

مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات

فصلنامه اطلاع‌رسانی. دوره ۱۷، شماره ۲

نوشته: لیلا مرتضائی

عضو هیأت علمی مرکز اطلاعات و مدارک علمی ایران

چکیده

بانک‌های اطلاعاتی فارسی، پیش از آن که فرهنگستان زبان معیارهای لازم را برای کاربرد اصطلاحات علمی و رسم‌الخط فارسی تعیین کند شکل گرفتند. مجریان بانک‌های اطلاعاتی و نمایه‌سازان، خواسته یا ناخواسته - با مسائل واژه‌گزینی و جنبه‌هایی از زبانشناسی درگیر شدند. در کار واژه‌گزینی، اطلاع‌رسانان - به لحاظ ماهیت حرفه‌خود - واژه‌های رایج در جامعه تولیدکنندگان و استفاده‌کنندگان از اطلاعات را مد نظر دارند و خود را مجاز به واژه‌سازی و اعمال سلیقه نمی‌دانند. واژه‌های تازه ساخت نیز تازمانی که در جامعه مقبولیت لازم را به دست نیاورده باشند و در مدارک به کرات دیده نشوند، در نظام‌های ذخیره و بازیابی اطلاعات یا ظاهر نمی‌شوند و یا میهمان چندروزه‌اند. بخش قابل توجهی از مشکلات نمایه‌سازان از رواج و کاربرد واژه ناشی می‌شود. متخصصان برای یک مفهوم واحد اصطلاحات متفاوت به کار می‌برند. حتی متخصصانی که در یک رشته و در یک جامعه کوچک کار می‌کنند خود را ملزم به هماهنگی در کاربرد واژه‌های تخصصی نمی‌بینند. به علاوه برای بسیاری از اصطلاح‌های وارداتی معادل‌های متفاوت در زبان فارسی وجود دارد که در مواردی همه، کم و بیش، به یک اندازه کاربرد دارند. این گونه مطالب به علاوه مسائل رسم‌الخط فارسی، آوانویسی اسامی عناصر و ترکیبات شیمیایی، سرواژه‌ها و کوتاه‌نوشته‌ها سبب شده است تا ذخیره اطلاعات به زبان فارسی با کندی صورت گیرد و جستجو و بازیابی کارایی مطلوب را نداشته باشد. در این مقاله سعی خواهد شد نمونه‌هایی از تجربه‌های واژه‌گزینی در ذخیره اطلاعات ارائه شود تا با توجه به آنها، شاهد تسریع و تسهیل ذخیره و بازیابی اطلاعات به زبان فارسی باشیم.

کلیدواژه‌ها: بانک‌های اطلاعاتی / زبان فارسی / رسم‌الخط فارسی / ذخیره و بازیابی اطلاعات

کتابداران از مدتها پیش دریافتند که رابطه‌ای مستقیم و الزامی بین تحلیل موضوعی مطالب و زبانشناسی وجود دارد. این رابطه با پیدایش علم اطلاع‌رسانی و کاربرد رایانه در این رشته، شتاب و اهمیت بیشتری یافت. امروزه اطلاع‌رسانی و اصطلاح‌شناسی شاخه‌هایی از علوم را تشکیل می‌دهند که ارتباطی نزدیک و مداوم بین آن‌ها برقرار است. دانش اطلاع‌رسانی در حوزه فعالیت خود عمدتاً با اطلاعات نوشتاری، که زبان وسیله اصلی انتقال آن است، سروکار دارد. متخصصان در مناسبات علمی از زبان‌های ویژه استفاده می‌کنند. هسته اصلی هر زبان ویژه، اصطلاحات علمی یا واژگان واژگان آن است. این اصطلاحات برای ارتباط علمی و انتقال صحیح اطلاعات به کار گرفته می‌شود و چنانچه دچار هرج و مرج و نابسامانی شود، طبعاً زبان تفهیم و تفاهم و جریان درست اطلاعات مختل می‌شود. کتابداران و اطلاع‌رسانان که رابط بین تولیدکنندگان و مصرف‌کنندگان اطلاعات هستند، پیش از سایر متخصصان ضرورت استاندارد کردن واژگان علوم را دریافتند و همزمان با توسعه بانک‌های اطلاعاتی به رعایت آن اصرار ورزیدند.

در کشورهای غربی بیش از ۳۰ سال است که رایانه‌ها در ذخیره و بازیابی اطلاعات و ایجاد پایگاه‌های اطلاعات کتابشناختی به کار گرفته شده‌اند. در این مدت اصطلاحنامه‌های تخصصی متعددی که بعضی از آنها شهرت و اعتبار دارند، با همکاری متخصصان موضوعی و زبانشناسان تدوین گردیدند. رایانه‌ها از جهت ساخت‌افزایی و نرم‌افزاری پیشرفت

کردند و ما امروزه شاهد استفاده از سیستم‌های هوشمند در ذخیره و بازیابی اطلاعات هستیم. در نظام‌های اطلاعاتی، واژه‌ها به دو گونه "زبان آزاد" و "زبان مهارشده" ظاهر می‌گردند. در استفاده از زبان آزاد، اصطلاح‌های تخصصی بدون آن که هیچ گونه کنترلی در مورد آن‌ها اعمال شود، به همان صورت که در مدارک آمده‌اند، در سیستم وارد و ذخیره می‌شوند. مسوولیت توجه به کلیه اصطلاحات معادل و شکل‌های متفاوت نوشتاری یک اصطلاح هنگام بازیابی اطلاعات به عهده کاربر است. حسن این روش کاهش زمان آماده‌سازی و پردازش اطلاعات و تقلیل نیاز به نیروی متخصص است. انواع نمایه‌های گردان (۱) که براساس چرخش عنوان‌های مدارک و الفبایی کردن هر یک از واژه‌های عنوان طراحی شده‌اند، از این نوع اند.

در استفاده از زبان مهار شده، نظام‌های ذخیره و بازیابی اطلاعات با بهره‌گیری از متخصصان موضوعی و صرف وقت و نیروی قابل ملاحظه‌ای می‌کوشند تا اطلاعات را در قالب زبانی مهار شده و مقید ذخیره کنند تا کاربر و رابط اطلاعات (۲) در زمانی کوتاه‌تر و با نیروی کمتر، درصد بالایی از اطلاعات پایگاه را بازیابی نمایند. اصطلاحنامه‌های تخصصی (۳) ابزار کار این روش‌اند. نظام‌های بازیابی تمام متن (۴) و آزاد متن (۵) که با رواج استفاده از شبکه جهانی وب شناخته شدند، از زبان آزاد و زبان مهار شده، تماماً در سیستم‌های هوشمند سود می‌جویند.

در ایران این امور سابقه چندانی ندارد. از عمر پایگاه‌های اطلاعاتی فارسی بیش از ۱۰ سال نمی‌گذرد. پیش از آن که فرهنگستان زبان معیارهای لازم را برای کاربرد اصطلاحات علمی و رسم‌الخط فارسی تعیین کند، پایگاه‌های اطلاعاتی شکل گرفتند و مجریان آن‌ها و نمایه‌سازان، خواسته یا ناخواسته، با مسائل واژه‌گزینی و جنبه‌هایی از زبانشناسی درگیر شدند. البته در این مدت به مدد رسانه‌های ارتباطی و بهره‌گیری از دانش و تجربه کشورهای پیشرفته، بسیاری از مراحل را شتابان پیمودیم و شاید بسیاری از مشکلات را چون دیگر مشکل به حساب نمی‌آمدند، حس نکردیم. ولی بتدریج که بر حجم اطلاعات فارسی افزوده شد، دشواری‌های خط و زبان فارسی خودنمایی کرد، از محاسن روش‌های ذخیره و بازیابی بازیابی کاست و بر معایب آنها افزود، حل آن‌ها روزبه‌روز مشکل‌تر شد و اعمال بعضی روش‌های ماشینی ممکن نگردید. کتابداران و اطلاع‌رسانان که به لحاظ ماهیت حرفه خود با واژه‌های رایج در جامعه، تولیدکنندگان و مصرف‌کنندگان اطلاعات سر و کار دارند، خود را مجاز به اعمال سلیقه نمی‌دانند. خط و زبان هم مطلبی نیست که بتوان با اجرای الگوهای غربی بر مسایل آن فایق آمد. تنها با استمداد از نهادهای مسوول و یاری آنها می‌توانند به رفع، و یا حداقل مهار این مشکلات بپردازند. نویسنده مقاله به عنوان یکی از کسانی که با مسائل پایگاه‌های اطلاعاتی مدارک فارسی سر و کار داشته و آن‌ها را تجربه کرده است، تلاش خواهد کرد در حد حوصله و وقت مقاله نمونه‌هایی ارائه دهد تا مشخص گردد دشواری‌های زبان و خط فارسی چگونه سبب می‌شود اولاً - در نظام‌هایی که براساس زبان آزاد طراحی شده‌اند به دلیل تعدد اصطلاح‌های معادل و پراکندگی آن‌ها در محل‌های الفبایی مختلف، مشخص نبودن حد کلمه در واژه‌های مرکب و استاندارد نبودن شکل نوشتاری کلمات، نتیجه جستجو جامعیت مطلوب را نداشته باشد. ثانیاً در نظام‌هایی که از زبان مهارشده بهره می‌گیرند به دلیل همین مسایل، نیرو و زمانی بیش از آنچه تصور می‌رود برای واژه‌گزینی و معادل‌یابی، هماهنگی و یکسان‌سازی شکل نوشتاری اصطلاحات صرف شود.

۱ - گوناگونی معادل‌های علمی

متخصصان در بیان و انتقال یک مفهوم از اصطلاحات متفاوت استفاده می‌کنند. نظری اجمالی به یکی دو واژه‌نامه، تخصصی که براساس کاربرد اصطلاحات در منابع تهیه شده‌اند نشان می‌دهد که بازار واژه‌سازی و به قول یکی از زبان‌شناسان واژه‌بازی رواج دارد. به عنوان نمونه واژگان کتابداری و اطلاع‌رسانی (۶) نشان می‌دهد متخصصان این رشته، ۶ معادل برای 9 Manual معادل برای 12 Online معادل برای Layout و ۱۳ معادل برای Cross refrence بکار برده‌اند. از این گونه نمونه‌ها در تمام رشته‌ها فراوان است که مواردی از آنها در ضمیمه مقاله آورده می‌شود. متأسفانه فرهنگستان زبان هم با تصویب برخی معادل‌های نامأنوس (مانند پروتجا به جای فایل؛ پروندان به

جای زونکن) سهمی در این بازار آشفته دارد. حال آنکه در کشورهای پیشرفته، اصل اصطلاح - چه خوش ساخت و چه بدساخت - پذیرفته می‌شود و بدون اعمال سلیقه، به همان صورت، به کار می‌رود. به راستی نمایه ساز باید کدام یک از اصطلاحات معادل را اصل قرار دهد و از بقیه به آن ارجاع بسازد؟ چه معیاری در دست دارد؟ رابط اطلاعات، یا به قول فرهنگستان کارور، اگر بخواهد با استفاده از منطق بول (۷) بین دو یا سه اصطلاح رابطه منطقی برقرار کند چه تدبیری باید ببیندیشد تا حداکثر نسبت بازیافت (۸) را داشته باشد؟

۲ - ضبط اسامی

در برگردان اسامی افراد، سازمان‌ها، عناصر و ترکیبات شیمیایی، ابزار و تجهیزات، محل‌های جغرافیایی و مانند آن‌ها از زبان‌های بیگانه به فارسی، قاعده خاصی وجود ندارد. هر متخصص، نویسنده و مترجمی بنا به ذوق و سلیقه، میزان آشنایی با زبان مبدأ و دانش و تخصص خود، آن‌ها را به فارسی برگردانده و در متون بکار برده است. این نابسامانی حتی در انتشارات سازمان‌های علمی و فرهنگی کشور نیز دیده می‌شود.

نمونه:

پستالزی، ژوهان هنریش / پستالوزی، یوهان هنریش / پستالوزی، ژان هانری

فلیشیا / فلیسیا / فلیشا / فلیسا

راینسون / روبینسون / رینسون / روبنسن

پیرسون / پی‌یرسون / پی‌یرسن

اف. آی. دی / فید

دبلیو. اچ. او / هو

پتاسیم / پتاسیوم / پوتاسیوم / پوتاسیم

کادمیوم / کادمیم / کادیوم

ئیدروژن / هیدروژن

آلزامر / الزایمر

آفریقا / افریقا

آمریکا / امریکا

آیا می‌توان تمام شکل‌های حرف نویسی و آوانویسی اسامی را پوشش داد و پایگاه اطلاعاتی را با ارجاعات متعدد انباشت؟

۳ - سرهم‌نویسی، جدانویسی، و بی‌فاصله‌نویسی

شیوه خط فارسی چنان است که بسیاری از واژه‌ها را می‌توان به چند صورت نوشت. این چندگونگی شکل واژه‌ها، برای رایانه قابل درک نیست. چرا که واژه‌ها را تنها به همان صورتی که ذخیره کرده است می‌شناسد و بازیابی می‌کند. لذا در مقابل سایر شکل‌های نوشتاری یک اصطلاح ناآگاه است و در هنگام جستجوی اطلاعات پاسخ منفی می‌دهد. رابط‌های اطلاعات برای پرهیز از این مشکل، عموماً از فهرست کلید واژه‌ها استفاده می‌کنند که این امر سبب شده است تا از امکانات منطق بول در بازیابی اطلاعات به خوبی بهره گرفته نشود. در مواقعی که بازیابی از محدوده فیلد کلید واژه‌ها، که اصطلاحات مهارشده‌اند، فراتر می‌رود و فیلدهای عنوان، پدیدآورنده و ناشر را شامل می‌شود، این ناهماهنگی کاملاً به چشم می‌خورد. گاه یک واژه مرکب براساس شکل نگارش آن در چند محل الفبایی مختلف، جدا از هم قرار می‌گیرد. علامت جمع "ها" که به صورت سرهم یا جدا نوشته شود نیز، همین وضع را در فهرست‌های رایانه‌ای ایجاد می‌کند.

نمونه:

آب‌بند / آب‌بند
آب‌شش / آب‌شش
آب‌کاری / آب‌کاری
آب‌گرم‌کن / آب‌گرم‌کن / آب‌گرم‌کن / آب‌گرم‌کن
بی‌خوابی / بی‌خوابی
بی‌حس‌کننده / بی‌حس‌کننده
بیماری‌زا / بیماری‌زا
دستگاه‌یخ‌ساز / دستگاه‌یخ‌ساز
دستگاه‌هم‌زن / دستگاه‌هم‌زن
ماشین‌ظرف‌شویی / ماشین‌ظرف‌شویی
یخ‌زدگی / یخ‌زدگی
یخ‌بندان / یخ‌بندان
یون‌ساز / یون‌ساز
خاک‌برداری / خاک‌برداری
نام‌گذاری / نام‌گذاری
خشک‌شویی / خشک‌شویی
غلام‌حسین / غلام‌حسین
علی‌رضا / علی‌رضا
حسن‌علی / حسن‌علی
کتاب‌درسی / کتاب‌درسی
برنامه‌درسی / برنامه‌درسی
نیروی‌انسانی / نیروی‌انسانی
دانشگاه‌ها / دانشگاه‌ها
کوه‌ها / کوه‌ها



۴ - انواع جمع

تعدد علائم جمع (ها؛ ان؛ ات؛ ین؛ ون) و وجود جمع بی‌قاعده در زبان فارسی سبب گردیده است در پایگاه‌هایی که کلید واژه‌ها را به صورت جمع به کار می‌برند، مشکلی بر مشکلات بالا افزوده شود. نمایه‌ساز در هنگام نمایه‌سازی در انتخاب بین مدارس / مدرسه‌ها، اساتید / استادان / استادها، محققان / محققین و مانند آن‌ها، مردد است. رابط اطلاعات در موقع بازیابی باید شکل‌های مختلف جمع کلیدواژه‌ها را در نظر داشته باشد و یا، با استفاده از علائم قراردادی، واژه را برش (۹) بزند. در هر دو صورت، بازهم احتمال پوشش ندادن بعضی از جمع‌های بی‌قاعده وجود دارد.

آن دسته از پایگاه‌های اطلاعاتی که کلید واژه‌ها را به صورت مفرد بکار می‌برند، باین مشکل مواجه نیستند. البته مسایل جزئی وجود دارد که به نوعی حل می‌کنند، از آن جمله اصطلاحاتی که در شکل جمع، مفهومی متمایز از شکل مفرد دارند (مانند تشکیلات، تجهیزات، امکانات، تسهیلات، ارتباطات) و یا برخی اصطلاحات که به صورت مفرد نام‌نوس‌اند (مثل گروه همسالان، فرصت‌های شغلی، خدمات مشاوره، اوقات فراغت). در این گونه موارد یا معادلی مناسب را جایگزین جایگزین می‌کنند (مانند سازمان به جای تشکیلات و یا رسانه همگانی به جای وسایل ارتباط جمعی) یا چنانچه ممکن

باشد شکل مفرد واژه را، به امید آن که پذیرفته شود برمی‌گزینند (مانند وقت فراغت به جای اوقات فراغت) در غیر این صورت همان شکل جمع را بکار می‌برند.

۵ - صورت‌های مختلف نوشتاری

همزه، الف مقصوره، تشدید و دوگانگی شکل نوشتاری واژه‌ها و اسامی، سبب ناهماهنگی‌هایی در ورود داده‌ها و پراکندگی پراکندگی اطلاعات پردازش شده می‌گردد.

نمونه:

هیأت مدیره / هیئت

مسأله اجتماعی / مسئله اجتماعی

مسئولیت والدین / مسئولیت والدین

عطایی / عطائی

رؤوف / رؤف

اسماعیل / اسمعیل

اسحاق / اسحق

آیینه / آینه

طومار / تومار

موّحدی / موحدی

داود / داوود

طاوس / طاووس

آیت‌اللهی / آیت الهی

لیلا / لیلی



حاصل سخن

زبان علم را "زبانی ارتباطی و اطلاعاتی، روشن و سراسر است و عاری از ابهام" (۱۰) تعریف کرده‌اند، آیا زبان فارسی در حیطه علم چنین ویژگی‌هایی دارد؟ آیا در نقش ارتباطی و اطلاعاتی خود موفق بوده است؟ تصور نمی‌کنم پاسخ چنین پرسش‌هایی مثبت باشد. طبعاً پایگاه‌های اطلاعاتی که با استفاده از این زبان به ذخیره و بازیابی اطلاعات می‌پردازند، نمی‌توانند کارایی مطلوب را داشته باشند. عواملی که پیش‌تر بدان اشاره شد سبب کندی مراحل ذخیره و بازیابی اطلاعات می‌شوند، نسبت بازیافت اطلاعات را کاهش می‌دهند و همواره می‌توان نسبت به جامعیت نتیجه یک جستجو شک کرد.

پایگاه‌های اطلاعاتی مدارک فارسی با وجود عمر کوتاه‌شان با مشکلات بسیاری دست‌بگیرانند که اگر هر چه زودتر چاره اندیشی نشود، با توجه به هجوم اطلاعات، دیگر مهار آن‌ها آسان نخواهد بود. در این زمینه پیشنهاد می‌شود:

- فرهنگستان برای یکسان‌سازی واژه‌های علمی و جلوگیری از ناهماهنگی بیشتر گام‌های سریع‌تر و مؤثرتری بردارد. همزمان با ظهور و ورود هر پدیده و یا فرآورده علمی، پیش از آن که معادل‌های گوناگون رواج یابند، اصطلاح مناسب را انتخاب و اعلام نماید.

- دستورالعمل‌هایی در مورد شیوه نگارش اصطلاحات و واژه‌های فارسی که مورد تأیید اهل فن باشد، تدوین و برای اجرا به کلیه واحدهای چاپ و نشر ابلاغ شود.

- برای تدوین اصطلاحنامه‌های تخصصی در زبان فارسی، که حاوی اصطلاحات معیار در هر رشته و شیوه نوشتاری مورد قبول باشد، اقداماتی مؤثر، هماهنگ و حساب‌شده از طرف سازمان‌های ذیربط صورت گیرد.
- متخصصان رایانه در جهت استفاده از امکانات این پدیده قرن و هوشمندکردن سیستم‌ها برای پردازش خط فارسی، چاره‌جویی و هم‌اندیشی بیشتری داشته باشند.

پی‌نوشت‌ها:

1. Permuted Index
2. Information intermediary
3. Thesaurus
4. Full-Text
5. Free-Text

۶. هاشمی، ابوالفضل (۱۳۷۶). واژگان کتابداری و اطلاع‌رسانی. تهران، دبیرخانه هیأت‌امنا کتابخانه‌های کشور.

7. Boolean logic
8. Recall ratio
9. Truncation

۱۰. حق شناس، علی محمد (۱۳۷۲). در جست و جوی زبان علم. مجموعه مقالات سمینار زبان فارسی در زبان علم. تهران: مرکز نشر دانشگاهی. ص ۱۳-۶.

منابع:

- آشوری، داریوش (۱۳۷۵). بازاندیشی زبان فارسی؛ ده مقاله، ویرایش دوم. تهران: نشر مرکز.
اکبری نژاد، سعید (۱۳۷۶). فاصله خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات. فصلنامه کتاب. (بهار و تابستان). ص ۴۹-۵۶.
امامی، کریم (۱۳۷۱). لزوم بازنگری در شیوه نگارش خط فارسی. آدینه. ۷۳/۷۴ (شهریور) ص ۱۹-۱۸.
باطنی، رضا (۱۳۷۱). نگاهی تازه به شیوه خط فارسی. آدینه. ۷۵ (آبان). ص ۴۴-۴۵.
بهزادی، ماندانا (۱۳۷۵). شیوه ضبط اعلام انگلیسی در فارسی. تهران: مرکز نشر دانشگاهی؛ کتابخانه ملی جمهوری اسلامی ایران.
حری، عباس (۱۳۷۲). کامپیوتر و رسم‌الخط فارسی، پیام کتابخانه. سال سوم. شماره ۱. (بهار) ص ۱۱-۶.
حق شناس، علی محمد (۱۳۷۲). در جست و جوی زبان علم. مجموعه مقالات سمینار زبان فارسی در زبان علم. تهران: مرکز نشر دانشگاهی. ص ۱۳-۶.
صنعتی، محمد (۱۳۷۱). دشواری‌های زبان فارسی با کامپیوتر. آدینه. ۷۲ (مرداد). ص ۵۶-۵۷.
کابلی، ایرج (۱۳۷۱). فراخوان برای فارسی نویسی و پیشنهاد به تاجیکان. آدینه. ۷۲ (مرداد) ص ۵۵-۵۰.

مآخذ نمونه‌ها

- امینی، سید محمد (۱۳۷۰). واژگان فیزیک. تهران: مرکز نشر دانشگاهی.
- باقری، محمد (۱۳۷۲). واژگان ریاضی. تهران: نشر فرهنگان.
- بريجانيان، ماری (۱۳۷۱). فرهنگ اصطلاحات فلسفه و علوم اجتماعی، ویرایش بهاء‌الدین خرمشاهی. تهران: مؤسسه مطالعات و تحقیقات فرهنگی.
- پورجوادی، علی. واژگان شیمی و مهندسی شیمی، تهران: مرکز نشر دانشگاهی.
- هاشمی، سید محمد (۱۳۷۶). واژگان کتابداری و اطلاع‌رسانی. تهران: دبیرخانه هیأت امنای کتابخانه‌های عمومی کشور.
- همایون، همادخت (۱۳۷۱). واژه نامه زبانشناسی و علوم وابسته. تهران: مؤسسه مطالعات و تحقیقات فرهنگی.



پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی