نما

مجله الكترونيكي مركز اطلاعات و مدارك علمي ايران

شماره اول دوره پنجم

تاريخ انشار اين مقاله: ۵ شهريور ۱۳۸٤

مهارت در جستجوي اطلاعات فارسي از اينترنت[۱]

محمد صابر راثي ساربانقلي[۲]

چكيده:

خط فارسي داراي مشكلات مختلفي مي‌باشد كه در جستجو و بازيابي اطلاعات مسائل و مشكلات فراواني را فراروي كاربران اينترنت قرار مي‌دهد. به خصوص با رشد سريع انتشارات الكترونيكي بر روي وب در شكلهاي مختلف پايگاه‌هاي اطلاعاتي، وبلاگ و ... و اينكه هيچ قاعده مشخص و ثابتي براي رسم‌الخط فارسي وجود ندارد باعث شده است كه جستجوگران مطالب فارسي با مشكلات فراواني روبرو بشوند. اين مقاله سعي دارد تا با اشاره به موارد مختلفي كه مي‌تواند در جستجو و بازيابي اطلاعات سرعت و دقت و جامعيت و مانعيت جستجو را بالابرد موجب افزايش مهارت كاربران اينترنت فارسي بشود.

كليد واژه ها: اينترنت، خط فارسي، جستجو و بازيابي اطلاعات.

مقدمه

اينترنت به عنوان يك محمل اطلاعاتي عظيم، منابع اطلاعاتي را در مقياسي وسيع در دسترس مخاطبان بالقوه قرار داده است. اغلب سهولت دسترسي به منابع اطلاعاتي اعم از متن و ساير رسانه‌ها عمده‌ترين مزيت اينترنت محسوب مي‌شود. اما اين توانايي كه هركس ناشر آثار خود باشد عواقب ناخواسته‌اي را نيز در پي خواهد داشت و آشكارترين معضل، آن است كه انبوهي از منابع بسيار متنوع و غير قابل مديريت را فراهم مي‌آورد. افزايش سريع منابع اينترنتي نيازمند يك سازمان‌دهي مفيد و موثر است. هرچند در حال حاضر راهنماهايي براي منابع اينترنتي تهيه شده است كه براساس فايل‌هاي مقلوب ساخته شدهِ توسط موتورهاي جستجو و با استفاده از قابليت‌هاي مختلف اين موتورها از جمله : استفاده از عملگرهاي بولي، جستجوي دقيق عبارت، محدود كردن يك جستجو به بخش خاصي از ركورد (مانند عنوان ، آدرس) ، كوتاه‌سازي كلمات، جستجوي نزديك‌يابي واژه‌ها، ايجاد محدوديت زماني و منطقه‌اي و زباني، و .... به جستجوي اطلاعات كمك مي‌كند، اما بايد تاكيد كرد كه در امر بازيابي اطلاعات از اينترنت بدون نمايه‌سازي نظام يافته نمي‌توان انتظار بازيابي مفيد و موثر را داشت. هرچند بيش‌تر اطلاعات موجود بر روي اينترنت به زبان انگليسي است، ولي حجم اطلاعات به زبان فارسي نيز با سرعت در حال افزايش است و كاربران به دلايل مختلفي علاقه زيادي به اطلاعات فارسي نشان مي‌دهند و از آنجائي‌كه زبان غالب در اينترنت انگليسي است جستجو به زبان‌هاي غير انگليسي از جمله فارسي، مسايل و مشكلات مختلفي را جداي از مشكلات عمومي اينترنت دارد.

خط فارسي

اشكال و نقصي كه در همه خطوط جهان است دو علت دارد كه يكي در اصل خط است و ديگري بر اثر تغيير و تحول زبان ايجاد مي‌شود. دقت فراوان در ثبت همه دقايق تلفظ اغلب موجب دشواري شيوه خط است و اين دقت زماني ضرورت مي‌يابد كه زباني توسعه بسيار بيابد و در كشورهاي ديگري كه به آن زبان سخن نمي‌گويند رايج شود. به عنوان مثال در خط عربي نقطه و علامت‌هاي حركات وقتي به وجود آمد كه زبان عربي نزد ملت‌هاي غير عرب معمول شد، در خط يوناني نيز نشانه‌هاي آهنگ و تكيه[۳] پس از رواج آن زبان در مصر ايجاد شد تا كساني كه زبان مادري‌شان يوناني نبود و با تلفظ آن مانوس نبودند بتوانند كلمات و عبارات يوناني را هر چه درست‌تر ادا كنند. با اين حال هيچ خطي هر قدر دقيق و شماره علامات آن فراوان باشد، ممكن نيست كه كاملاً نشانه شيوه تلفظ باشد. و با كمك علامات متعدد علم حروف نيز تا كسي چگونگي تلفظ زباني را نشنود نمي‌تواند عبارت و كلمات آنرا مانند اهل آن زبان ادا كند.

اما نقصي كه بر اثر تحول زبان و به تدريج در خط حاصل مي‌شود، مشكلي است كه همه ملت‌ها با آن رو به رو هستند. بعضي از حروف و اصوات زبان در طي زمان تغيير مي‌پذيرند و اين تغيير در گفتار حاصل مي‌شود، اما خط هميشه صورت كهن تلفظ را حفظ مي‌كند، و از اينجا ميان "گفتار" و "نوشتار" اختلاف روي مي‌دهد. ديگر آن كه هر زباني ناگزير لغاتي از زبان‌هاي ديگر به عاريت مي‌گيرد و اگر علائم خط در اين دو زبان يكي باشد كلمه خارجي به همان املاي اصلي در نوشتن به كار مي‌رودكه اغلب با املاي كلمه مشابه در زبان ثانوي تفاوت دارد و از اينجا براي اصوات واحد علائم خطي متعدد پديد مي‌آيد. در خط فارسي نمونه همه اين موارد را مي‌توان يافت. چون خط عربي براي نوشتن فارسي به كار رفت كلماتي كه از آن زبان اخذ شده بود به همان صورت اصلي نوشته شد. حال آنكه به يقين در هيچ دوره‌اي حروف خاص عربي را فارسي زبان‌ها درست مثل اصل تلفظ نكرده‌اند. در زبان‌هاي ديگر نيز اين گونه موارد نمونه‌هاي متعدد دارد. شايد دو زبان انگليسي و فرانسه بيش از همه زبان‌هاي جهان دچار اختلاف تلفظ و خط باشند. به طور كلي نقائص و معايبي كه در خطوط معمول جهان است را مي‌توان به طريق زير طبقه‌بندي كرد:

۱. شكل واحدي اصوات مختلف را بيان مي كند. چنانكه در فارسي حرف "ي" را گاهي براي حرف لين بكار مي‌بريم (يك) و گاهي براي حرف مد (بي) و گاهي به جاي الف (عيسي) و گاهي براي نشان دادن مصوت مركب (ري). و يا حرف «و» در كلمات (سوار، سود، تو)

۲. اصوات واحد به صورت‌هاي مختلف نوشته مي شود. در فارسي حرف "س" سه صورت (س – ص – ث) و حرف "ز" چهار صورت (ز – ذ- ض- ظ) دارد؛ در زبان فرانسه حروفي كه "سن" خوانده مي‌شود پنج رسم الخط دارد كه اگر صورت‌هاي جمع را نيز به حساب بياوريم ده شكل مي‌شود از اين قرار (saint, ceint, sein, seing, sain)

۳. بسياري از حروف نوشته مي‌شود ولي خوانده نمي‌شود. يعني علاماتي بي‌فايده در نوشتن به كار مي‌رود در فارسي نوشتن "واو معدوله" و "هاء غير ملفوظ" از اين قبيل است. در انگليسي نمونه اين مورد بسيار است مانند high كه دو حرف آخر آن به كلي از تلفظ ساقط است. و يا "K" در كلمه "Know" .

۴. اصواتي هستند كه تلفظ مي‌شود اما در خط نشانه‌اي براي آن‌ها نيست. در فارسي سه مصوت كوتاه (َ ِ ُ) از اين قبيل است هم چنين الف در كلمات اسحق و الله كه در كتابت نمي‌آيد. [٤]

زبان و خط فارسي نيز مشكلات خاصي را دارا مي‌باشد و نظام نوشتاري فارسي براي ثبت دقيق گفتار، نارسائي دارد و قواعد نگارش آن مدون نيست، از اين رو فاصله ميان گفتار و نوشتار در فارسي قابل توجه است. بيش‌ترين مشكلات نيز به جهت نبود يك رسم الخط واحد كه عموم اساتيد و اهل فن روي آن اجماع كرده باشند به وجود آمده است. به طوري كه در حال حاضر جداي از چندين شيوه‌نامه رسمي همچون" شيوه‌نامه سمت، نشر دانشگاهي، فرهنگستان، آموزش و پرورش" به تعداد افراد جامعه، رسم الخط و شيوه نگارش زبان وجود دارد، هر ناشري براي خود به قاعده‌اي دلخواه عمل مي‌كندكه اين تعددها موجب پريشاني و پراكندگي شده و با يك‌ديگر تفاوت‌هايي دارند. از ديگر دلايل مي‌توان به عاريتي بودن خط فارسي و چاره‌انديشي براي حركات و عدم تطابق واج‌ها با حروف اشاره كرد. متصل و منفصل‌نويسي نيز يكي ديگر از حوزه‌هاي مورد اختلاف است از ديگر مشكلات: گوناگوني معادل‌هاي علمي، انواع مختلف ضبط اسامي خارجي، سرهم‌نويسي، جدانويسي، بي‌فاصله‌نويسي، انواع جمع‌ها، صورت‌هاي مختلف نوشتاري، آوانويسي اسامي عناصر و تركيبات شيميايي، سرواژه‌ها و كوته‌نوشت‌ها مي‌باشد.

به طور كلي نقص‌هايي كه براي زبان فارسي شمرده‌اند به شرح زير مي‌توان عنوان كرد:

۱.سه مصوت كوتاه يعني حركات زير و زبر و پيش (َ ِ ُ) را از نوشتن ساقط مي‌كنيم. و اين باعث مي‌شود به جاي اين كه از خط و نوشتار پي به معني ببريم بايستي از معني كلمه و جايگاه آن در جمله آن‌را درست بخوانيم مانند كلمات (كَرَم، كَرَم، كِرم، كُرُم، كِرْم) و (مَلَك، مَلِك، مُلك، مِلك) و يا سه كلمه (حَكَم، حُكم، حِكَم) و نيز ننوشتن مصوت‌هاي كوتاه در داخل متن باعث مي‌شود كه براي تلفظ صحيح اجباراً لاتين كلمات به صورت پانويس متن آورده شود كه همين امر باعث اتلاف وقت و انرژي مي‌شود. كه البته همين لاتين‌نويسي هم قاعده خاصي ندارد و هر ناشر و نويسنده‌اي سليقه خاص خودش را براي آوانويسي حروف فارسي به لاتين دارد. كه به عنوان نمونه براي نشان دادن حركت فتحه و الف و آ هيچ‌گونه هماهنگي در كتاب‌ها و خصوصا فرهنگ‌هاي مختلف ديده نمي‌شود. "هر چند برخي معتقدند همين ننوشتن حركات مزيتي است و موجب تندنويسي مي‌شود"[٥].

۲.براي يك حرف چند علامت مختلف داريم مانند علامت‌هاي (س،ص، ث) كه هر سه در فارسي يكسان خوانده مي‌شوند و هم چنين (ذ، ز، ض، ظ) و نيز (ت، ط). البته اين امر در زبان انگليسي هم وجود دارد چنان كه «ف» ممكن است به شكل‌هاي « F. GH. PH. V» باشد.

۳.يك علامت را براي دلالت بر چند حرف مختلف استعمال مي‌كنيم مانند "و" كه پنج مورد نوشتن دارد يكي براي بيان ضمه در كلمات "خوش" و "تو". ديگر بيان مصوت ممدود يا "واو ماقبل مضموم" مانند "شور" و "او". سوم بيان حرف صامت "واو" در

كلماتي چون "آواز" و "والي" و "عفو" . چهارم بيان حرف مصوت مركبي كه در كلمات "نو" و "جوشن" و مانند آن‌هاست. پنجم حرفي كه در زبان كنوني خوانده نمي‌شود مانند "واو معدوله" در كلمات "خواهر" و "خواستن" و "واو" در كلمه "عمرو"[٦]

۴.حرف‌هايي هم هست كه در كلمات خاصي از نوشتن حذف مي‌شود مانند "الف" در كلمات "اسحق" و "اسمعيل" و "اللّه"

۵.نقطه‌هايي متعدد در بالا و پائين حرف كه هم سبب دشواري و هم موجب اشتباه در خواندن مي‌شود. اهميت بيش از حد نقطه درخط فارسي هنگام تشخيص نوري كاراكترها[٧] توليد اشكال اساسي مي‌كند. به عنوان مثال در نظر بگيريد كه تفاوت ‹ر› و ‹ز› و يا تفاوت ‹د› و ‹ذ› و يا تفاوت ‹ب› ‹ت› ‹ پ› ‹ث› فقط در نقطه است و چون نقطه جزء بسيار كوچكي است در اين امر مشكلات زيادي را فرا روي متخصصين قرار مي‌دهد. و يا كلمات زير را در نظر بگيريد كه با يك يا چند نقطه عوض مي‌شوند (بُر، بَر، پُر، پَر، تَر، پُز، پَز، بُز، تِز).

۶. يك عيب ديگر هم كه براي خط فارسي ذكر كرده‌اند اين است كه از راست به چپ نوشته مي‌شود. و براي اين مورد دلايل مختلفي ذكر شده است از جمله عدم هماهنگي و ايجاد مشكل در نوشتن متون رياضي و شيمي و نت‌هاي موسيقي و دستورات شطرنج و اين كه خط تصويري يعني علائم گرافيكي كه در كل جهان استفاده مي‌شود مانند علائم راهنمائي و رانندگي تماماً از چپ خوانده مي‌شوند.

۷.پيوسته‌نويسي و جدانويسي كلمات مركب كه در اكثر موارد به صورت سليقه‌اي عمل مي‌شود مانند تنوع استفاده از ‹مي› چسبان و غير چسبان و يا تنوع نحوه به كار بردن «علامت‌هاي جمع ‹ها، ان، جات› ، هم، هيچ، كه، (ضماير شخصي متصل مان، تان، شان)، شناسي، را، چه، چون، تر، ترين، بي (پيشوند نفي)، به، اي (نشانه ندا)، آن و اين» در كلمات به صورت پيوسته و يا جدا گانه: (آنچه ، آن چه)؛ (همچنانكه، همچنان‌كه) ؛ (جنابعالي، جناب‌عالي)؛ (هيچكس، هيچ‌كس)؛ (ميتواند، مي‌تواند)؛ (آن ها، آنها) در اين مورد كلماتي كه پيشوند و يا پسوند دارند نيز در شكل‌هاي مختلف نوشته مي‌شوند. برخي از كلمات در دو شكل متصل‌نويسي و منفصل‌نويسي به دو شكل مختلف ظاهر مي‌شوند، مانند «علاقمند و علاقه‌مند؛ انديشمند و انديشه‌مند». مصدرها و فعل‌هاي مركب و اسم‌هاي مشتق از آنها نيز به دو صورت متصل و منفصل نوشته مي‌شوند مانند «نگه‌داشتن و نگهداشتن». در جستجوي مطالب از اينترنت اين مورد توليد اشكال مي‌كند چنانكه جستجوي «هيچ‌كس» نتايج متفاوتي را با جستجوي «هيچكس» مي‌آورد و يا جستجوي «كتاب‌شناسي» و «كتابشناسي» در موتور جستجوي گوگل نتايج متفاوتي را ارائه مي‌كند. اين گونه كلمات با اين كه در خواندن متن اشكال كمي به وجود مي‌آورند و هر آشناي به زبان فارسي به راحتي مي‌تواند آن را بخواند اما در فن‌آوري امروزه و تجزيه و تحليل كلمات به كمك رايانه اشكال اساسي توليد مي‌كند و شايد اگر قاعده‌اي جامع و مانع براي آن وضع گردد، بتوان گفت بزرگ‌ترين مشكل خط فارسي حل شده است. منظور اين كه، براي مثال خواندن سه كلمه «بي‌حوصلگي، بيحوصلگي، بي‌حوصله‌گي» مشكلي ايجاد نمي‌كند. اما در محيط الكترونيكي و شبكه اينترنت براي بازيابي اين كلمه بايستي براي تمام اشكال اين كلمه، جستجو را انجام دهيم، البته اگر آگاهي از تمام اشكال نوشتاري آن داشته باشيم.آآ

۸.سي و دو حرف الفباي فارسي همراه با چهار علامت مد، همزه، تنوين، تشديد به ۱۳۰ شكل مختلف ظاهر مي‌شوند و تفاوت اين اشكال در اتوماسيون خط فارسي توليد اشكال مي‌كند. « تنوع و تعدد نويسگان، يادگيري زبان و خط فارسي را براي آموزگار و آموزنده دشوار و براي نوآموز توان‌فرسا مي‌سازد. تعداد زياد نويسگان در رابطه با اتوماسيون زبان توسط رايانه مشكلاتي در خصوص تعداد و ترتيب قرار گرفتن نويسگان در جداول كد ايجاد مي‌نمايد و طراحان كد در جاي دادن اين تعداد نويسه در جداول با مساله كمبود جا رو به رو هستند. هر چند كه مشكل جا با كد ١٦ بيتي حل شده است اما مسايل ديگري همچنان باقي مي‌مانند كه احتياج به برطرف شدن دارند»[٨]

۹. نوشتن ك و گ (كـ گـ ك گ گ ك) در اشكال مختلف نيز باعث سردرگمي و عدم جستجوي صحيح مي‌شود.

۱۰. در اغلب اوقات يك فاصله اضافي معني متفاوتي و يا متضادي را مي‌دهد (مثل مادر ، ما در).

۱۱. سه كرسي مختلف براي حرف‌هاي مختلف الفبا باعث مي شود كه در مقايسه با اكثر زبان‌ها تعداد سطرهاي هر صفحه به مراتب بيش‌تر گردد چون برخي حروف روي خط كرسي قرار مي‌گيرند و برخي پائين خط كرسي و برخي بالاي خط كرسي مثل (ا ب م)

۱۲.از آنجائيكه حروف در نوشتن غالباً به صورت چسبيده و پيوسته نوشته مي‌شوند و اين امر تشخيص حرف به حرف نوشته به وسيله رايانه را، دچار مشكل مي‌كند.

۱۳. در او. سي. آر. فارسي هم چنين اعداد نيز مشكل ساز هستند چنانچه صفر در فارسي يك نقطه كوچك است كه مي‌تواند رايانه را به اشتباه بيندازد و نيز اعداد ۱ و ۲ و ۳ بسيار شبيه هم هستند و تفاوت‌شان در يك دندانه كوچك است.

۱۴. تنوع املائي يا تنوع در رسم الخط بعضي از كلمات كه همه شكل‌هاي آن نيز درست است مانند(اتاق و اطاق) و يا (امپراتور و امپراطور). و كلماتي كه فقط يك شكل آنها صحيح مي‌باشد ولي شكل ناصحيح آن نيز زياد استفاده مي‌شود مانند «ذغال و زغال؛ خوشنود و خشنود». البته اين جداي از تنوع در مفهوم كلمات است كه در ديگر زبان‌ها نيز وجود دارد، يعني