازند. در ضمن، بر این باوریم که کتاب حاضر برای دانشجویان رشته هوش مصنوعی نیز می‌تواند مفید باشد.

انتخاب موضوعات

از آنجا که نظریه معنا \leftrightarrow متن، نه تنها سنت پیشین در زبان‌شناسی عمومی را رد نکرده و باعث غنی‌سازی و بهبود آن گردیده، ما نیز در توصیف و تشریح واقعیت‌ها و ویژگی‌های زبان‌های طبیعی عمدتاً از آن نظریه تبعیت می‌نماییم.

برای پرهیز از پیچیدگی و طولانی شدن کلام، از پرداختن به همه ویژگی‌های نظریه معنا \leftrightarrow متن صرف نظر شده است و تنها به ویژگی‌هایی توجه شده است که یا در نظریه‌های دیگر وجود ندارند و یا در آن نظریه‌ها به شکلی دیگر توصیف شده‌اند. ضمناً، در مواردی که آشنایی با مفاهیم و منابع مکتب چامسکی ضروری باشد، در کنار ویژگی‌های نظریه معنا \leftrightarrow متن، اطلاعات موازی در قالب دستور ساخت گروهی هسته‌بنیان نیز ارائه می‌گردد.

ذکر این نکته ضروری است که نظریه معنا \leftrightarrow متن تنها یک ابزار توصیف زبانی است. در این نظریه، پژوهشگر خود را به استفاده از الگوریتم‌ها یا فرمت‌های خاص نمایش داده‌های زبانی محدود نمی‌کند. این آزادی عمل در کتاب‌های جدید در حوزه زبان‌شناسی و حتی در دستور

ساخت‌گروهی هسته‌بنیان نیز به چشم می‌خورد، اگر چه دستور ساخت‌گروهی هسته‌بنیان با رویکردهای رایانه‌ای مناسب‌تری دارد. امید می‌رود روند تحولات در این دو رویکرد در جهت تکمیل و تقویت یکدیگر در آینده باشد.

منابع وبی برای کتاب حاضر

کتاب حاضر از طریق وبگاه www.Gelbukh.com/clbook قابل دسترسی است. در این وبگاه، همچنین می‌توانید به اطلاعاتی تکمیلی در خصوص موضوعات مورد بحث در کتاب حاضر و پیوند به منابع اینترنتی دسترسی پیدا نمایید. در ضمن، تعداد زیادی از آثار نویسندگان حاضر در وبگاه www.Gelbukh.com گردآوری گردیده و قابل دسترسی است.

سپاس‌گزاری

در آغاز از پروفیسور ایگور ملچوک که علاوه بر در اختیار قرار دادن جدیدترین آثار خود، پیش‌نویس کتاب حاضر را نیز به دقت مطالعه نمود و انتقادهایی بسیار صریح اما سازنده ارائه کرد بسیار سپاس‌گزاریم. همچنین از پروفیسور یوری آپرسیان و تیم تحقیقاتی‌اش در مسکو، پروفیسور لئونید تسینمان [۵۹] و پروفیسور ایگور بوگوسلاوسکی [۶۰] به خاطر ارائه اطلاعاتی مفید در خصوص کاربردهای نظریه معنا→متن و نقطه نظرات سازنده، متشکریم. دکتر پاتریک کاسیدی [۶۱] از شرکت MICRA قسمت‌هایی از نوشته حاضر را مطالعه نمود و نکاتی ارزشمند را متذکر شد. در ضمن، از سوفیا ان. گالیسیا هارو [۶۲] دانشجوی مقطع دکترای خود که در امور تدریس و تصحیح اوراق یاری‌مان داد و از این طریق این امکان را فراهم آورد تا با فراغ بال و طیب خاطر تهیه و نگارش کتاب حاضر را پیش ببریم صمیمانه سپاس‌گزاریم. وی در هر آنچه که به زبان اسپانیولی مربوط می‌شد، مشاور اصلی ما بود. در پایان، از خانواده‌هایمان به خاطر صبر و شکیبایی‌شان و نیز حمایت‌های همیشگی‌شان تشکر می‌نماییم.