واژگان زایای زبان فارسی*
(Persian Generative Lexicon)

محرم اسلامی (۱ و ۳)
مسعود شریفی آتشگاه (۲ و ۳)
صدیقه علیزاده لمجیری(۳)
طاهره زندی (۳)

(۱) دانشگاه زنجان (۲) دانشگاه تهران (۳) پژوهشکدهٔ پردازش هوشمند علائم

چکیده

واژگان (lexicon) به بخشی از دستور زبان اطلاق می‌شود که حاوی مجموعه واژه‌ها در ذهن اهل زبان است . هر کدام از واژه‌ها در واژگان ذهنی دارای اطلاعات واجی، صرفی، نحوی و معنایی منحصربه‌فردی است و با هیچ قاعده‌ای نمی‌توان اطلاعات واژگانی را صورت‌بندی کرد و این اطلاعات از جمله توانش زبانی اهل زبان به حساب می‌آید (بلومفیلد ۱۹۳۳ ص ۲۴۷ ، چامسکی ۱۹۶۵ ص ۸۷). واژه (lexeme) واحد واژگان است و واژه‌ها به عنوان واحدهای انتزعی و ذهنی برای ورود به واژگان باید حائز شرایطی باشند . از جملهٔ این شرایط داشتن ویژگی منحصربه‌فرد زبانی در مقایسه با سایر واژه‌هاست. با تکیه بر اصل فوق و دیگر اصول حاکم بر نظام‌های ارتباطی درمی‌یابیم که صورت‌های تصریفی کلمه در واژگان وارد نمی‌شوند و اهل زبان در کنار واحدهای واژگانی قواعدی را نیز در ذهن خود دارند که مطابق آن قواعد صورت‌های مختلف تصریفی کلمه را تولید و درک می‌کنند. عملاً ذخیره‌کردن همهٔ صورت‌های کلمه در واژگان نه امکان‌پذیر است و نه ضرورت دارد. مثلاً " کتاب، کتابها، کتابی، کتاب‌هایی، کتابم ، کتابش، کتاب‌هایم و غیره" صورت‌های مختلف تصریفی واژهٔ "کتاب" هستند و با احتساب تمام تصریف‌ها " کتاب" بالقوه ۱۸۹ صورت تصریفی خواهد داشت . اهل زبان تنها صورت واژگانی " کتاب" را در ذهن خود دارند و صورت‌های تصریفی آن را به مراجعه به واژگان و مطابق قواعد تصریفی تولید و درک می‌کنند.

دیدیم که انسان در درک و تولید گفتار نیازمند واژگان است. در پردازش ماشینی زبان نیز اعم از تبدیل رایانه‌ای متن به گفتار، بازشناسی گفتار، ترجمهٔ ماشینی، تحلیل متن و غیره ناگزیر از داشتن یک واژگان هستیم. به هر میزان واژگان ما دقیق‌تر طراحی شده باشد کارایی آن نیز بیشتر خواهد بود. در طراحی واژگان زایای زبان‌فارسی سعی کرده‌ایم که از یک طرف واژگان ذهنی فارسی زبانان را مدل‌سازی کنیم (اسلامی و بی‌جن خان ۱۳۷۹)، و از طرف دیگر با صورت‌بندی قواعد تصریف کلمه در زبان‌فارسی واژگان زایای فارسی را طراحی کنیم که نوعاً واژگان ذهنی فارسی زبان‌ها را نمایندگی می‌کند. واژگان زایای زبان فارسی حدود ۵۵ هزار مدخل واژگانی دارد و این تعداد واحد واژگانی در چارچوب قواعد تصریف کلمه می‌توانند صورت‌های تصریفی متفاوت داشته باشند. برای عملیاتی‌کردن واژگان زایا برنامهٔ رایانه‌ای تهیه شده است که این برنامهٔ رایانه‌ای با ارجاع به واژگان و نیز قواعد تصریف کلمه در زبان فارسی می‌تواند واحدهای زبانی (نوشتار یا گفتار) را به لحاظ صرفی پردازش کند و خوانش صحیح صرفی به زنجیرهٔ ورودی برنامه اختصاص دهد. اصطلاحاً این برنامه رایانه‌ای را برنامهٔ واحدساز (tokenizer) صرفی نامگذاری کرده‌ایم. واژگان زایا حاوی چند نوع اطلاع زبانی و غیرزبانی برای هر مدخل است؛ مانند: صورت املایی مدخل، صورت واجی مدخل، مقولهٔ واژگانی مدخل، الگوی تکیهٔ مدخل، بسامد مدخل در یک پیکرهٔ زبانی و غیره. البته طراحی واژگان به گونه‌ای صورت گرفته است که امکان تغییر و یا افزایش اطلاعات دیگر امکان‌پذیر است.

---

* ارائه شده در اولین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۳

۱. مشخصات واژگان زایا

در مسیر اجرای طرح تبدیل رایانه‌ای متن به گفتار (Text – to – Speech: TTS) در زبان فارسی متوجه شدیم که بدون وجود فهرستی از کلمات زبان در قالب یک دادگان (database)، امکان تبدیل متن به گفتار وجود ندارد. پیرو این احساس نیاز بر آن شدیم تا واژگانی برای برنامهٔ بازسازی گفتار فارسی تهیه کنیم. در این راستا کوشیدیم تا نوعی الگوبرداری از واژگان ذهنی اهل زبان بکنیم. برای این منظور یک پیکرهٔ متنی ده میلیون کلمه‌ای را ملاک استخراج واژه‌های واژگان خود قرار دادیم. در برآورد اولیه متوجه شدیم که پیکرهٔ مورد نظر از حدود یکصد هزار کلمه با بسامدهای متفاوت تشکیل شده است. بعد از حذف صورت‌های تصریفی از فهرست فوق، حدود ۴۴ هزار واژه به مفهوم علمی آن به‌دست آمد. در بررسی فهرست واژه‌های به‌دست‌آمده از پیکرهٔ متنی متوجه شدیم که برخی واژه‌های عامیانه و برخی واژه‌های کاملاً علمی در فهرست ۴۴ هزاری مدخلی ما غایب هستند. برای رفع این کاستی، در ادامه، فهرست فوق را با *فرهنگ فارسی امروز* (صدری‌افشار ۱۳۸۱) مقایسه کردیم و با این فعالیت، حدود ۱۱ هزار مدخل جدید به فهرست واژه‌ها اضافه شد و واژگان ۵۵ هزار مدخلی به‌دست آمد. در واژگان زایای زبان فارسی هر مدخل با اطلاعات مربوط به صورت نوشتاری واژه در خط فارسی، ساخت واجی، مقولهٔ واژگانی، الگوی تکیه، بسامد واژه در پیکرهٔ متنی ده میلیون کلمه‌ای (جهت حل بخشی از مسئله هم نویسه‌ها) مشخص شده است. علاوه بر آن واژگان به گونه‌ای طراحی شده است که امکان تغییر و یا افزودن اطلاعاتی به واژگان جهت استفاده خاص از آن وجود دارد.

انواع کلمه یا به عبارت علمی‌تر مقوله‌های واژگانی مورد استفاده در واژگان زایا ۳۳ مقوله است که فهرست آنها به همراه کدهای ویژه برای هر کدام در جدول زیر آمده است.

شماره	مقوله‌های واژگانی	کد
۱	اسم	N1
۲	اسم خاص مکان	N2
۳	اسم خاص اشخاص	N3
۴	اسم فامیل	N4
۵	ضمیر فاعلی	N5
۶	ضمیر اشاره	N6
۷	ضمیر تأکیدی / انعکاسی	N7
۸	ضمیر مشترک / متقابل	N8
۹	اسم موجود در ساختمان حرف اضافه و حرف ربط گروهی	N9
۱۰	صفت پرسشی	A3
۱۱	صفت مبهم	A2
۱۲	شاخص	Spec
۱۳	صفت	A
۱۴	صفت اشاره	A1
۱۵	عدد اصلی	Nu
۱۶	عدد واحد	Nu1
۱۷	بن مضارع	V1

۱۸	بن ماضی	V2
۱۹	فعل اسنادی	V3
۲۰	فعل وجه‌نما	V4
۲۱	فعل کمکی	V5
۲۲	پیشوند فعل	Vpr
۲۳	حرف اضافه	Pr
۲۴	حرف اضافه پسین	Po
۲۵	حرف ربط	C
۲۶	حرف ربط گروهی	C1
۲۷	عبارت	Exp
۲۸	صوت	Intj
۲۹	علامت	sign
۳۰	قید	Adv
۳۱	حرف اضافهٔ گروهی	Pr1
۳۲	حروف الفبا	Al
۳۳	اختصارات	Ab

۲. تصریف کلمه در زبان فارسی

پیش‌تر گفتیم که صورت‌های تصریفی کلمه را اهل زبان از روی قواعد صرفی خاص تولید و درک می‌کنند و ذخیرهٔ همهٔ صورت‌های تصریفی برای یک کلمه در ذهن نه لازم است و نه امکان‌پذیر. اهل زبان اگر مجبور می‌شدند همهٔ صورت‌های تصریفی یک کلمه را به همراه تمام اطلاعات واژگانی مربوط به آن در ذهن خود حفظ کنند، قطعاً بار سنگین بر حافظهٔ آنها وارد می‌شد، چه بسا که امکان ارتباط زبانی از بین می‌رفت. اصل اقتصاد در زبان به اهل زبان این امکان را می‌دهد که تنها قواعد ساخت واحدهای زبانی باقاعده (صرفی یا نحوی) را فراگیرند و آن واحدها را از روی قاعده تولید و درک کنند. اصولاً این اصل در همهٔ نظام‌های ارتباطی حاکم است که با صرف کمترین هزینه ارتباط برقرار شود.

با تکیه بر استدلال فوق در تهیهٔ واژگان زایا زبان فارسی، ساخت تصریفی انواع کلمه را به شکل زیر صورت‌بندی کردیم: (اسلامی ۱۳۸۰)

۱.۲ اسم

اسم یک مقولهٔ واژگانی است که وندهای تصریفی آن تنها شکل پسوندی دارند. به عبارت دیگر وندهای تصریفی تنها به آخر اسم اضافه می‌شوند. ساخت تصریفی اسم را به شکل زیر می‌توان صورت‌بندی کرد:

[اسم] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) (تکواژ بند موصولی) (واژه‌بست‌های شخصی) (کسرهٔ اضافه)] + [(واژه بست‌های ربطی)]

مثال: کتاب‌ها، کتابی، کتابهایی، کتاب‌هایی (که)، برادرهایشانند،کتاب‌های (او).

در توضیح ساخت تصریفی اسم (و سایر مقوله‌ها) باید خاطرنشان شویم که هر کدام از وندهای تصریفی جایگاه معینی نسبت به ستاک دارند و اگر اسم فاقد یکی از وندهای تصریفی باشد، جایگاه آن خالی می‌ماند. از طرف دیگر وندهای تصریفی همه اختیارپند و به همین دلیل در داخل ( ) آمده‌اند. همچنین وندهای تصریفی که در یک جایگاه آمده‌اند ، در توزیع تکمیلی (complementary distribution) هستند. به این معنا که حضور یکی از آن وندها در آن جایگاه مانع حضور بقیه وندهای آن جایگاه می‌شود. مثلاً در "کتابم" واژه "کتاب" با واژه‌بست شخصی همراه شده است و از آنجائیکه جایگاه واژه‌بست‌های شخصی و یای نکره یکسان است، لذا نمی‌توانیم یای نکره را به "کتابم" اضافه کنیم یا بالعکس نمی‌توانیم /am-/ را به "کتابی" اضافه کنیم.

۲ . ۲ صفت

ساخت تصریفی صفت در زبان فارسی ساده است. و صفت‌های مدرج مطلق با گرفتن پسوندی "تر و ترین" تصریف می‌شوند. البته این کل ماجرا نیست. در زبان فارسی صفت‌ها اعم از مطلق، تفضیلی و عالی با یک اشتقاق صفر (zero derivation) به راحتی می‌توانند مقولهٔ اسم پیدا کنند و تصریف‌های اسم را بپذیرند. اشتقاق صفر ناظر بر نوعی فرایند واژه‌سازی است که در آن بدون افزوده شدن وند اشتقاقی به یک کلمه، مقولهٔ واژگانی کلمه تغییر می‌کند و واژه جدیدی ساخته می‌شود. مثلاً "دانشمند و بزرگتر " در ترکیب‌های " آدم دانشمند" و "برادر بزرگتر" صفت هستند، ولی در ترکیب‌های "دانشمندان ایرانی" و "بزرگترهای مجلس" اسم هستند و تصریف‌های اسم را گرفته‌اند. بنابراین در واژگان زایا، صفت‌های دارای ساخت تصریفی زیر هستند:

صفت + [(تکواژ صفت تفضیلی‌ساز) / (تکواژ صفت عالی‌ساز)] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) / (تکواژ بند موصولی) / (واژه‌بست‌های شخصی) / (کسرهٔ اضافه)] + [(واژه‌بست‌های ربطی)]

مثال: زرنگتر، زرنگترین، زرنگترها، زرنگی، زرنگی (که)، درسخوان ترهایشان، زرنگند.

۲ . ۳ قید

قید دارای ساخت تصریفی ساده زیر است :

قید + [(تکواژ قید تفضیلی‌ساز)]

مثال : او سریعتر می راند.

۲ . ۴ حرف اضافه

حروف اضافه واحدهای واژگانی هستند که گاه وند تصریفی می‌گیرند. ساخت تصریف حروف اضافه به شکل زیر است:

[(واژه‌بست‌های شخصی)] + حرف اضافه

مثال : ازش خواستم بیاید.

۲ . ۵ عدد

اعداد اصلی که در واژگان وارد شده‌اند و با گرفتن تکواژهای عدد ترتیبی‌ساز می‌توانند در صورت‌های تصریفی مختلف بکار روند. نقش عددها در زبان فارسی شبیه به صفت‌هاست. به همین دلیل می‌توانند همانند صفت، به اسم تبدیل شوند و ساخت تصریفی اسم را بپذیرند. با توجه به این واقعیت ساخت تصریفی عددها به شکل زیر است.

عدد اصل (+ [(تکواژهای عدد ترتیبی‌ساز)] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) (تکواژ بند موصولی) (واژه‌بست‌های شخصی) (کسرهٔ اضافه)] + [(واژه‌بست‌های ربطی)]

مثال : اولین، اولی، سومی، سومی‌ها، اولین‌ها، (شاگرد) چهارم‌های (مدرسه)، چهارمی ، چهارمه ، چهاری (که)،

۲ ۰ ۶ ضمیر

برای ضمیر به دلیل تفاوت در رفتار انواع زیر قابل تصور است:

ضمیر ← جانشین ← فاعلی / اشاره‌ای / تاکیدی
ضمیر ← پرسشی
ضمیر ← نامعین
ضمیر ← متقابل
ضمیر ← انعکاسی
ضمیر ← مشترک

[ضمیر] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) (تکواژ بند موصولی) (واژه‌بست‌های شخصی) (کسرهٔ اضافه)] + [(واژه‌بست‌های ربطی)]

۲ . ۷ فعل

فعل برخلاف بقیه انواع کلمات در زبان فارسی می‌تواند با گرفتن پیشوند نیز تصریف شود. ساخت تصریف فعل را در زیر می‌بینیم.

[(تکواژ نفی)] + [(تکواژ وجه امری و التزامی) (تکواژ استمراری)] + [فعل] + [(تکواژ ماضی نقلی‌ساز)] + [(شناسه‌های شخصی)] + [(واژه‌بست‌های شخصی) (واژه‌بست‌های فاعلی)]

۸.۲ مصدر

در واژگانی که برای بازسازی گفتار در نظر گرفته شده است، مصدرها مدخل های جداگانه و مستقلی ندارند بلکه قاعده ساخت مصدر در برنامه گنجانده می شود. این قاعده را به شکل زیر می توان نشان داد:

مصدر = بن ماضی + تکواژ اشتقاقی/an-/

حاصل اشتقاق بالا مصدر است. مصدرها از نظر مقوله اسمند و تمام تصریف های اسم را می‌پذیرند. اما تفاوت مصدرها با اسم‌ها در این است که مصدرها در عین حال برخی از تصریف های فعلی را نیز می‌پذیرند. همین مسئله باعث شده است که برای مصدرها ساختمان تصریفی جداگانه‌ای در نظر بگیریم که از یک نظر شبیه اسم و از نظر دیگر شبیه فعل می باشد. در واحدسازی باید این ویژگی تصریفی مصدرها مدنظر باشد و واحدساز در ارائه خوانش صحیح به مواردی مانند " بودنش، نبودنش، آمدن‌ها، نیامدنت، آمدنت و ... " دچار مشکل نشود. ساختمان تصریفی مصدر به شکل زیر است:

[(تکواژ نفی)] + [مصدر] + [(تکواژ جمع)] + [(تکواژ نکره‌ساز) / (تکواژ بند موصولی) / (واژه‌بست‌های شخصی) / (کسرهٔ اضافه)] + [(واژه‌بست‌های ربطی)]

۳. برنامهٔ واحدساز صرفی

به منظور عملیاتی کردن واژگان زایای فارسی به عنوان یک محصول مستقل جهت استفاده در جنبه‌های مختلف پردازش زبان، برنامهٔ رایانه‌ای تهیه کرده‌ایم که می‌تواند با مراجعه به واژگان و قواعد تصریف کلمه خوانش صحیح به صورت تصریفی کلمات اختصاص دهد. این برنامه مثلاً در تبدیل متن به گفتار، متن ورودی برنامه را به واحدهایی تقسیم می‌کند. سپس آن واحدها را با مراجعه به واژگان و ساخت تصریفی کلمات به لحاظ صرفی پردازش می‌کند . مثلاً "نمی‌رفتند" که یک واحد خارج از واژگان است. در چنین مواقعی از "نمی" عبور می‌کند و اگر بعد از آن بن فعل پیدا کرد، آنگاه مطابق ساخت تصریفی فعل، "نمی" را در ابتدا و "ند" در انتها مربوط فعل می‌داند و هر کدام از اجزای تشکیل دهندهٔ آن کلمه را جداگانه با عنوان خاص در خروجی مشخص می‌کند. این برنامه قابلیت‌هایی دارد که صورت لزوم برخی ایرادهای چاپی را اصلاح و سپس شروع به کار می‌کند. مثلاً وجود فاصله بین اجزای کلمه و غیره. همچنین قواعد اشتقاق کلمه (اسلامی ۱۳۸۱) در این برنامه گنجانده شده است تا از رهگذر آن کلمات مشتق تازه ساخته شده که در واژگان نیامده‌اند، خوانش صحیح پیدا کنند.

منابع

اسلامی، محرم و محمود بی‌جن‌خان. ۱۳۷۹. « یک انگارهٔ شناختی از واژگان ذهنی گویشور فارسی »، اولین کنفرانس بین‌المللی علوم شناختی، دانشگاه تهران و مؤسسهٔ علوم‌شناختی ایران، تهران، ایران.

اسلامی، محرم. ۱۳۸۰ . « ساخت تصریفی کلمه در زبان فارسی »، پژوهشکدهٔ پردازش هوشمند علائم، تهران، ایران.

اسلامی، محرم. ۱۳۸۱ . « ساخت اشتقاقی واژه و بازسازی گفتار »، پژوهشکدهٔ پردازش هوشمند علائم.

اسلامی، محرم. ۱۳۸۱ . « دشواری‌های پردازش رایانه‌ای خط فارسی »، *نشر دانش*، سال نوزدهم، شمارهٔ سوم، پاییز ۱۳۸۱، مرکز نشر دانشگاهی، تهران، ایران.

صدری‌افشار، غلامحسین و همکاران. ۱۳۸۱ . *فرهنگ معاصر فارسی امروز*. فرهنگ معاصر، تهران، ایران.

Bloomfield , Leonard . 1933 . *Language* . New York : Holt , Rinehart , Winston.

Chomsky, Noam.1965. *Aspects of the theory of syntax*.. Cambridge. MA: MIT Press.