



پژوهشگاه علوم انسانی و مطالعات فرهنگی
رتال جامع علوم انسانی

طرح پایگاه متون زبان و ادب فارسی

مقاله

آنها و عدم امکان گردآوری تمامی آنها حتی در کتابخانه‌های بزرگ، سبب شده است که کتابخانه‌ای را در این سرزمین پهناور نتوان سراغ کرد که تمامی این آثار را در خود جای داده باشد. پراکندگی این آثار در کتابخانه‌های مختلف اعم از خصوصی و عمومی و بزرگ و کوچک مانع بزرگی بر سر راه پژوهش محققان، دانشجویان و دانش‌آموختگان ماست و وقت بسیاری را از ایشان به هدر می‌دهد. اکنون، در آغاز قرن بیست و یکم، با پیشرفت تکنولوژی این امکان فراهم آمده است که کتابخانه‌ای بزرگ را بتوان در حجمی بسیار اندک به اندازه یک کف دست جای داد و در کوتاه‌ترین زمان و با دقتی به کمال، به جست و جو در آنها پرداخت

متون زبان و ادب فارسی حاصل کوششهای طاقت فرسا، شور و شوق و نبوغ دانشمندان، فرهیختگان و ادیبان ایران بزرگ، در طول قرون، به عنوان میراثی گرانبها به ما رسیده است. میراثی که به تصدیق همگان به لحاظ غنا و عظمت و عمق و وسعت در ادبیات جهان بی نظیر است و سهم ایران و ایرانی را در پیشرفت علم، تمدن، فرهنگ و انسانیت ممتاز گردانیده است. این میراث عظیم چنانکه باید و شاید در دسترس فارسی‌زبانان و علاقه‌مندان این فرهنگ غنی، نیست. بسیاری از آنها با همت و کوشش محققان ارجمند در سده اخیر چاپ و منتشر شده و بسیاری دیگر هنوز در انتظار چاپ است. تنوع و فراوانی آثار چاپ شده و تولید و توزیع اندک شمار

پس از گذشت یک سال و منتفی شدن اقدام فرهنگستان در این خصوص، نگارنده در سال ۱۳۸۰ طرح ایجاد پایگاه متون را به شورای گسترش زبان و ادب فارسی وزارت فرهنگ و ارشاد اسلامی عرضه نمود. نظر موافق مدیر محترم شورای گسترش و پیگیری جدی ایشان خوشبختانه به نتیجه رسید و در پی موافقت وزیر محترم فرهنگ و ارشاد اسلامی و تصویب طرح، کار از تیرماه ۱۳۸۱ به سفارش شورای گسترش زبان و ادب فارسی آغاز شد.

بنابر طرح مصوب قرار است که تمامی کتابهای چاپ شده به زبان فارسی - در هر زمینه و موضوعی - که تاریخ تألیف آنها پیش از ۱۳۰۰ هجری شمسی باشد،^۲ در این پایگاه وارد شود، اما به دلیل زیاد بودن حجم اطلاعاتی که باید وارد نرم افزار شود (بیش از پانصد هزار صفحه)، ورود اطلاعات در چندین مرحله و به صورت دوره‌ای یا قرن به قرن انجام می‌شود و کتابهایی انتخاب می‌شوند که در دسترس هستند و فعلاً در این مرحله برای جست و جوی کتابهای مشمول طرح که احياناً در کشورهای دیگر چاپ شده‌اند و از آنها اطلاعی در دست نیست اقدامی نخواهد شد.

با توجه به موارد مذکور و نیز برای طراحی، تولید و آزمایش نرم افزار مورد نظر و رفع اشکالات احتمالی آن، در مرحله نخست تنها کتابهایی که تا نیمه قرن پنجم هجری قمری تألیف شده‌اند، انتخاب شد که بالغ بر ۲۰۰۰۰ صفحه می‌شود. این حجم برای آزمایش نرم افزار و نیز روشن شدن جزئیات، مناسب به نظر می‌رسید. انتخاب این کتابها (بهترین چاپ)، با توجه به نظر استادان زبان و ادب فارسی (مشاوران طرح) صورت گرفته است. فهرست این کتابها در پایان خواهد آمد.

اکنون که مرحله نخست طرح رو به اتمام است گزارشی از کار انجام شده و ویژگیهای نرم افزار تولید شده جهت آگاهی و اظهار نظر پژوهشگران و دست‌اندرکاران زبان و ادب فارسی تقدیم می‌شود: این پایگاه که علاوه بر استفاده عام، برای استفاده پژوهشگران زبان فارسی، ادبیات، تاریخ و دیگر رشته‌های علوم انسانی ایجاد خواهد شد می‌بایست در نهایت امر، پژوهش در زمینه‌های گوناگون فرهنگی به ویژه حوزه‌های زبان‌شناسی، دستور زبان، تاریخ زبان فارسی، سبک‌شناسی، نقد ادبی، فرهنگ‌نویسی (به ویژه فرهنگ تاریخی زبان فارسی)، آموزش زبان، واژه‌گزینی و تحقیقات ادبی را به نحو قابل توجهی تسهیل کند و ضریب اطمینان تحلیلها و استنتاجها را حتی المقدور افزایش دهد. بنابراین، نرم افزار مورد نظر به گونه‌ای طراحی شده است که علاوه بر امکان استفاده عمومی از آن، قابل استفاده پژوهشگران باشد. برای تأمین این منظور موارد زیر مورد توجه بوده است:

۱. ورود اطلاعات: در مرحله ورود اطلاعات، در متن هیچ گونه

تغییری داده نشده است. حتی غلطهای چاپی به همان صورت باقی مانده‌اند. شماره صفحات نیز مطابق با متن اصلی است و تنها متن کتابها تایپ شده‌اند و زیرنویسها، نسخه‌بدلها، مقدمه مصحح و ... تایپ نشده‌اند. این تصمیم ضمن اینکه برای صرفه‌جویی در وقت گرفته شده، به دلیل محدودیت در حروف یا به اصطلاح کاراکترهای موجود در نرم افزار بوده است. مثلاً برای 'ف' با دو نقطه، 'ب' که بالای آن نیز نقطه داشته یا 'ز' که زیر آن نقطه داشته باشد، 'س' با سه نقطه در زیر آن و ... کاراکتری موجود نیست و اگر برنامه‌نویسان طراحی هم بکنند از نظر مرتب کردن کلمات (یعنی تعیین جای این نشانه‌های جدید در جدول کاراکترها برای سورتینگ) و تهیه فهرستهای مورد نظر از حوزه واژگان متن، برنامه مختل خواهد شد. البته در آینده اگر این مشکلات برطرف شود امکان افزودن بخشهای حذف شده وجود خواهد داشت.

برای هر کتاب شناسنامه‌ای ایجاد شده که اطلاعات زیر را در بر دارد: نام کتاب، موضوع، نویسنده، مترجم، مصحح، تاریخ تألیف، محل تألیف، نوع اثر (نظم، نثر، داستانی، علمی و ...)، محل چاپ، ناشر و تاریخ چاپ که در نهایت می‌توان این اطلاعات را جست و جو کرد و نیز فهرستهای لازم از آن تهیه نمود.

۲. اطمینان از صحت اطلاعات وارد شده: اطمینان از درستی و صحت متنی که در نرم افزار مورد استفاده قرار می‌گیرد اهمیت بسیار زیادی دارد. ویژگی عمده و امتیاز این نرم افزار نسبت به کتاب و کتابخانه‌های معهود، در جست و جوهای آسان و سریعی است که می‌توان در آن انجام داد و نیز در گزارشها و فهرستهای متنوعی که می‌توان از متون تهیه کرد؛ بنابراین اگر اطمینانی به صحت متن وجود نداشته باشد، صحت گزارشها و نتیجه جست و جویها نیز مخدوش بوده و عملاً نرم افزار پایگاه متون به کاری نخواهد آمد. برای تأمین این منظور، از هر کتاب دو نسخه تهیه شد و نسخه‌ها جداگانه به دو گروه تالیست و واگذار شد که آنها را در محیط واژه‌پرداز مورد نظر با دستور العمل ویژه‌ای تایپ نمایند. پس از ورود اطلاعات، دو متن تایپ شده در رایانه، وارد محیط نرم افزاری شد که برای مقابله این دو متن تهیه شده است: به عبارت دیگر نمونه‌خوانی ماشینی صورت گرفته است. نرم افزار مورد نظر موارد اختلاف دو متن را تعیین می‌کند و کارشناس زبان و ادبیات فارسی با استفاده از کتاب، شکل غلط را حذف و عبارت درست را ابقا می‌کند یا کلمات و عبارات زاید را حذف و احياناً کلمات و عبارات جافا شده را تایپ می‌کند. در ابتدا نیز حجم اطلاعات وارد شده دو متن تایپ شده و شماره صفحات با هم و متن اصلی مطابقت داده شد و اگر در متون تایپ شده، افتادگی در حد پاراگراف و صفحه بود، تشخیص داده شد. پس از ویرایش و آماده‌سازی متن، مرحله‌ای دیگر برای تصحیح اطلاعات در نظر گرفته شده است، به

این ترتیب که متن آماده ورود در نرم افزار نهایی، ابتدا در نرم افزاری وارد می شود که برای جلوگیری از ورود کلماتی که در زبان وجود ندارد، تهیه شده است. کارشناس در این مرحله هر کلمه را فقط یک بار می بیند و آن کلمه و اردبانک کلمات نرم افزار می شود. پس از آن در موارد مشابه، نرم افزار آن کلمه را درست و در حوزه واژگان زبان فارسی تشخیص می دهد و وارد پایگاه متون می کند. به این ترتیب کلماتی که در زبان وجود ندارد و احیاناً هر دو تاییست مانند هم تایپ کرده اند و طبعاً نمونه خوانی ماشینی هم در این موارد بی فایده است، تصحیح می شود.

یک بار هم به وسیله نرم افزار، متن تبدیل به فهرست کلماتی می شود که بسامد آنها جلوی آنها درج شده است، کلماتی که یک یا دو بار و سه بار در متن به کار رفته اند بازبینی می شوند، زیرا کلماتی که بسامد اندکی دارند احتمال اینکه غلط باشند بیشتر است. متن تصحیح شده که باید همانند کتاب اصلی باشد در نرم افزار ذخیره می شود. مرحله ای دیگر برای تصحیح وجود دارد که بعداً ذکر خواهد شد.

۳. از کتاب ذخیره شده در نرم افزار، یک نمونه چاپی تهیه و برای ویرایش و آماده سازی متن به کارشناسان زبان و ادب فارسی سپرده می شود.

ویرایش و آماده سازی متن شامل موارد زیر است:

۱. **تصحیح غلطهای چاپی و غیر چاپی** (به ویژه با استفاده از یادداشتهای کارشناسان مرحله ۶ و مقالات تقدیمتون مورد نظر که تا زمان اجرای این طرح به چاپ رسیده، گردآوری و کارشناسی شده است).

غلطهای غیر چاپی که شامل بدخوانی و تشخیص نادرست مصحح و... می شود با حفظ هر دو عبارت اصلی (غلط) و تصحیح ناقد (شکل صحیح) و با تأیید استادان مشاور طرح، جای یکی در متن و دیگری در حاشیه تعیین می شود.

۲. **تنظیم فاصله ها** (در این برنامه، رایانه هر حرف یا مجموعه حروف را که بین آنها فاصله نباشد یک کلمه تشخیص می دهد. یعنی دو طرف هر کلمه باید یک فاصله داشته باشد.) بنابراین باید این کار با دقت و حوصله انجام شود.

اجزایی از کلمه که به لحاظ نحوی مستقل اند مانند: ضمائر متصل، یای نکره، الف ندا و... (پی چسبها) که توسط کارشناسان کدگذاری شده اند (بند ۶)، در نرم افزار قابل شناسایی اند و طبعاً قابل جست و جو و...

۳. **نشانه گذاری:** نقطه، نشانه پرسش، دو نقطه، نشانه تعجب و... این کار عملاً یعنی تعیین جملات و بعد تعیین هویت جملات و در نتیجه تسهیل جست و جو و فهرست گیریهای گونه گون. در



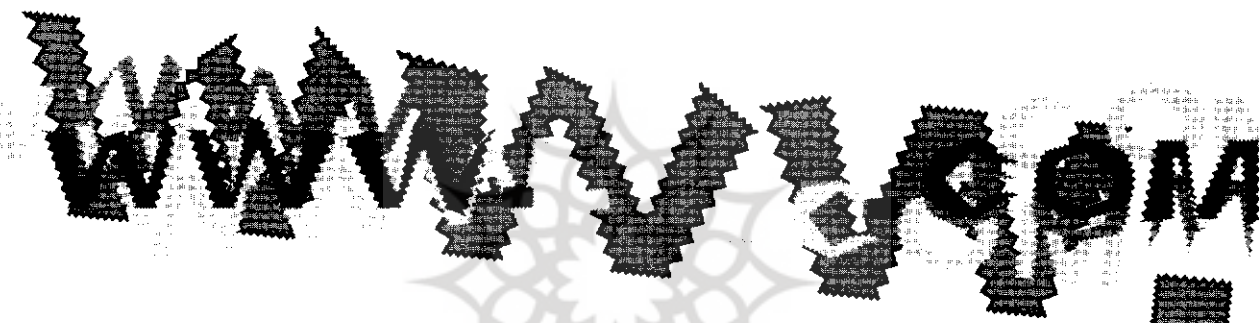
این صورت ماقادر خواهیم بود تمام جملات پرسشی و... را از متن انتخاب کنیم، به دلخواه مرتب کنیم و در آنها به جست و جو پردازیم. دو نقطه (: نشانۀ گفت و گو است و با تعیین آن می توان فقط در گفت و گوهای یک متن به جست و جو پرداخت یا از آنها فهرستهایی به دلخواه تهیه کرد.

(در متون منظوم و نیز منثور، بیت یک واحد به شمار می آید و) نیز مصراع، البته اگر تک بود) و جملات داخل آن با نقطه تعیین نمی شود، ولی نشانه های پرسش، تعجب، امر، گفت و گو و... را خواهد داشت.

۴. حرکت گذاری کلمات در متن فارسی: به منظور ایجاد امکان جست و جوی دقیق و نیز امکان فهرست گیری و ایجاد تمایز بین کلمات مشابه در املا و متفاوت در تلفظ و معنا، این کلمات حرکت گذاری می شوند. کلماتی مانند: بُرد، کند که چهارگونه خوانده می شوند، یا کلماتی مانند: بین، شوید، جو، حکم، خرد، درد که سه گونه خوانده می شوند،^۳ در صورتی که بدون حرکت

نخانه ای و نخانه یی انتخاب و به ترتیب جایگزین نخانه نی و نخانه شد. یا کلماتی از قبیل دَرین، برین، ازین، دران، بران، ازان، به شکل کامل آن و با فاصله: دَر این، بر این، از این، در آن، بر آن، از آن تبدیل شد که حتی اگر در شعر هم باشد خللی ایجاد نخواهد کرد. این تغییر ضمناً مانع از بدخوانی و نیز تسهیل در جست و جو خواهد شد. مثلاً برین به معنی برترین یا برین به معنی روی این یکی نمی شود.

پس از انجام این موارد روی کاغذ و اعمال اصلاحات در فایل های ذخیره شده، یک نمونه چاپی دیگر از متن نهایی تهیه می شود و در اختیار گروه بازبینی نهایی قرار می گیرد.^۵ سپس اصلاحات نهایی در متون ذخیره شده اعمال می شود و متن ها برای اعمال کد گذاری (بند ۶) وارد نرم افزار می شوند. (همه اصلاحات موارد پنجگانه در کپی فایل های اصلی اعمال می شود و نسخه ای از فایل های اصلی برابر با اصل کتابها جداگانه ذخیره و نگهداری می شود تا در صورت لزوم دسترسی به فایل های اصلی ممکن باشد.)



۶. کدگذاری کلمات و عبارات به منظور ایجاد محیطی مناسب در متون برای جست و جوهای تخصصی پژوهشگران و استادان و نیز تهیه بانکهای مورد نیاز اقشار مختلف کاربران:

۱- ۶. تعیین عبارات عربی (آیات قرآنی، احادیث، سخنان بزرگان، امثال و حکم، مصراع، شعر)

۲- ۶. تعیین عبارات غیر فارسی و عربی "زبانها و لهجه ها" (سخنان بزرگان، امثال و حکم، مصراع، شعر)

۳- ۶. تعیین عبارات فارسی: مصراع و شعر در متون نشر، ترجمه آیات قرآنی، ترجمه احادیث، سخنان بزرگان، امثال و حکم، هر کدام با کد جداگانه.

۴- ۶. کدگذاری پای نکره، الف ندا و ضمائر منفصل و متصل (منظور از ضمیر متصل در این طرح، هر ضمیری است که به کلمه ای چسبیده باشد مانند: "و در کو" (= که او); این کار برای تسهیل و دقت هر چه بیشتر در جست و جو و تهیه آمار واژگانی و... انجام شده است.

۵- ۶. کدگذاری فعل. تمام فعلهای هر متن روی کاغذ کد گذاری شده است. پس از ورود اطلاعات (متون) بانکی از افعال کلیه متون موجود در نرم افزار، تهیه می شود. در این بانک افعال که قاعدتاً به تدریج کامل تر می شود و شامل تمام ساختهای افعال اعم از زمانی و غیر زمانی خواهد شد، ساختهای گونه گونه هر فعل ذیل مصدر خود قرار می گیرد و در جست و جو نیز به آن مربوط می شود، مثلاً اگر کاربر رفتن را در یک متن جست و جو کند، تمام

ثبت شوند، با توجه به حجم بسیار زیاد متون (بالغ بر پانصد هزار صفحه)، عملاً کار جست و جو و فهرست گیریهای فایده می شود. رفع مشکلات جست و جو برای کاربری که در مورد کلمات مشابه حضور ذهن ندارد نیز در نرم افزار پیش بینی شده است بدین صورت که هر گاه کاربر کلمه ای را برای جست و جو و گزارش گیری تعیین کرد، اگر آن کلمه مشابه داشته باشد، کلمات مشابه با حرکت های لازم در صفحه نمایشگر ظاهر می شود تا کاربر بتواند فقط کلمه مورد نظر یا همه کلمات یا هر تعداد از آنها را که خواست به دلخواه انتخاب نماید تا از نرم افزار گزارش صحیح و دلخواه خود را دریافت کند.^۴

۵. تصحیح رسم الخط و یکدست سازی در موارد خاص برای تسهیل در امر جست و جو. رسم الخط متون کهن فارسی در نسخه های چاپی نیز بسیار متفاوت است. این تفاوتها تا جایی که در کار جست و جو خللی ایجاد نکنند ایرادی ندارد و ضروری است که به جهت حفظ اصالت متن حتی الامکان حفظ شود، اما در این متون نخانه ای را به صورت نخانه یی، نخانه نی و نخانه نیز نوشته اند، برای جست و جوی این کلمه باید چهار بار جست و جو صورت گیرد یا هر چهار گونه برای جست و جو انتخاب شود. فهرست حاصل جست و جو نیز بی جهت طولانی خواهد شد و به خصوص اگر موردی باشد که گونه ها از نظر ترتیب الفبایی تفاوت فاحش داشته باشند، بررسی گزارش دشوار خواهد شد. در مورد مذکور برای حفظ اصالت متن (تا حد ممکن)، با توجه به تلفظ، دو گونه

ساختهای فعلی رفتن را (در صورت لزوم در جمله) با قید شماره صفحه و نام کتاب می‌تواند ببیند و به هر صورت که خواست (کپی و پرینت) در اختیار داشته باشد. این امکان اساساً برای تأمین نیاز فرهنگ‌نویسان ایجاد شده است، یعنی معادل فیش نویسی روی کاغذ است که ذیل مدخل، شاهد رومی نویسند. در مورد فعلها چون شکل مصدری فعل، در فرهنگ مدخل می‌شود، امکان تطبیق آن با ساختهای فعلی در رایانه به طور معمول وجود ندارد و رایانه نمی‌تواند ارتباط برود، رفته بود، می‌روم، برو و... را با رفتن تشخیص دهد. این کار، یعنی کدگذاری افعال، کلمات مشابه با آنها را - که فعل نیستند - نیز از آنها متمایز می‌کند و برای تسهیل در جست و جو مفید است. مثلاً بُرد (فعل) و بُرد (نوعی پارچه)، بُرید (فعل امر) و بُرید (پست)، نهشت (گذاشت) و نهشت (جنت) از این طریق از هم متمایز می‌شوند.

- تعیین هویت‌های دیگر نیز ممکن است، اما به جهت وقتگیر بودن، در این مرحله از آنها صرف نظر شده، ولی توجه شده است که امکانات آن در نرم افزار فراهم باشد که در صورت لزوم بعدها بتوان کدهای لازم دیگری را برای هویت‌های دیگر افزود. هویت‌های از قبیل: اسم، صفت، صفت برتر، صفت برترین، صفت عددی، تمیز عددی و واحد شمارش، قید، صوت، جمع مکسر، اسم جمع، کلمات پرسشی، نام اشخاص، نام مکانها، نام کتابها و...^۶

موارد پنج گانه فوق (بند ۶) روی همان دو نسخه متن اصلی کتابها (که برای تایپ تهیه شده بود) انجام گرفته است و در متون آماده سازی شده (بند های ۱ تا ۵) و وارد شده در نرم افزار، اعمال می‌شود. این مرحله یعنی اعمال کدگذاریهای انجام شده روی کاغذ در نرم افزار که توسط کارشناسان زبان و ادبیات فارسی صورت می‌گیرد، آخرین مرحله تصحیح (پیش از عرضه نرم افزار) نیز خواهد بود، بدین صورت که کارشناسان مذکور در مطابقت متن اصلی و کلمات و عبارات کدگذاری شده آن با متن موجود در نرم افزار، خواه ناخواه با غلطها و افتادگیهای احتمالی مواجه خواهند شد - که گاهی آنرا هر دو تایپیست مانند هم عمل کرده اند^۷ - و به تصحیح آنها اقدام خواهند کرد.

امکانات نرم افزار

۱. قابلیت ایجاد بانک واژگان از همه متون ذخیره شده در نرم افزار یا متون انتخابی و مرتب کردن آنها به روشهای زیر:

الف) فهرست کردن الفبایی کل واژگان متن یا متون انتخابی بر اساس آغاز واژه همراه با درج شماره بسامد آن در مجموعه مورد جست و جو به دو صورت: از الف تا ی و از ی تا الف به هر منظوری مثلاً تعیین مدخلهای یک فرهنگ عام یا خاص، بزرگ یا کوچک و...، در اختیار داشتن تمام کلمات زبان از آغاز تا ۱۳۰۰ برای واژه گزینی و نام گذاری در موارد عام و خاص، علمی و غیر علمی. ب) فهرست کردن الفبایی کل واژگان متن یا متون انتخابی بر اساس پایان واژه همراه با درج شماره بسامد آن در مجموعه مورد جست و جو به دو صورت: از الف تا ی و از ی تا الف به منظورهای مختلف مثلاً شناختن پسوندهای زبان فارسی.

ج) فهرست بسامدی کل واژگان متن یا متون انتخابی به دو صورت صعودی و نزولی (از کم بسامدترین به پر بسامدترین و

برعکس) که کاربردهای بسیاری خواهد داشت مثلاً انتخاب پربسامدترین واژگان در زبان فارسی با رعایت مسائل مربوط، برای تألیف کتابهای آموزش زبان فارسی به کودکان، نوجوانان، جوانان، کم سوادان، غیر فارسی زبانان و نیز تعیین مدخلهای فرهنگهای یک زبان به یاد دوزبانه برای گروههای یاد شده.

۲. امکان شمارش کل کلمات هر متن یا همه متون انتخابی و شمارش کلمات (با حذف بسامد آنها) برای تعیین حوزه کمی واژگان هر متن و در نهایت واژگان زبان فارسی در هر دوره یا از آغاز تا ۱۳۰۰ هجری به منظور تحلیل و بررسیهای زبانی و... مثلاً تعیین تفاوت دو کتاب در نسبت حوزه واژگان (بدون بسامد) به کل واژگان متن و بررسی تفاوت‌های دیگری که می‌توان آنها را ناشی از این تفاوت تلقی کرد.

۳. امکان مشاهده، کپی و تهیه پرینت از بیت یا جمله ای که واژه مورد جست و جو در آن به کار رفته با درج شماره صفحه و نام متن یا کدی که برای آن در نظر گرفته شده است.

۴. امکان مرتب سازی بیتها یا جمله های حاوی واژه مورد جست و جو به طرق مختلف و مورد نیاز پژوهشگر، مثلاً مرتب سازی به ترتیب تاریخ تألیف متون از قدیم به جدید و برعکس، الفبایی، نظم یا نثر بودن خواهد و...^۸

۵. امکان وصل شدن از جمله در فهرست جمله ها (بند ۴) به همان جمله در متن اصلی و ادامه جست و جو.

۶. امکان انتخاب یک واژه از درون متن اصلی برای مشاهده موارد دیگر کاربرد آن (ادامه جست و جو) در همان متن و دیگر متون انتخابی کاربر.

۷. امکان جست و جو بر اساس ساخت پایه مصدر برای فعلها که از این طریق مثلاً می‌توان دریافت که در متن مورد نظر چه ساختهایی از یک مصدر وجود دارد و بررسی تحولات ساختها در تاریخ زبان فارسی؛ و مهم تر از آن در اختیار داشتن شواهد برای هر مصدر که عموماً به شکل فعل در جمله به کار می‌رود و تسهیل و تسریع در کار فرهنگ نویسی.

۸. امکان تبدیل متن به ابیات و جملات متوالی (هر جمله از ابتدای سطر آغاز شود) و امکان مرتب کردن این جملات یا ابیات به صورت الفبایی بنا بر آغاز جمله یا بیت و پایان جمله^۹ یا بیت. این امکان در مورد ابیات کاربرد خیلی روشنی دارد؛ با ترتیب الفبایی به سادگی ابیات مشابه در متون مختلف را می‌توان کنار هم دید و تاریخ اولین بار ثبت بیت در متون را دریافت و گاهی نسبتهای مشکوک را تشخیص داد و به گوینده واقعی بیت رسید. از سوی دیگر تفاوت‌های ضبط بیت در متون مختلف و یافتن قدیم ترین کاربرد آن در تصحیح متون (نسخه های خطی) نیز مفید خواهد بود و می‌توان تصور کرد ارتباط این مرتب سازی را با مقوله های سبک شناسی، تاریخ زبان فارسی، فرهنگ تاریخی زبان فارسی و نیز تاریخ اندیشه در زبان فارسی^۹ و... در مورد جملات نیز مرتب سازی کم و بیش فوایدی خواهد داشت مانند بررسی نسبت وجود عناصر دستوری در آغاز یا پایان جملات در سبکهای مختلف، دوره های مختلف، موضوعات مختلف، انواع مختلف ادبی (حماسی، غنایی، تعلیمی و...).

۹. امکان جست و جوی جملات پرسشی، تعجبی و امری در کلیه متون یا متون انتخابی پژوهشگر و امکان مرتب کردن یافته ها به

صورت الفبایی بنا بر آغاز جمله یا بیت و نیز پایان جمله یا بیت یا بنا بر تاریخ تالیف متون.

۱۰. امکان جست و جوی گفت و گوهای کلیه متون یا متون انتخابی پژوهشگر و امکان مرتب سازی فهرست به دست آمده بنا بر ترتیب تاریخ تالیف متون یا ترتیب الفبایی.

۱۱. امکان وارد کردن معنی لغات. می توان بعدها از این امکان استفاده کرد و پس از وارد کردن معنی لغات بدون هیچ کار اضافه ای، این لغتها و معنیها را با مثال و شاهدشان به صورت یک فرهنگ، در نرم افزار در اختیار داشت.

امکانات دیگر

با توجه به بند ۶ و ویرایش و آماده سازی متن، پس از کامل شدن اطلاعات وارد شده (حدود هزار عنوان کتاب^{۱۱} در بیش از پانصد هزار صفحه) مجموعه های باارزش زیر را نیز در اختیار خواهیم داشت که به راحتی قابل تبدیل به کتاب و لوحهای فشرده مجزای نیز خواهند بود:

فرهنگ جامع امثال و حکم فارسی

فرهنگ جامع سخنان بزرگان (به فارسی)

مجموعه ترجمه احادیث

فرهنگ اقوال عربی در متون فارسی

فرهنگ احادیث عربی به کار رفته در متون فارسی

فرهنگ جامع اشعار (و مصراعهای) عربی به کار رفته در متون

فارسی

فرهنگ امثال و حکم عربی به کار رفته در متون فارسی

فرهنگ جامع اشعار فارسی به کار رفته در متون نشر فارسی (که

بسیاری از آنها در هیچ دیوانی یافت نمی شود).



حوزه و وظایف خود نمی بیند^{۱۲} انتظار می رود مسئولان نهادهایی که به نحوی سازمان خود را مربوط به این طرح می بینند به این امر توجه نمایند. به سرانجام رساندن این طرح که مورد تأیید شفاهی و کتبی استادان طراز اول ادبیات و زبان شناسی این مرز و بوم است و هم با اشراف و راهنمایی ایشان آغاز شده و تا کنون ادامه یافته است،^{۱۳} باعث سربلندی و افتخار، رشد فرهنگی، علمی و ادبی جامعه، حفظ میراث مکتوب، توسعه، تثبیت و تعمیق فرهنگ ایرانی - اسلامی و معرفی آن به جهانیان با رایج ترین و آسان ترین وسیله ممکن (لوح فشرده، اینترنت)، گسترش زبان و ادب فارسی در جهان، بازیابی هویت فرهنگی - تاریخی، معرفی بزرگان ایران اسلامی در طول تاریخ، کمک قابل توجه به مراکز و نهادهای پژوهشی کشور و جلوگیری از اتلاف وقت پژوهشگران، جلوگیری از اتلاف سرمایه و صرفه جویی در هزینه های پژوهشی و... خواهد بود و هر نهاد یا فردی که عهده دار به انجام رساندن این طرح شود بی شک نام خویش را جاودان کرده است. همین جا باید تصریح کنم که به عنوان طراح و مجری مرحله نخست این طرح، مطمئن ام هر کس دیگری

که ادامه این کار را به عهده بگیرد از آنجا که به تنهایی کاری از پیش نخواهد برد و ناگزیر باید از راهنمایی و مشورت استادان و اهل فن استفاده کند، نتیجه یکی خواهد بود و من هم تجربه ناچیز خود را با کمال میل در اختیار خواهم گذاشت. به گمان من ادامه این طرح نیاز به ایجاد یک بنیاد علمی دارد که بتواند با استفاده از امکانات و نیروی انسانی و علمی جامعه دانشگاهی این کار را سامان دهد و به همه جوانب آن بپردازد. طبیعتاً ادامه این طرح به تألیف فرهنگ جامع زبان فارسی، فرهنگهای تخصصی، کتابهای آموزش زبان فارسی، تحقیقات بنیادی زبان شناسی و... منجر خواهد شد که هر کدام خود طرحی مفصل و گسترده اند. نهادهایی مانند وزارت علوم، وزارت ارشاد، بنیاد ایران شناسی و به خصوص دانشگاه فرهنگستان، پژوهشگاه علوم انسانی و دیگر نهادهای عهده دار فرهنگ نویسی که در گیر این نوع مسائل هستند، بهترین گزینه ها برای ادامه این طرح می توانند باشند. امر خطیری است و حساسیت بسیار می طلبد، امیدوارم که توانسته باشم در این گزارش حق مطلب را ادا کنم.

و همگی با درج تفاوتها و ضبطهای مختلف آنها و... و ناگفته پیداست که با تولید این اثر چه امکانات وسیعی برای پژوهشهای زبانی، ادبی، تاریخی، اجتماعی و... ایجاد خواهد شد و ضمن بی نیاز ساختن ما از کارهای پراکنده و به ناچار ناقص در این زمینه ها سبب می شود که پس از این دیگر عمر شریف محققان این سرزمین صرف امور وقتگیری از این دست نگردد و دوران پژوهشهای اساسی کارگشا و نقد و تحلیل آغاز شود و شاهد تحقیقات علمی جدی تری در علوم انسانی به خصوص در حوزه زبان و ادبیات فارسی باشیم.^{۱۱} اکنون با توجه به اینکه بانی طرح، شورای گسترش زبان فارسی، ادامه طرح در مراحل بعدی و حمایت از آن را در توان یاد

فهرست متون قرن چهارم تا نیمه قرن پنجم هجری

۱. آثار منظوم رودکی
۲. تاریخ سیستان (قسمت اول)
۳. ترجمه السواد الاعظم
۴. تاریخنامه طبری
۵. ترجمه تفسیر طبری
۶. ترجمه رساله حی بن یقظان
۷. ترجمه قرآن موزه پارس
۸. تفسیر قرآن پاک
۹. تفسیر قرآن کمبریج
۱۰. تفسیری بر عشری از قرآن مجید
۱۱. حدود العالم من المشرق الی المغرب
۱۲. اشعار پراکنده قدیم ترین شعرای فارسی
۱۳. دانشنامه علایی
۱۴. دانشنامه میسری
۱۵. اشعار باباطاهر همدانی
۱۶. دیوان دقیقی طوسی
۱۷. دیوان عسجدی مروزی
۱۸. دیوان عنصری بلخی
۱۹. دیوان فرخی سیستانی
۲۰. دیوان منوچهری دامغانی
۲۱. رگ شناسی یا رساله در نبض
۲۲. زمین الاخبار گردیزی
۲۳. اشعار کسایی مروزی
۲۴. سخنان منظوم منسوب به ابو سعید
۲۵. (شرح احوال و) اشعار شاعران بی دیوان
۲۶. شاهنامه فردوسی
۲۷. کشف المحجوب سجستانی
۲۸. لسان التنزیل
۲۹. مقدمه شاهنامه مشهور ابو منصور
۳۰. وامق و عذرا (عنصری)
۳۱. ویس و رامین
۳۲. هدایه المتعلمین فی الطب
۳۳. ترجمه آهنگین از دو جزء، قرآن مجید
۳۴. الابنیه عن حقائق الادویه
۳۵. رساله جودیه
۳۶. کنوز المعزمین
۳۷. معیار العقول
۳۸. تاریخ بلعمی
۳۹. التئور
۴۰. پندنامه ماتریدی
۴۱. التفهیم لا وائل صناعه التنجیم
۴۲. بازنامه نسوی
۴۳. بخشی از تفسیری کهن
۴۴. بخشی از تفسیری کهن به پارسی
۴۵. تاریخ بیهقی

پانوشتها:

۸. منظور از جمله در این طرح، در حال حاضر لزوماً به معنی دستوری جمله نیست، بلکه جمله تاجایی که مقصود مورد نظر کاملاً بیان شود ادامه خواهد داشت، یعنی کامل بودن شاهد برای واژگان فرهنگ مد نظر بوده است. البته امکان تعیین جمله دستوری نیز وجود دارد که در صورت لزوم، در مراحل بعد می توان بدان پرداخت.
۹. تذکر جناب آقای دکتر غلامحسین صدری افشار.
۱۰. تاکنون نزدیک به ۹۰۰ عنوان کتاب شناسایی و فهرست شده است و برآورد می شود که این رقم با توجه به محدوده ای که پیش از این تعیین شد، از ۱۰۰۰ عنوان بیشتر نشود.
۱۱. در باب مسائل حقوقی این طرح نیز راهکارهایی اندیشیده شده است که حقوق مادی و معنوی مصححان محترمی که کتابشان در این طرح مورد استفاده قرار می گیرد، محفوظ بمانند. نپرداختن بدان در این مقاله به این جهت است که نحوه عرضه این نرم افزار که در این مورد تعیین کننده است، هنوز معلوم نیست.
۱۲. مرحله نخست طرح نیز تاکنون به دلیل نبود امکانات به اتمام نرسیده است. اگر امکانات فراهم شود این مرحله حداکثر تا دو ماه دیگر به پایان خواهد رسید، ولی با ادامه وضع کنونی شاید بیش از ده ماه طول بکشد. (مشکل اصلی کمبود رایانه است)
۱۳. استادان محترم عبارت اند از: دکتر حسن انوری، دکتر تقی پورنامداریان، دکتر علی محمد حق شناس، دکتر حسین سامعی، دکتر محمدرضا شفیعی کدکنی، دکتر علی اشرف صادقی، دکتر پریوش صفا و دکتر مصطفی عاصی. با تشکر از حمایتها و راهنماییهای بی دریغشان به خصوص آقای دکتر سامعی که با دقت و حوصله فراوان و توجه به جزئیات، نکات مهمی را یادآوری کردند.

۱. کتابخانه الوراق بانسانی WWW.ALWARAQ.COM
۲. از آنجا که طرح پایگاه داده های زبان فارسی در پژوهشگاه علوم انسانی به مدیریت آقای دکتر مصطفی عاصی شامل متون تألیف شده پس از ۱۳۰۰ می شود، برای پرهیز از دوباره کاری و صرف هزینه مجدد، متون پس از این تاریخ از این طرح کنار گذاشته شد.
۳. فهرستی از این کلمات (مشابه در املا و متفاوت در تلفظ و معنی: تاکنون ۲۵۰ مورد) و (مشابه در املا و تلفظ و متفاوت در معنی: تاکنون ۹۰ مورد) تهیه شده که به تدریج کامل می شود. برای کاربرد این فهرست ← پانوشت بعد.
۴. به این موارد گونه های اسم نیز افزوده می شود. مثلاً با انتخاب آنوشیروان برای جست و جو، تمام گونه های دیگر این نام که در متون موجود است (آنوشروان، انوشه روان، نوشروان، نوشیروان، نوشین روان و...) در صفحه نمایشگر ظاهر می شود تا کاربر ضمن اشراف بر گونه های دیگر این نام، جست و جوی خود را کامل و با دقت به انجام برساند.
۵. این گروه کارشناسان متخصص زبان و ادب فارسی (فارغ التحصیلان دوره های کارشناسی ارشد و دکتری) هستند.
۶. اینجایاباید از کوشش، دقت نظر و همدلیهای مهندسان گروه تحقیق در عملیات که مسئولیت برنامه نویسی و پشتیبانی فنی این نرم افزار را بر عهده دارند صمیمانه تشکر کنم که در ایجاد این امکانات بسیار مساعدت کرده اند.
۷. طبیعی است که در مواردی دو تاپیست مانند هم اشتباه کنند، مانند مواردی که در دو سطر متوالی یا با فاصله کلمات مشابهی باشد و هنگام تاپ، چشم از روی سطر نخست به سطر مشابه بعدی بلغزد و یک یا چند