مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات

فصلنامه اطلاع رسانی. دوره ۱۷، شماره ۱و 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 ـ با مسائل واژه گزینی و جنبه‌هایی از زبانشناسی درگیر شدند. در کار واژه گزینی، اطلاع‌رسانان ـ به لحاظ ماهیت حرفه‌خود ـ واژه‌های رایج در جامعه تولیدکنندگان و استفاده کنندگان از اطلاعات را مد نظردارند 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 متخصصانی‌که 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 دارد که در مواردی همه، کم و بیش، به یک اندازه کاربرد دارند. این گونه مطالب به علاوه مسائل رسم‌الخط فارسی، آوانویسی اسامی عناصر وترکیبات شیمیائی، سرواژه‌ها و کوته‌نوشته‌ها سبب شده است تا ذخیره اطلاعات به زبان‌فارسی با کندی صورت گیرد و جستجو و بازیابی کارایی مطلوب را نداشته باشد.

در این مقاله سعی خواهد شد نمونه هایی از تجربه‌های واژه گزینی 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 حوزهء 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 ء اصلی هر زبان ویژه، اصطلاحات علمی یا واژگان واژگان آن است. این‌اصطلاحات 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 مصرف کنندگان اطلاعات هستند، پیش از سایر متخصصان ضرورت استانداردکردن 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، باهمکاری متخصصان موضوعی و زبانشناسان تدوین گردیدند. رایانه‌ها از جهت‌سخت‌افزاری و نرم‌افزاری پیشرفت

کردند و ما امروزه شاهد استفاده از سیستم‌های هوشمند در ذخیره و بازیابی اطلاعات هستیم.
در نظام‌های اطلاعاتی، واژه‌ها به دوگونه "زبان آزاد" و "زبان مهارشده" ظاهر می‌گردند. در استفاده از زبان آزاد، اصطلاح‌های تخصصی بدون آن که هیچ گونه کنترلی درمورد 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 یک‌اصطلاح 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 وبا نیرویی 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 رواج‌استفاده از شبکه جهانی وب شناخته شدند، از زبان آزاد و زبان مهار شده، تؤاماً 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 برحجم 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 جامعهء 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 این‌مشکلات بپردازند. نویسندهء 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 درحد حوصله و وقت مقاله نمونه‌هایی ارائه دهد تا مشخص گردد دشواری‌های زبان و خط فارسی چگونه سبب می‌شود اولاً ـ 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 معادل یابی، هماهنگی و یکسان سازی شکل نوشتاری اصطلاحات صرف شود.

۱ ـ گوناگونی معادل‌های علمی

متخصصان در بیان و انتقال یک مفهوم از اصطلاحات متفاوت استفاده می‌کنند. نظری اجمالی به یکی دو واژه‌نامهء تخصصی که براساس کاربرد اصطلاحات در منابع تهیه شده‌اند نشان می‌دهد که بازار واژه‌سازی و به قول یکی از زبان‌شناسان واژهء بازی رواج دارد. به عنوان نمونه واژگان کتابداری و اطلاع‌رسانی(۶) نشان می‌دهد متخصصان این رشتهء ۶ معادل برای Manual 9 معادل برای Online 12 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 ضمیمهء مقاله آورده می‌شود. متأسفانه فرهنگستان زبان هم با تصویب برخی معادل‌های نامأنوس (مانند پروَنجا به جای فایل؛ پروندان به

جای زونکن) سهمی در این‌بـــــازار آشفته دارد. حال آنکه در کشورهای پیشرفته، اصل اصطلاح ـ چه خوش‌ساخت و چه بدساخت ـ پذیرفته می‌شود و بدون اعمال سلیقه، به همان صورت، به کار می‌رود.
به راستی نمایه ساز بـاید کدام یک از اصطلاحات معادل را اصل قـرار دهـد و از بقیه به‌آن ارجاع بسازد؟ چه معیاری در دست دارد؟ رابـط اطلاعات، یـا به قول فرهنگستان کـارور، اگر بخواهد با استفاده از منطق بول(۷) بین دو یا سه اصطلاح رابطه منطقی برقرار کند چه تدبیری باید بیاندیشد تا حداکثر نسبت بازیافت(۸) را داشته باشد؟

۲ ـ ضبط اسامی

در برگردان اسامی افـراد، سازمان‌هـا، عناصـر و ترکیبات شیمیایی، ابزار و تجهیزات، محل‌های جغرافیایی و مانند آن‌ها از زبان‌های بیگانه به فارسی، قاعده خاصی وجـــود ندارد. هـــر متخصص، نویسنده و مترجمی بنـــا به ذوق و سلیقه، میزان آشنایی بـا 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،ژوهان‌هنریش/پستالوزی،یوهان‌هانریش/پستالوزی، ژان هانری
فلیشیا/ فلیسیا/ فلیشا/ فلیسا
رابینسون/ روبینسون/ ربینسون/ روبنسن
پیرسون/ پی‌یرسون/ پی‌یرسن
اف. آی. دی/ فید
دبلیو. اچ. او/ هو
پتاسیم/ پتاسیوم/ پوتاسیوم/ پوتاسیم
کادمیوم/کادمیم/کادیوم
ئیدروژن/ هیدروژن
آلزایمر/ 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 ـ سرهم‌نویسی، جدانویسی،و بی‌فاصله نویسی

شیوه خـط فارسی چنان است که بسیاری از واژه‌هـا را می‌توان به چـند صورت نوشت. این چندگونگی شکل واژه‌ها، برای رایانه قابل درک نیست. چـرا که واژه‌هـا را تنها به همان صورتی که ذخیره کرده است می‌شناسد و بازیابی می‌کند. لذا در مقابل ســایر شکل‌هــای نوشتاری یک اصطلاح ناآگاه است و در هنگام جستجوی اطلاعات پاسخ منفی می‌دهد. رابط‌های اطلاعات برای پرهـیز از این مشکل، عمومـاً از فهرست کلـید واژه‌هـا استفاده‌می‌کنـند که این امـر سبب شده است تـا از امکانات منطق بول در بازیابی اطلاعات به خوبی بهره گرفته نشود. در مواقعی که بازیابی از محدوده فیلد کلید واژه‌ها، که اصطلاحات مهارشده‌انـد، فراتـر می‌رود و فیلدهای عنوان، پدیدآورنده و ناشـر را شامل می‌شـود، این ناهماهنگی کاملا به چشم می‌خورد. گاه یک واژه مرکب براساس شکل نگارش آن در چـــند محل الفبایی مختلف، جـــدا از هم قـــرار 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
دستگاه یخ‌ساز/ دستگاه یخساز
دستگاه هم‌زن/ دستگاه همزن
ماشین ظرف‌شویی/ ماشین ظرفشویی
یخ‌زدگی/ یخزدگی
یخ‌بندان/ یخبندان
یون‌ساز/ یونساز
خاک‌برداری/خاکبرداری
نام‌گذاری/نامگذاری
خشک‌شویی/خشکشویی
غلام حسین/غلامحسین
علی‌رضا/علیرضا
حسن‌علی/حسنعلی
کتاب‌درسی/کتاب درسی
برنامه‌درسی/برنامه درسی
نیروی‌انسانی/نیروی انسانی
دانشگاه‌ها/دانشگاهها
کوه‌ها/کوهها

۴ ـ انواع جمع

تعـدّد علائم جمع (ها؛ ان؛ ات؛ ین؛ ون) و وجود جمع بی‌قاعده در زبان فارسی سبب گردیده است در پایگاه‌هایی که کلید واژه‌هـا را به صورت جمع به کـار می‌بـرند، مشکلی بر مشکلات بالا افزوده شود. نمایه‌ساز در هنگام نمایه‌سازی در انتخاب بین مدارس / مدرسه‌هـا، اساتید/ استادان/ اسـتادها، محققان/ محققین و مـانند آن‌ها، مردد است. رابط اطلاعات در موقع بازیابی بـاید شکل‌های مختلف جمع کلیدواژه‌هـا را در نظر داشته باشد ویا، 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 کلـید واژه‌ها را به صورت مفرد بکار می‌برند، بااین مشکل مواجه نیستند. البته مسایل جزئی وجـود دارد که به نوعی حل می‌کنـند، از آن‌جمله اصطلاحاتی که در شکل جمع، مفهومی متمـایز از شکل مفرد دارنـد (مـانند تشکیلات، تجهیزات، امکانات، تسهیلات، ارتباطات) و یـا برخی اصطلاحات که به‌صورت مفـرد نامأنوس‌اند (مثل گروه همسالان، فرصت‌های شغلی، خدمات مشاوره، اوقات فراغت). در این گونه موارد یا معادلی مناسب را جایگزین جایگزین می‌کنـند (مـانند سازمان به جای تشکیلات و یا رسانه همگانی به جای وسایل ارتباط جمعی) یا چنانچه ممکن

باشـد شکل مفـرد 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 ـ 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
مسؤولیت والدین/ مسئولیت والدین
عطایی/ عطائی
رؤوف/ رئوف
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 سرراست و عاری از ابهام"(۱۰) تعریف کرده‌اند، آیا زبان فارسی در حیطه علم چنین ویژگی‌هایی دارد؟ آیـــا در نقـــش ارتباطی و اطلاعاتی خـــود موفق بوده است؟ تصـــور نمی‌کنم پاسخ چنین پرسش‌هایی مثبـت باشد. طبعـاً پایگاه‌های اطلاعاتی که بـا استفاده از این زبان به ذخیره و بازیابی اطلاعـات می‌پردازند، نمی‌توانـــند کارایی مطلوب را داشته باشند. عواملی که پیش‌تـــر بدان اشاره شـــد سبب کندی مراحل ذخیره و بازیابی اطلاعات می‌شـوند، نسبت بازیافت اطلاعات را کاهش می‌دهـند و همواره می‌توان نسبت به جامعیت نتیجه یک جستجو شک کرد.

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

ـ فرهنگستان برای یکسان‌سازی واژه‌های علمی و جلوگیری از ناهماهنگی بیشترگام‌های سریع‌تـــــر و مؤثرتری بردارد. همزمان بـا ظهـور و ورود هر پدیده و یا فرآورده علمی،پیش از آن که معادل‌های گوناگون رواج یابند، اصطلاح مناسب را انتخاب و اعلام نماید.

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

ـ برای تدوین اصطلاحنامه‌های تخصصی در زبان فارسی، که حاوی اصـطلاحات معیار در هر رشته و شیوه نوشتاری مورد قبول باشد، اقداماتی مؤثر، هماهنگ و حساب‌شده از طرف سازمان‌های ذیربط صورت گیرد.
ـ متخصصان رایانه در جهت استفاده از امکانات این پدیده قرن و هوشمندکردن سیستم‌هـــا برای پردازش خـــط 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 اطلاع رسانی. تهران، دبیرخانه هیأت‌امنای کتابخانه‌های کشور.

7. Boolean logic
8. Recall ratio
9. Truncation

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

منابع:

آشوری، داریوش (۱۳۷۵). بازاندیشی زبان فارسی؛ ده مقاله، ویرایش دوم. تهران: نشرمرکز.

اکبری نـــژاد، ســـعید (۱۳۷۶). فاصله خالی میان واژه‌هـــا در ذخیره و بازیابی رایانه‌ای‌اطلاعات. فصلنامه کتاب. (بهـــار و تابستان). ص ۵۶-۴۹.

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 ۴۵-۴۴.

بهزادی، ماندانـــا (۱۳۷۵). شیوه ضـــبط اعلام انگلیسی در فارسی. تهران: مرکـــز نشـــر 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 محمـــــد (۱۳۷۲). در جست و جوی زبان علم. مجموعه مقالات سمینارزبان فارسی در زبان علم. تهران: مرکز نشر دانشگاهی. ص ۱۳-۶.

صنعتی، محمد (۱۳۷۱). دشواری‌های زبان فارسی با کامپیوتر. آدینه. ۷۲ (مرداد).
ص ۵۷-۵۶.

کابلی، ایرج (۱۳۷۱). فراخوان برای فارسی نویسی و پیشنهاد به تاجیکان. آدینه. ۷۲
(مرداد) ص ۵۵-۵۰.

مآخذنمونه‌ها

امینی، سید محمد (۱۳۷۰). واژگان فیزیک. تهران: مرکز نشر دانشگاهی.

باقری، محمد (۱۳۷۲). واژگان ریاضی. تهران: نشر فرهنگان.

بریجانیان، 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 هیأت‌امنای کتابخانه‌های عمومی کشور.

همایون، همادخت (۱۳۷۱). واژه نامه زبانشناسی و علوم وابسته. تهران: مؤسسه مطالعاتو تحقیقات فرهنگی.