**نام مقاله: بررسي مشكلات ريخت شناسي زبان فارسي در سه پايگاه اطلاعاتي مركز منطقه اي اطلاع رساني علوم و فناوري، پژوهشگاه اطلاعات و مدارك علمي ايران و جهاد دانشگاهي**

**نام نشريه: فصلنامه كتابداري و اطلاع رساني (اين نشريه در www.isc.gov.ir نمايه مي شود)**

**شماره نشريه: 50 \_ شماره دوم، جلد 13**

**پديدآور: مرضيه گل تاجي، سعيده بذرگر**

**چكيده**

**اين پژوهش به بررسي مسائلي پرداخته است كه پايگاه­هاي مقاله‌هاي فارسي در جستجوي ريختهاي مختلف يك كلمه با آن روبرو هستند. براي پاسخگويي به سؤالهاي پژوهش، از روش پيمايش مقايسه‌اي استفاده شده است. جامعه پژوهش عبارت است از سه پايگاه مقاله‌هاي فارسي «مركز منطقه‌اي اطلاع‌رساني علوم و فناوري»، «جهاد دانشگاهي»، و «پژوهشگاه اطلاعات و مدارك  علمي ايران». محققان سياهه‌اي شامل 17 كليدواژه را با دقت در متون فارسي انتخاب نموده‌اند كه هركدام از آنها نمايانگر يك مورد از چالشهاي زبان فارسي در برخورد با فناوري نوين هستند. سپس كليدواژه‌ها در جعبه جستجوي پايگاه‌هاي مذكور وارد و نتايج هر كدام نيز ثبت گرديد. اين بررسي نشان داد چالشهاي ريختيِ شناخته شدة زبان فارسي، تأثير زيادي بر بازيابي اطلاعات در هر يك از سه پايگاه مورد نظر دارد. همچنين، هيچ‌كدام از اين سه پايگاه  به شيوه اي جامع و قابل ملاحظه به حل مسائل ريخت شناسي واژگان فارسي نپرداخته اند و هر پايگاه به صورت جداگانه از ميان 17 چالش پيش رو تنها به رفع تعداد محدودي از آنها پرداخته است.**

**كليدواژه­ها: بازيابي اطلاعات، ريخت­شناسي، پايگاه مركز منطقه‌اي اطلاع‌رساني علوم و فناوري، پايگاه پژوهشگاه اطلاعات و مدارك علمي ايران، پايگاه جهاد دانشگاهي، زبان فارسي.**

**مقدمه**

كتابداران از مدتها پيش دريافتند بين تحليل موضوعي مطالب و زبانشناسي، رابطه‌اي مستقيم و الزامي وجود دارد. اين رابطه با پيدايش علم اطلاع‌رساني و كاربرد رايانه در اين رشته، شتاب و اهميت بيشتري يافت. امروزه اطلاع‌رساني و اصطلاح‌شناسي شاخه‌هايي از علوم را تشكيل مي‌دهند كه ارتباطي نزديك و مداوم بين آنها برقرار است. دانش اطلاع‌رساني در حوزة فعاليت خود عمدتاً با اطلاعات نوشتاري، كه زبان وسيلة اصلي انتقال آن است، سر و كار دارد. هسته اصلي هر زبان ويژه، اصطلاحات علمي يا واژگان آن است. اين اصطلاحات براي ارتباط علمي و انتقال صحيح اطلاعات به كار گرفته مي شود و چنانچه دچار هرج و مرج و نابساماني شود، زبان تفهيم و تفاهم و جريان درست اطلاعات مختل مي‌گردد. كتابداران و اطلاع‌رسانان كه رابط بين توليدكنندگان و مصرف‌كنندگان اطلاعات هستند، پيش از ساير متخصصان ضرورت استاندارد كردن واژگان علوم را دريافتند و همزمان با توسعه بانكهاي اطلاعاتي، به رعايت آن اصرار ورزيدند (مرتضايي، 1381).

ما اكنون در دوراني به سر مي بريم كه با حجم عظيمي از اطلاعات در موضوعات متنوع روبرو هستيم. اين كثرت اطلاعات در محيطهاي الكترونيكي و بخصوص وب، گرچه باعث تسهيل دستيابي كاربران به اطلاعات مورد نياز شده، مستلزم به‌ كارگيري شيوه‌ها و تمهيدات خاص در بازيابي آنهاست.

از آنجا كه زبان فارسي، در مواجهه با محيط الكترونيكي، از جهت شيوه نگارشي، داراي مشكلاتي است كه بر كيفيت كاوش در محتويات آن تأثير مي‌گذارد، تأًثير برطرف سازي اين موانع در طراحي هر پايگاه اطلاعاتي فارسي زبان بر ميزان بازيابي ركوردهاي مرتبط، چه از لحاظ كميّت و چه از لحاظ محتواي ركوردهاي بازيابي شده، بركسي پوشيده نيست. از اين رو، پژوهش حاضر سعي دارد با در نظر گرفتن عمده‌ترين مشكلات نگارشي در زبان فارسي و نتايج بازيافت هر كدام در سه پايگاه اطلاعاتي «مركز منطقه‌اي اطلاع‌رساني علوم و فناوري»، «پژوهشگاه اطلاعات و مدارك علمي ايران» و «پايگاه اطلاعات علمي جهاد دانشگاهي» توانايي اين پايگاه‌ها را در شيوه برخورد با اين مشكلات و يا برطرف كردن چالشهاي مربوط، مقايسه كند.

**تعريف ريخت‌شناسي****[[1]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn1" \o "):** بررسي ساخت كلمه و اجزاي تشكيل‌دهندة آن كه شامل پايه، پيشوند، ميانوند و پسوند مي‌شود.

**معايب خطوط رايج جهان**

به طور كلي، نقايصي در خطوط رايج جهان وجود دارد كه مي‌توان آنها را بدين‌گونه طبقه‌بندي كرد:

1- صداهاي يكسان به وسيله حروف مختلفي نوشته مي شود. در فارسي صداي «س» به سه صورت (س، ص، ث) و صداي «ز» چهار صورت (ز، ذ، ض، ظ) دارد؛ در زبان فرانسه كلمه «سن» پنج شيوه نگارش دارد كه اگر صورتهاي جمع را نيز به حساب بياوريم، ده شكل مي‌شود: (saint, ceint, sein, seing, sain).

2- بسياري از حروف نوشته مي‌شوند، ولي خوانده نمي‌شوند؛ يعني نشانه‌هايي بي‌فايده در نوشتن به كار مي‌رود. در فارسي نوشتن «واو معدوله»[[2]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn2" \o ") 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، يا K در كلمه know (ناتل خانلري، 1373).

3- چند صداي مختلف با يك حرف نمايانده مي شود:

city/cook or ginder/girl

4- يك صداي واحد، به صورت مجموعه‌اي از چند حرف نگاشته مي‌شود:

 shoot/character/nation/physics/coat

5- بعضي صداها معادل حرفي ندارند. مثلا صداي y قبل از u شنيده مي‌شود، ولي نوشته نمي‌شود:

 cute/futle/utility

6- گاهي يك حرف نماينده مجموعه‌اي از صداهاست؛ مانند حرف x در كلمه box كه صداي /ks/ مي‌دهد (يارمحمدي، 1364).

**ويژگيهاي زبان و خط فارسي**

زبان و نگارش فارسي از ويژگيهاي متنوعي برخوردار است كه هر كدام از آنها به نوعي مي‌تواند بر بازيابي اطلاعات در پايگاه­هاي فارسي زبان تأثير بگذارد. در ادامه، به برخي ويژگيهاي شاخص و تأثيرگذار در بازيابي اطلاعات به زبان فارسي اشاره مي شود:

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

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

3. 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».

4. حرفهايي وجود دارد كه در بعضي از كلمات هنگام نوشتن حذف مي‌شود؛ مانند «الف» در كلمات «اسحق» و "اسمعيل" و "الله". در عين حال حذف اين حروف دائمي نيست و بيشتر به سليقه نگارشي افراد بستگي دارد؛ مانند اسماعيل يا اسحاق.

5. نقطه‌‌هايي متعدد در بالا و پايين حرف كه هم سبب دشواري و هم موجب اشتباه در خواندن مي‌شود. اهميت بيش از حد نقطه درخط فارسي هنگام تشخيص نوري كاراكترها [[3]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn3" \o ") (اُ. سي. آر. )اشكال اساسي توليد مي‌كند. به عنوان مثال، كلمات زير را در نظر بگيريد كه با يك يا چند نقطه عوض مي‌شوند (بُر، بَر، پُر، پَر، تَر، پُز، پَز، بُز، تِز).

6. خط فارسي از راست به چپ نوشته مي‌شود و اين امر نيز به نوبة خود مشكلاتي به وجود مي‌آورد، از جمله نبود هماهنگي و ايجاد مشكل در نوشتن متون رياضي و شيمي، نت ‌هاي موسيقي، دستورات شطرنج؛ خط تصويري يعني علايم گرافيكي كه در كل جهان استفاده مي‌شود؛ مانند علايم راهنمايي و رانندگي همگي از چپ به راست نگاشته مي شود.

7. پيوسته‌نويسي و جدا‌نويسي كلمات مركب كه در اكثر موارد به صورت سليقه‌اي اعمال مي‌شود مانند تنوع استفاده از «مي» چسبان و غير چسبان و يا تنوع نحوة 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 جداگانه: (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.

8. سي و دو حرف الفباي فارسي همراه با چهار علامت مد، همزه، تنوين، تشديد به 130 شكل مختلف ظاهر مي‌شوند و تفاوت اين اشكال در اتوماسيون خط فارسي توليد اشكال مي‌كند. «تنوع و تعدد نويسگان[[4]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn4" \o ")، يادگيري زبان و خط فارسي را براي آموزگار و آموزنده دشوار و براي نوآموز توانفرسا مي‌سازد. تعداد زياد نويسگان در رابطه با خودكارسازي زبان توسط رايانه مشكلاتي در خصوص تعداد و ترتيب قرار گرفتن نويسگان در جدولهاي كد ايجاد مي‌كند و طراحان كد در جاي دادن اين تعداد نويسه در جدولها با مسئله كمبود جا رو به رو هستند. هر چند مشكل جا با كد 16 بيتي حل شده است، اما مسائل ديگري همچنان باقي مي‌مانند كه احتياج به برطرف شدن دارند» (محقق زاده و زارعيان،1383).

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

10. در اغلب اوقات يك فاصله اضافي معناي متفاوت و يا متضادي را مي‌دهد (مثل مادر ، ما در).

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

12. از آنجا كه حروف در نوشتن اغلب به صورت چسبيده و پيوسته نوشته مي‌شوند، تشخيص حرف به حرف نوشته به وسيلة رايانه را، دچار مشكل مي‌كند.

13. در اُ. سي. آر.فارسي همچنين اعداد نيز مشكل‌ساز هستند، چنانكه صفر در فارسي يك نقطه كوچك است كه مي‌تواند رايانه را به اشتباه بيندازد و نيز اعداد 1 و 2 و 3 بسيار شبيه هم هستند و تفاوتشان در يك دندانه كوچك است.

14. تنوع املايي يا تنوع در رسم‌الخط 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 است، ولي شكل ناصحيح آن نيز زياد استفاده مي‌شود، مانند «ذغال و زغال؛ خوشنود و خشنود». البته اين جدا از تنوع در مفهوم كلمات است كه در ديگر زبانها نيز وجود دارد، يعني براي بعضي از مفاهيم ممكن است كلمات متنوعي استفاده شود؛ مانند «كامپيوتر و رايانه».

15. به كار بردن همزه درصورتهاي مختلف مانند (مسأله، مسئله)‍؛ (مسئول، مسوول).

16. استفاده از « ا» و « آ » به جاي يكديگر مانند (فرايند و فرآيند).

17. شكلهاي مختلف ضبط نامهاي بيگانه در فارسي: ورود واژه ‌هاي بيگانه معمولا از راه ورود پديده‌هاي فرهنگي نو در عرصه ‌هاي مختلف فني، علمي، اجتماعي، سياسي و هنري و .... و يا از طريق افراد دو زبانه انجام مي‌گيرد كه به وام‌گيري زبان معروف است و كم و بيش در تمام زبانها وجود دارد. واژه‌هاي بيگانه اغلب براي پركردن خلأ واژه‌هاي علمي و يا ارتباطي سودمندند، اما وجود آنها مسائلي از قبيل چگونگي ضبط آنها در زبان وام را به‌ وجود مي‌آورد. براي ضبط واژه‌هاي به وام گرفته شده به سبب اختلاف فاحش نشانه‌هاي الفباي فارسي با نشانه‌هاي الفباي خارجي، مشكلات جدي وجود دارد. از جمله اينكه الفباي فارسي آوانگار نيست و به همين جهت در ضبط دقيق تلفظ واژه‌هاي زبان فارسي نيز ناتوان است. اين ناتواني در ضبط واژه‌هاي بيگانه به مراتب بيشتر است. در مورد برگردان اسامي خارجي به خط فارسي نيز قاعده خاصي وجود ندارد و هر كس بنا بر سليقه و ذوق خود اين كار را انجام مي‌دهد، در نتيجه يك كلمه واحد به شكلهاي مختلف نوشته مي‌شود. براي مثال (اتومبيل و اتوموبيل)؛ (كلسيم، كلسيوم، كالسيوم) و يا اسم Franklin به صورت (فرانكلين، فرانكلن، فرنكلين، فرنكلن) ضبط شده است.

18. استفاده يا استفاده نكردن از «ي» در كلمات مختوم به «الف» مانند (موسي و موسا).

19. استفاده يا استفاده نكردن از «ء» براي كلمات مختوم به هاي بيان حركت در حالت مضاف مانند (خانه مسكوني و خانهء مسكوني و يا خانه‌ي مسكوني).

20. استفاده يا استفاده نكردن از اعراب براي كلمات.

21. انواع مختلف جمع براي يك واژه مفرد: به عنوان مثال، جمع بستن يك واژه با علايم جمع فارسي وعلايم جمع عربي مانند (معلم، معلمين، معلمان، معلم‌ها).

22. تنوينهاي زبان عربي نيز از جمله دشواريهاي رعايت اصل همخواني نوشتاري و گفتاري هستند.

23. در نگارش ياء وحدت يا نكره در آخر كلماتي كه به هاء مختفي يا غير ملفوظ ختم مي‌شوند، سه نوع املا‌ ديده مي‌شود. (خانه‌اي، خانه‌يي، خانة).

24.كلمه‌هاي عربي در شكلهاي گوناگون در زبان فارسي نوشته مي‌شوند. (مبدا، مبداء)؛ (ابتدا، ابتداء)؛ (نسبتاً، نسبته، نسبتا) و ....

25. وجود دندانه‌هاي متعدد در كلمات، خواندن كلمات و بخصوص در  او.سي.آر. فارسي اشكال ايجاد مي‌كند؛ مانند كلمات: نشستن و استشهاد.

26. حروف فارسي اغلب مشابهند و با اندكي غفلت به جاي هم نوشته مي‌شوند و مطلب را به كلي دگرگون مي‌كنند، مانند (در، رد، ور) (راثي ساربانقلي، 1384).

بنابراين، با در نظر گرفتن موارد فوق مي توان چنين استنباط كرد كه اين ويژگيها در زبان فارسي  با  وجود اينكه در خواندن متن اشكال كمي به وجود مي‌آورند و هر آشناي به زبان فارسي به راحتي مي‌تواند آنها را بخواند، در فناوري امروزه و تجزيه و تحليل كلمات به كمك رايانه اشكال اساسي توليد مي‌كنند و چنانچه قاعده‌اي جامع و مانع براي آنها وضع گردد، بزرگ‌ترين مشكل خط فارسي حل مي‌شود. منظور اينكه، براي مثال خواندن سه كلمه «بي‌حوصلگي، بيحوصلگي، بي‌حوصله‌گي» مشكلي ايجاد نمي‌كند. اما در محيط الكترونيكي و شبكه اينترنت براي بازيابي اين كلمه بايد براي تمام شكلهاي آن، جستجو را انجام دهيم (البته اگر از تمام شكلهاي نوشتاري آن آگاهي داشته باشيم).

**بيان مسئله و اهميت پژوهش**

امروزه روش غالب در جستجوي اطلاعات از پايگاه­هاي اطلاعاتي، روش كليدواژه اي است. اما جستجو به اين روش، دشواريهاي خاص خود را دارد. چنانچه فردي به دنبال اطلاعاتي در مورد «كتابشناسي» باشد، اين كليدواژه را مي تواند به سه شكل بنويسد: «كتابشناسي، كتاب‌شناسي و كتاب­شناسي». از آنجا كه پايگاه­هاي اطلاعاتي، نظامهايي تطبيق دهنده هستند، دقيقاً همان كلمه­اي را بازيابي خواهند كرد كه وارد جعبه جستجو شده است. بنابراين، براي هركدام از اين شكلها، تعداد نتايج متفاوتي بازيابي خواهد شد. چنانچه كاربري تنها يك شكل از اين سه مورد را به كار ببرد، اطلاعاتي را كه به شكلهاي ديگر نوشته شده است، از دست خواهد داد. از اين رو، در اين مقاله سعي خواهد شد تا مشكلات ريخت‌شناسي زبان فارسي در سه پايگاه اطلاعاتي مركز منطقه‌اي اطلاع‌رساني علوم و فناوري، پژوهشگاه اطلاعات و مدارك علمي ايران و پايگاه اطلاعاتي جهاد دانشگاهي بررسي و در پايان پيشنهادهايي براي بهبود كارايي پايگاه‌هاي اطلاعاتي فارسي ارائه شود.

**هدفهاي پژوهش**

1.    بررسي برخي چالشهاي ريختي شناخته شده زبان فارسي، بر ميزان بازيابي اطلاعات در سه پايگاه اطلاعاتي فارسي مورد بررسي

2.                  مقايسه ميزان توانايي سه پايگاه در رفع چالشهاي ريختي مدنظر.

**پيشينة پژوهش**

از آنجا كه پيشينه­هاي يافت شده براي اين پژوهش به دو دسته مشكلات نگارشي در زبان فارسي و مشكلات نگارشي در ساير زبانها قابل دسته‌بندي است، هر دو گروه را جداگانه بررسي خواهيم كرد.

**الف) پيشينه‌هاي مرتبط در زبان فارسي**

«حرّي» (1372) در مقالة خود با عنوان «كامپيوتر و رسم‌الخط فارسي» بيان مي‌دارد كه يكي از متغيرهاي عمده در ذخيره و بازيابي اطلاعات فارسي، رسم الخط يا شيوه خط فارسي است. گرچه حروف و كلمات به عنوان ورودي و خروجي هر سيستم رايانه‌اي در هر زبان اهميت دارند، خط فارسي به دليل ويژگي آن، در رويارويي با رايانه داراي مسائل پيچيده تري است. وي معتقد است پنج مورد كه اختصاصاً به مسئله پيوند ميان زبان فارسي و رايانه مربوط مي‌شود، از اين قرار است: هماهنگ كردن حروف، استفاده از تكواژها، استفاده از سياهه آماده، پيوند ساختگي ميان كلمات، هماهنگي رسم‌الخط.

«سمائي و همكاران» (1379) در طرح پژوهشي با عنوان «يكسان‌سازي شيوه رسم‌الخط اسامي تركيبات شيميايي در زبان فارسي» تلاش كرده‌اند شيوة نگارش اسامي تركيبات شيميايي و بخصوص تركيبات آلي در زبان فارسي و معضلات مربوط به آن را بررسي و الگوهايي براي يكسان نويسي آنها پيشنهاد كنند.

«نشاط» (1379)، در بررسي خود با عنوان «مسائل رسم الخط فارسي در رويارويي با فناوري نوين اطلاعاتي» سعي دارد با استفاده از شواهد موجود و الزامهاي مربوط به زبان نظامهاي رايانه‌اي به عنوان وجه غالب فناوريهاي نوين و نيز ناسازگاري ميان اين دو ، تصويري از وضع موجود را عرضه و راه‌ حلهاي ممكن را ارزيابي كند.

«مرتضايي» (1381) در مقاله‌اي با عنوان «مسائل زبان و خط فارسي در ذخيره و بازيابي اطلاعات» مشكلات گوناگوني را كه در جريان ذخيره و بازيابي اطلاعات و ايجاد پايگاه هاي اطلاعاتي به زبان فارسي به وجود مي‌آيد، بررسي كرده است.

«محقق‌زاده و زارعيان» (1383) در مقاله‌اي با عنوان «ارائه راه‌حل براي برخي مسائل اتوماسيون و نگارش فارسي» ضمن برشمردن ايرادهايي كه در مورد پردازش خط فارسي به وسيلة رايانه به وجود مي‌آيد، پيشنهادهايي را براي اين مشكل ارائه مي‌كند.

«بي‌جن خان» (1383) در مطالعه خود نقش پيكره هاي زباني را در نوشتن دستور زبان بررسي و نقدهايي را بر پيكره هاي زباني مطرح كرده است. وي به رابطة دستور زبان و پيكره زباني اشاره و نرم‌افزاري را براي انواع جستجو در پيكره‌ها، تجزيه و تحليل آماري داده‌ها و در نهايت گزارش‌گيري از داده‌هاي آماده‌سازي شده، معرفي مي‌كند. وي در نتيجة تحقيق خود بيان مي‌دارد كه با استفاده از اين روش علاوه بر اينكه مي‌شود ساخت احتمالي نظامهاي زباني را مطالعه كرد، يافته‌هاي زبانشناسي نظري را هم مي‌توان در حوزة دستور زبان در قالب فرضيه هاي زماني محك زد.

«راثي ساربانقلي» (1384) در مقاله خود با عنوان «مشكلات جستجو و بازيابي اطلاعات به زبان فارسي در اينترنت، مطالعه موردي: كاربران مركز اينترنت دانشگاه آزاد اسلامي واحد شبستر» مشكلات جستجو و بازيابي اطلاعات به زبان فارسي در اينترنت توسط كاربران مركز اينترنت دانشگاه آزاد اسلامي شبستر را بررسي نمود. نتايج پژوهش نشان داد بيشتر مشكل كاربران در جستجو، توجه نكردن به شكلهاي مختلف نوشتاري واژه و استفاده نكردن از عملگر OR  مي‌باشد.

«عبدالهي نورعلي» (1386) در پژوهش خود با عنوان «كندوكاو مسائل ريخت‌شناسي زبان فارسي در بازيابي اطلاعات از جستجوگرهاي وب» به بررسي مسائلي پرداخته است كه جستجوگرهاي فارسي در كاوش ريخت‌شناسي مختلف يك كلمه با آن روبرو هستند. براي اين مهم از سه جستجوگر بين‌المللي گوگل، ياهو و آلتاويستا[[5]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn5" \o ") كه امكان جستجو به زبان فارسي را دارند، استفاده شد. نتايج نشان داد هيچ كدام از جستجوگرها، چالشهاي زبان شناختي زبان فارسي را در جهت بهبود كاوش، مورد توجه قرار نداده اند. با توجه به دستاوردهاي حاصل از پژوهش، الگويي براي ايجاد اصلاحات در شيوة نگارش فارسي ارائه شد تا از اين طريق بتوان پردازش متون رايانه اي را تسهيل نمود.

**ب) پيشينه‌هاي مرتبط در ساير زبانها**

«هدلاند[[6]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn6" \o ") و ديگران» (2000) ويژگيهاي زبان سوئدي را از نظر بازيابي بررسي كردند. آنها مطالعه‌اي مقايسه‌اي بر روي زبانهاي سوئدي، فنلاندي و انگليسي انجام دادند تا ميزان ابهامهاي لغوي را در اين زبانها معيّن كنند. محققان پيشنهاد مي‌كنند برچسب‌گذاري ادات سخن جهت بازيابي كلمات هم نگاشت، مي‌تواند مفيد باشد.

«مونز و دوريكه»[[7]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn7" \o ") (2002) با تمركز بر اثرات تحليلهاي ريخت‌شناسي همچون ريشه‌سازي و جداسازي كلمات مركب، كارآيي بازيابي اطلاعات را بررسي كردند. اين مطالعه بر روي زبانهاي هلندي، آلماني و ايتاليايي انجام شده است. نتايج نشان داد بازيابي اطلاعات حدود 25% براي زبان آلماني، 69% براي زبان هلندي و 25% براي زبان ايتاليايي بهبود يافت.

«درويش»[[8]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn8" \o ") (2002) روشي را براي ايجاد يك تحليلگر ريخت‌شناسي ارائه مي‌دهد. اين تحليلگر توانايي توليد ريشه‌هاي احتمالاتي يك كلمه را خواهد داشت. در اين نظام، قواعد ريشه‌سازي خودكار مورد استفاده قرار گرفته است. محقق براي ارزيابي اين نظام، آن را با يك تحليلگر ريخت‌شناسي عربي موجود در بازار مقايسه كرده است.

«مقداد»[[9]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn9" \o ") (2005) در يك پژوهش، عملكرد سه ابزار جستجوي عمومي را با سه جستجوگر عربي (كه اختصاصاً مسائل زبان شناختي عربي را لحاظ مي‌كنند) مقايسه كرد. نتايج نشان داد جستجوگرهاي عمومي، نظير آلدوب[[10]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn10" \o ")، آلتاويستا و گوگل در بازيابي مدارك عربي، ناقص عمل مي‌كنند. همچنين، نتايج اين تحقيق، نياز به تحقيقات بيشتر در زمينة عملي بودن ابزارهاي جديد بازيابي اطلاعات در جستجوگرها را نشان داد.

«تاث»[[11]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn11" \o ") (2006) به بررسي قابليتهاي زبان شناختي جستجوگرهاي انگليسي و مجاري پرداخت. محقق سه ابزار جستجوي انگليسي به نامهاي گوگل، آلتاويستا و آلدوب را با پنج جستجوگر محلي مقايسه نمود. تحليل داده‌ها بر پايه چند شاخص انجام شد كه عبارت بودند از: ريشه‌سازي، بازيابي لهجه‌هاي مختلف، كوتاه‌سازي و جستجوي مترادفها. نتايج حاكي از آن بود كه جستجوگرهاي محلي، مسائل زبان مجاري را بهتر از جستجوگرهاي انگليسي مورد توجه قرار داده‌اند. ابزارهاي انگليسي زبان، لهجه‌هاي مختلف زبان مجاري را به خوبي پشتيباني نمي‌كردند، كه اين امر به بازيابي ضعيف اطلاعات منجر مي‌شد.

**پرسشهاي پژوهش**

1-    چالشهاي ريختي شناخته شده زبان فارسي چه تأثيري بر بازيابي اطلاعات در هر يك از سه پايگاه مورد نظر داشته است؟

2-    كدام‌ يك از سه پايگاه مورد نظر، چالشهاي ريختي مورد نظر را در الگوريتمهاي بازيابي خود مورد توجه قرار داده اند؟

**روش­شناسي پژوهش**

اين پژوهش با استفاده از روش پيمايش مقايسه­اي انجام پذيرفته است. داده‌هاي جدولها نيز بر اساس آمار توصيفي بررسي شده است. زمان گردآوري داده­ها مهر 87 بود. از آنجا كه اين سه پايگاه از جمله پايگاه‌هاي مهمي هستند كه مقاله‌هاي فارسي را نمايه مي‌كنند، در پژوهش حاضر مورد بررسي قرار گرفتند.

شيوة اجراي تحقيق بدين شكل است كه ابتدا سياهه­اي مشتمل بر 17 چالش نگارشي در زبان فارسي با استفاده از متون موجود شناسايي شد. سپس براي هريك از آنها مصداقهاي موجود در زبان فارسي انتخاب و بررسي گرديد. مصداقها به صورت كليدواژه هايي در سه پايگاه جستجو شد تا اطمينان حاصل شود دست كم يك پيشينه براي آن چالش وجود داشته باشد. اين كليدواژه ها به عنوان وسيله گردآوري داده ها به كار گرفته شده اند و نتايج هر يك از جستجوها در قالب تعداد ركوردهاي يافت شده براي هر واژه در هر سه پايگاه، در جدول شماره 1 ارائه شده است.

شايان ذكر است، براي اطمينان از اينكه بازيابي‌هاي هم تعداد يك محتوا دارند، ركوردهاي بازيابي شده به صورت گزينشي با هم مقايسه گرديد.

در بعضي موارد نيز براي كنترل رخداد يك واژه از كنترل مدارك همپوشان در سه پايگاه استفاده شد؛ بنابراين تا حد امكان از نبود رويداد يك واژه در سه پايگاه اطمينان حاصل شد، لذا مقدار صفر در جدول يك به معناي يافت نشدن ركورد براي ريخت مورد نظر در پايگاه است.

در جدول شماره 2، نسبت تعداد رخدادهاي مختلف واژگان به صورت دو به دو براي هر واژه در هر يك از سه پايگاه محاسبه گرديد. همان‌طور كه مشاهده مي‌شود، در صورت يكسان بودن تعداد نتايج، برچسب «يك» و در غير اين صورت برچسب «غير از يك» به هر كدام داده شد. بنابراين، مفهوم «يك» در آن جدول به احتمال قوي به معناي يكسان بودن ركوردهاي بازيابي شده از دو صورت واژه در پايگاه مورد نظر است**.** براي مثال، در مركز منطقه اي اطلاع رساني علوم و فناوري، براي واژه «محمد» 13982 ركورد و براي واژه «محمّد» نيز 13982 ركورد بازيابي شد؛ لذا نسبت «يك» ميان اين دو واژه در اين پايگاه برقرار است.

**جدول 1. آمار نتايج بازيابي شده براي هركدام از مشكلات زبان فارسي به تفكيك پايگاه ها**

| **رديف** | **مشكل** | **مشكل ريخت‌شناسي** | **پايگاه** | | |
| --- | --- | --- | --- | --- | --- |
| **مركز منطقه‌اي** | **ايرانداك** | **جهاد دانشگاهي** |
| 1 | **تشديد** | محمد | 13982 | 12881 | 86 |
| محمّد | 13982 | 0 | 0 |
| 2 | **همزة پاياني** | املا | 8 | 2 | 62 |
| املاء | 12 | 1 | 1 |
| 3 | **نشانه‌هاي جمع** | معلمان | 401 | 140 | 103 |
| معلمين | 18 | 0 | 18 |
| 4 | **برگرداندن كلمات خارجي** | آمريكا | 4507 | 676 | 30 |
| امريكا | 572 | 33 | 6 |
| 5 | **هاي غير ملفوظ** | واژگان | 509 | 53 | 65 |
| واژه گان | 0 | 0 | 0 |
| 6 | **تنوين** | واقعاً | 107 | 0 | 1 |
| واقعا | 107 | 34 | 3 |
| 7 | **همزه متصل به ياي وحدت** | رضايي | 1149 | 721 | 1187 |
| رضائي | 222 | 140 | 6 |
| 8 | **استفاده از "ا" و"آ" به جاي هم** | درآمد | 970 | 172 | 146 |
| درامد | 1 | 2 | 3 |
| 9 | **الف مقصوره** | اسحاق | 97 | 37 | 1 |
| اسحق | 20 | 14 | 2 |
| 10 | **پيوسته نويسي ، بي فاصله نويسي يا جدانويسي تركيبات** | كتابشناسي | 683 | 42 | 5 |
| كتاب شناسي | 145 | 22 | 2 |
| كتاب‌شناسي | 683 | 22 | 0 |
| 11 | **تاي منقوط** | مشكات | 13 | 2 | 9 |
| مشكوه | 1098 | 26 | 21 |
| مشكوة | 8 | 0 | 1 |
| 12 | **صامت ميانجي«ي»** | دو استقامت | 2 | 3 | 2 |
| دو‌ي استقامت | 0 | 1 | 0 |
| 13 | **تنوع صورتهاي درست يك كلمه** | اتاق | 3589 | 103 | 41 |
| اطاق | 21 | 4 | 4 |
| 14 | **همزه به صورتهاي مختلف** | مسئول | 77 | 13 | 17 |
| مسؤول | 0 | 1 | 17 |
| 15 | **تنوع در تلفظ** | داود | 1154 | 441 | 5 |
| داوود | 356 | 172 | 2 |
| 16 | **خط تيره** | اقتصادي اجتماعي | 4881 | 295 | 13 |
| اقتصادي - اجتماعي | 1 | 295 | 22 |
| 17 | **نقطه بين سرنامها** | اچ آي وي | 14 | 23 | 5 |
| اچ. آي. وي | 0 | 23 | 0 |

**جدول2. مقايسه نسبت ريختهاي مختلف هر واژه در پايگاه هاي مختلف**

| **رديف** | **مشكل** | **نسبت ريخت كلمات به يكديگر** | **نسبت تعداد ريخت بازيابي شده كلمات به تفكيك پايگاه** | | |
| --- | --- | --- | --- | --- | --- |
| **مركز منطقه‌اي** | **ايرانداك** | **جهاد دانشگاهي** |
| **1** | **تشديد** | محمد/ محمّد | 1 | غير از يك | غير از يك |
| **2** | **همزه پاياني** | املا/ املاء | غير از يك | غير از يك | غير از يك |
| **3** | **نشانه هاي جمع** | معلمان/معلمين | غير از يك | غير از يك | غير از يك |
| **4** | **برگرداندن كلمات خارجي** | آمريكا/امريكا | غير از يك | غير از يك | غير از يك |
| **5** | **هاي غير ملفوظ** | واژگان/واژه گان | غير از يك | غير از يك | غير از يك |
| **6** | **تنوين** | واقعاً/واقعا | 1 | غير از يك | غير از يك |
| **7** | **همزه متصل به ياي وحدت** | رضايي/رضائي | غير از يك | غير از يك | غير از يك |
| **8** | **استفاده از "ا" و"آ" به جاي هم** | درآمد/درامد | غير از يك | غير از يك | غير از يك |
| **9** | **الف مقصوره** | اسحاق/اسحق | غير از يك | غير از يك | غير از يك |
| **10** | **پيوسته نويسي ، بي‌فاصله نويسي يا جدانويسي تركيبات** | كتابشناسي/كتاب شناسي | غير از يك | غير از يك | غير از يك |
| كتاب شناسي/ كتاب‌شناسي | غير از يك | 1 | غير از يك |
| كتابشناسي/كتاب‌شناسي | 1 | غير از يك | غير از يك |
| **11** | **تاي منقوط** | مشكات/مشكوه | غير از يك | غير از يك | غير از يك |
| مشكوه/مشكوة | غير از يك | غير از يك | غير از يك |
| مشكات/مشكوة | غير از يك | غير از يك | غير از يك |
| **12** | **صامت ميانجي«ي»** | دو استقامت/دوي استقامت | غير از يك | غير از يك | غير از يك |
| **13** | **تنوع صورتهاي درست يك كلمه** | اتاق/اطاق | غير از يك | غير از يك | غير از يك |
| **14** | **همزه به صورتهاي مختلف** | مسئول/مسؤول | غير از يك | غير از يك | 1 |
| **15** | **تنوع در تلفظ** | داود/داوود | غير از يك | غير از يك | غير از يك |
|  | **خط تيره** | اقتصادي - اجتماعي/ اقتصادي- اجتماعي | غير از يك | 1 | غير از يك |
| **16** | **نقطه بين سرنام** | اچ آي وي/ اچ. آي. وي. | غير از يك | 1 | غير از يك |
| **17** | **تعداد موارد حل شده** |  | 3 | 3 | 1 |

**يافته‌هاي پژوهش**

    با استفاده از آمارهاي داده شده در جدولهاي فوق، مي‌توان در پاسخ به سؤالهاي پژوهش چنين بيان داشت:

1-     چالشهاي ريختي شناخته شده زبان فارسي چه تأثيري بر بازيابي اطلاعات در هر يك از سه پايگاه مورد نظر داشته است؟

بايد گفت، بر اساس اطلاعات ارائه شده در جدول شماره 1، شاهديم كه هر شكل نوشتاري كلمه در زبان فارسي نتايج متعددي را در هر پايگاه اطلاعاتي در پي دارد. به طور مثال، به بررسي تأثيري كه الف مقصوره و لحاظ يا عدم لحاظ آن در كلمه «اسحاق» داشته است، خواهيم پرداخت:

طبق آمار به دست آمده از جدول شماره 1، براي كلمه «اسحاق» در دو شكل نوشتاري مختلف شاهديم كه چنانچه براي نوشتن اين كلمه از الف مقصوره استفاده نشود، نتايج به دست آمده در سه پايگاه مركز منطقه‌اي اطلاع‌رساني علوم و فناوري، پژوهشگاه اطلاعات و مدارك علمي ايران، و جهاد دانشگاهي به ترتيب تعداد ركوردهاي بازيابي شده 97، 37 و 1 مي باشد و چنانچه در حالتي ديگر براي نوشتن كلمه «اسحاق» از الف مقصوره استفاده شود و شكل نوشتن اين كلمه در جعبه جستجوي پايگاه به شكل «اسحق» باشد، نتيجه متفاوتي به دست خواهد آمد، به طوري كه در اين حالت ركوردهاي بازيابي شده در سه پايگاه مدنظر به ترتيب 20، 14، 2 مي‌باشد. بدين ترتيب، متوجه مي‌شويم مشكل ريختي الف مقصوره در هر سه پايگاه اطلاعاتي فارسي باعث اختلاف در تعداد ركوردهاي بازيابي شده، مي شود و چنانچه كلمه «اسحق» را به جاي كلمه «اسحاق» در جعبه جستجوي پايگاه اطلاعاتي مركزمنطقه اي اطلاع رساني علوم و فناوري وارد كنيم، 77 ركورد و در پايگاه اطلاعاتي پژوهشگاه اطلاعات و مدارك علمي ايران 23 ركورد اطلاعات را از دست خواهيم داد و بدين طريق اين عامل باعث ريزش ركوردهاي اطلاعاتي مفيد خواهد شد، اما در پايگاه جهاد دانشگاهي با شكل «اسحاق» 1 ركورد و با شكل «اسحق» 2 ركورد بازيابي شده است؛ يعني افزايش تعداد ركورد رخ داده است. بنابراين، شاهديم كه چگونه شكلهاي متنوع نوشتار كلمات مي تواند باعث كاهش يا افزايش تعداد ركوردهاي بازيابي شده در پايگاه­هاي فارسي زبان شوند.

2-    كدام‌ يك از سه پايگاه مورد نظر، چالشهاي ريختيِ ذكر شده را در الگوريتمهاي بازيابي خود مورد توجه قرار داده اند؟

بر اساس اطلاعات موجود در جدول شماره 2، شاهديم كه هيچ كدام از سه پايگاه فارسي مورد نظر، به شيوه‌اي جامع چالشهاي ريخت شناسي زبان فارسي را در جهت بهبود نتايج كاوش مورد توجه قرار نداده‌اند، اگرچه در بعضي موارد تساوي تعداد ركوردهاي بازيابي شده در ريختهاي مختلف يك واژه را به احتمال قوي مي‌توان به منزلة رفع آن چالش خاص در الگوريتم بازيابي پايگاه در نظر گرفت، اما   
نمونه‌هايي از اين دست براي هر پايگاه نسبت به حجم مشكلات ريخت‌شناسي موجود، درصد بسيار اندكي را به خود اختصاص مي دهد. به طور مثال، از ميان     17 چالش موجود كه در جدولهاي فوق طرح شد، پايگاه مركز منطقه اي اطلاع‌رساني علوم و فناوري، پژوهشگاه اطلاعات و مدارك علمي ايران، و جهاد دانشگاهي به ترتيب موفق به حل سه، سه و يك مورد از مسائل ريخت شناسي زبان فارسي شدند.پايگاه‌هاي مذكور از ميان تمامي چالشهاي ريخت شناسي مطرح شده تنها براي موارد زير چاره جويي نموده اند:

پايگاه مركز منطقه‌اي اطلاع‌رساني علوم و فناوري: تنوين، تشديد، پيوسته‌نويسي و بي‌‌فاصله‌نويسي؛ پژوهشگاه اطلاعات و مدارك علمي ايران: جدانويسي و بي فاصله نويسي، خط تيره، نقطه بين سرنام‌ها؛ پايگاه جهاد دانشگاهي: همزه به صورتهاي مختلف.

**نتيجه‌گيري**

بديهي است، پردازش بهتر و سريع­تر متون فارسي با استفاده رايانه در زمانه ما يك ضرورت اساسي به نظر مي رسد. پايگاه هاي اطلاعاتي كه با استفاده از زبان و شيوه خط كنوني به ذخيره و بازيابي اطلاعات مي پردازند، نمي توانند كارايي مطلوبي داشته باشند و اين شكلهاي متنوع نوشتار كلمات مي تواند باعث كاهش يا افزايش تعداد ركوردهاي بازيابي شده در پايگاه­هاي فارسي زبان شود. بر اين اساس، شاهديم كه پايگاه‌هاي اطلاعاتي فارسي با وجود عمر نسبتاً كوتاه، با مشكلات بسياري دست به گريبانند، كه اگر هر چه زودتر چاره‌انديشي نشود، با توجه به هجوم اطلاعات ديگر، مهار آن آسان نخواهد بود. نتايج بررسي نشان داد هيچ‌كدام از سه پايگاه فارسي مورد نظر، به شيوه‌اي جامع چالشهاي زبانشناختي زبان فارسي را در جهت بهبود نتايج كاوش مورد توجه قرار نداده‌اند. اگرچه در بعضي موارد تساوي تعداد ركوردهاي بازيابي شده در ريختهاي مختلف يك واژه را مي‌توان به منزلة رفع آن چالش خاص در نظر گرفت، اما نمونه‌هايي از اين دست براي هر پايگاه نسبت به تعدد مشكلات ريخت‌شناسي موجود، درصد بسيار اندكي را به خود اختصاص مي‌دهد. به طور مثال، از ميان 17 چالش موجود كه در جدولهاي فوق طرح شد، پايگاه مركز منطقه‌اي اطلاع‌رساني علوم و فناوري، پژوهشگاه اطلاعات و مدارك علمي ايران، و جهاد دانشگاهي به ترتيب موفق به حل سه، سه و يك مورد از مسائل ريخت‌شناسي زبان فارسي شدند.

**پيشنهادها**

با توجه به اينكه هر كدام از سه پايگاه مورد بررسي توانسته در سطحي محدود بر برخي از چالشهاي ريخت‌شناسي واژگان فارسي فايق آيد، و با در نظرگرفتن اين نكته كه مشكل حل شده هر پايگاه با ساير پايگاه­ها متفاوت است و اين سه پايگاه مورد نظر در كشور ما از پايگاه هاي علمي معتبر به شمار رفته و  هر روز نيز  بر تعداد كاربران آنها افزوده مي شود، متأسفانه اغلب كاربران از آنچه به هنگام جستجو در اين پايگاه ها رخ مي‌دهد، آگاهي ندارند. از اين‌رو، آگاهي نداشتن و همچنين تنوع صورت نوشتاري، تأثير زيادي بر از دست دادن مدارك مربوط در هر يك از سه پايگاه اطلاعاتي مورد بررسي دارد. بنابراين، پيشنهاد مي‌شود طراحان پايگاه‌هاي اطلاعاتي فارسي در نشستي پيرامون اين موضوع به ارائه تجربه‌ها و يافته‌هاي خود پرداخته و از دستاوردهاي ديگران در اين حيطه بهره ببرند. همچنين مي توان هنگام طراحي پايگاه‌ها، آنها را به اصطلاحنامه مجهز نمود تا كاربران از ريختهاي مختلف واژه به اصطلاح پذيرفته شده راهنمايي شوند. همين‌طور طراحان مي توانند تمهيداتي را دربارة چگونگي استفاده از پايگاه و الگوريتمهاي مرتبط با مسائل ريختي واژگان به كاربرده شده براي جستجو در اختيار كاربران قرار دهند تا از اين طريق آنها راحت‌تر به جستجو بپردازند و بتوانند حداكثر نتايج دلخواه خود را بيابند. به نظر مي‌رسد همكاري بين متخصصان زبان شناسي با متخصصان عرصه بازيابي اطلاعات به منظور جهت دهي تحقيقاتي در اين زمينه بسيار ضروري است.

**منابع**

- بي جن خان، محمود (1383). نقش پيكره­هاي زباني در نوشتن دستور زبان، معرفي يك نرم‌افزار رايانه‌اي. مجله زبان­شناسي، 19 (2)، 48-67.

- راثي ساربانقلي و محمد صابر (1384). مهارت در جستجوي اطلاعات فارسي از اينترنت. مجله الكترونيكي نما، 5 (1)، بازيابي 22 آبان 1387، از

<http://www.irandoc.ac.ir/Data/E_J/vol5/rasi.htm>

- حري، عباس (1372). كامپيوتر و رسم‌الخط فارسي. پيام كتابخانه. 3 (1)، 6-11.

- سمائي، مهدي (1379). يكسان‌سازي شيوه رسم‌الخط اسامي تركيبات شيميائي در زبان فارسي. طرح پژوهشي، مركز اطلاعات و مدارك علمي ايران، تهران.

- عبدالهي نورعلي، محمدصادق (1386). كندوكاو مسائل ريخت شناسي زبان فارسي در بازيابي اطلاعات از جستجو گرهاي وب. پايان نامه كارشناسي ارشد كتابداري و اطلاع رساني، دانشگاه شيراز، شيراز.

- محقق‌زاده، محمدصادق و كاظم زارعيان (1383). ارائه راه‌حل براي برخي مسايل اتوماسيون نگارش فارسي. فصلنامه اطلاع‌رساني، 19( 3-4 )، 1-10.

- مرتضايي، ليلا (1381). مسائل زبان و خط فارسي در ذخيره‌سازي و بازيابي اطلاعات**.** فصلنامه اطلاع‌رساني، 17 (1-2)، 1-7.

- ناتل خانلري، پرويز (1373). زبان‌شناسي و زبان فارسي. تهران: توس.

- نشاط، نرگس (1379). مسائل رسم‌الخط فارسي در رويارويي با فناوري نوين اطلاعاتي. در فهرست­هاي رايانه‌اي، كاربرد و توسعه.مجموعه مقالات همايش كاربرد و توسعه فهرست­هاي رايانه‌اي در كتابخانه­هاي ايران، آبان 27-28، (401-408). مشهد: دانشگاه فردوسي مشهد.

- يارمحمدي، لطف الله (1364). درآمدي به آواشناسي. تهران: مركز نشر دانشگاهي.

- Darwish, K. (2002). Building a Shallow Arabic Morphological Analyzer in One Day. Annual Meeting of the ACL, Proceedings of the ACL-workshop on Computational approaches to Semitic languages. Philadelphia, 19-28.

- Hedlund, T., Pirkola, A. ,& Kalervo, J. (2001). Aspects of Swedish morphology and Semantics from the perspective of mono- and cross-language information retrieval. Information Processing and Management,37,147-161.

- Retrieved November 3, 2008, from http://www.dcs.shef.ac.uk /nlp/clarity/papers/SWEIR-Hedlund-IPM01.pdf .

- Monz, C. , & De Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross language Evaluation Forum, CLEF 2001, Darmstadt, Germany.

- Moukdad, H. (2005). Lost In cyberspace: How Do Search Engines Handle Arabic Queries?**.**The international information & library review,37(4),237-394.

- Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search Engine for Various Queries. Libri, 56, 38-47.

1. Morphology.

1. واو معدوله، واوي است كه در اين زمان عموماً نوشته مي‌شود ولي خوانده نمي‌شود، مانند خواهش. اما در زمانهاي قديم آن را با كيفيت خاصي تلفظ مي‌كرده‌اند و چون در هنگام تلفظ ضمه به فتحه عدول مي‌كرده است، آن را واو معدوله ناميده‌اند. هنوز در برخي از لهجه‌ها تلفظ آن به صورت قديم مانده است. پيش از واو معدوله هميشه حرف «خ» و پس از آن يكي از حروف «د. ر. ز. س. ش. ن. و. ه. ي» آمده است.

1. optical character recognition(ocr).

1. نويسگان، جمع نويسه معادل Characters

[[5]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref5" \o "). Altavista.

[[6]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref6" \o "). Hedlund.

[[7]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref7" \o "). Monz & De Rijke.

[[8]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref8" \o "). Darwish.

[[9]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref9" \o "). Moukdad.

[[10]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref10" \o "). Altheweb.

[[11]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftnref11" \o ").Toth.