Iranian Association of Information and Communication Technology
IKT 2013
Shiraz University
Regular Paper

ایجاد و انتشار زیر ساخت وب معنایی برای قرآن کریم

احمد استیری، محسن کاهانی و هادی قائمی

آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد، ahmad.estiri@stu.um.ac.ir
آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد، kahani@um.ac.ir
آزمایشگاه فناوری وب، دانشگاه فردوسی مشهد، hadi.qaemi@stu.um.ac.ir

چکیده

زبان‌شناسی رایانه‌ای در سال‌های اخیر به یکی از دغدغه‌های اساسی محققان و پژوهشگران حوزه کامپیوتر و زبان‌شناسی تبدیل شده است. استفاده از رایانه و ابزارهای هوشمند باعث شده است که بتوان بسیاری از کارهای مرتبط با پردازش متن را با سرعت و دقت قابل توجهی انجام داد. پردازش زبان طبیعی در حوزه متن به پردازش پیکره‌های متنی به عنوان ابزاری برای بیان ویژگی‌های زبان می‌پردازد. پیکره‌های متنی در واقع نمادی از زبان هستند که با هدف خاصی تولید گردیده، می‌توان با تحلیل آنها به استخراج اجزا، قواعد و ساز و کار زبان پی برد و در مرحله بعد، با فرآوری و غنی‌سازی متون و با بکارگیری فناوری‌های رایانه‌ای، محیط پژوهشی مناسبی را در ارائه‌ی محتوای این متون به گونه‌ای کارآمد ایجاد نمود.

پیکره متنی و زیرساختی که تحت عنوان پیکره‌ی "فرقان" برای قرآن کریم تولید گردیده، حاصل بهره‌گیری از سامانه‌ای هوشمند است. این پیکره با بیش از ۵۸۷ مگابایت داده، حاوی کلیه‌ی اطلاعات قرآنی، آماری، متن و ترجمه فارسی و انگلیسی آیات و برچسب-گذاری صرفی و نحوی متن عربی، فارسی و انگلیسی آیات، ریشه‌یابی کلمات آنها و بسیاری موارد دیگر در قالب RDF است و امکان استفاده و کاوش را برای هرگونه پژوهش و پردازش هوشمند ایجاد کرده است.

کلید واژه

پردازش زبان طبیعی، پیکره، وب معنایی، قرآن کریم، RDF.

۱- مقدمه

زبان‌شناسی رایانه‌ای از پویاترین شاخه‌های علم زبان‌شناسی و علم کامپیوتر در جهان است که متاسفانه در کشور ما از دیگر حوزه‌های این علوم، ناشناخته‌تر بوده و کمتر به آن پرداخته شده است.

بدون تردید یکی از اهداف زبان‌شناسی رایانه‌ای که از ابتدای تکوین و گسترش این علم مد نظر بوده و پیوسته دنبال شده است، ارائه ابزارهایی برای تحلیل بر روی متون یا پیکره‌های زبانی[۱] می‌باشد. ورود رایانه به علم زبان‌شناسی، امکان وارد شدن به عرصه‌هایی که حتی تصور آنها نیز مشکل بود، فراهم کرده است.

زبان‌شناسی رایانه‌ای در حوزه متن به پردازش پیکره‌های متنی به عنوان نمادی از زبان می‌پردازد. پیکره‌ی زبانی[۲] عبارتست از مجموعه‌ای از متن‌های نوشتاری یا گفتاری آوانویسی شده که می‌توان در توصیف و تحلیل زبان از آن بهره گرفت [1]. در تعریف دیگری برای پیکره آمده است: "حجم زیادی از داده‌های زبانی که بر اساس معیارهای مشخص شده برای هدف معینی جمع‌آوری و ذخیره شده‌اند؛ بطوریکه نماینده‌ی زبان یا گویش مورد مطالعه باشد. [2]"

یکی از مهم‌ترین مسائلی که در تولید یک پیکره باید مورد توجه قرار گیرد، ویژگی‌ها و خصوصیات خاص آن پیکره است به گونه‌ای که حاوی خصوصیاتی باشد که نیل به اهدافی را که در تولید پیکره، مد نظر بوده است میسر سازد.

پیکره‌های متنی با اهداف متفاوتی نظیر تجزیه و تحلیل‌های آماری، سنجش صحت فرضیه‌ها و قواعد زبانی ، بررسی رخدادها، استخراج اطلاعات و دانش نهفته در متن و موارد مشابه، در حوزه‌ای مشخص تولید شده و بکار گرفته می‌شوند. پیکره‌های متنی به عنوان پایگاه دانش[۳] اصلی در زبان-شناسی رایانه‌ای خصوصا در حوزه خط و زبان شناخته می‌شوند. پیکره‌های متنی می‌توانند تک‌زبانی[۴] و یا چندزبانی[۵] باشند.

به منظور ساخت پیکره‌های متنی مناسب برای کاربردهای مفیدتر در مطالعات مختلف زبان‌شناسی، پیکره‌ها می‌باید پرورش یافته و به طور مناسبی غنی‌سازی گردند و به عنوان مثال حاشیه‌نویسی[۶] شوند که بدین صورت تعریف می‌گردد: تحلیل و افزودن برخی اطلاعات مانند نقش[۷]،ریشه[۸] و یا سایر ویژگی‌های کلمات موجود در متن به پیکره و یا سایر اطلاعاتی که می‌تواند در استفاده صحیح از پیکره به کاربران کمک نماید.

از اساسی‌ترین چالش‌های پیش‌رو در تولید پیکره‌های متنی این است که چه پیکره‌ای و با چه حجم و خصوصیاتی می‌تواند به بهترین شکل ممکن، بیانگر خصوصیات اصلی زبان باشد. امروزه حجم زیادی از مطالعات

حوزه‌ی زبان‌شناسی رایانه‌ای، تنها با استفاده از پیکره‌های زبانی میسر می‌گردد. به همین دلیل می‌توان پیکره‌های متنی را یکی از اساسی‌ترین دادگان ورودی در زبان‌شناسی رایانه‌ای قلمداد نمود. در حال حاضر، بکارگیری، بررسی، تجزیه و تحلیل و پردازش پیکره‌های استاندارد مختلف و ارائه‌ی نتایج بدست آمده، موضوع پژوهش‌ها و مقاله‌های متعددی در حوزه‌ی زبان‌شناسی رایانه‌ای می‌باشد.

در پیکره فرقان، حجم عظیمی از دادگان قرآنی به صورت ساختارمند در قالب RDF گردآوری شده‌اند و بستر مناسبی جهت پیاده‌سازی فناوری‌های کشف دانش‌های نهفته در قرآن از طریق ایجاد ابزارهای پردازش هوشمند متن فراهم گردیده است.

ساختار این مقاله به شرح زیر می‌باشد. در بخش دوم ادبیات موضوع و در بخش سوم توضیحات مربوط به روش بکار رفته در تولید پیکره‌ی قرآنی "فرقان" ارائه خواهد شد. در پایان نیز نتیجه‌گیری و کارهای آینده ذکر گردیده است.

۲- ادبیات موضوع

پیکره‌ی زبانی می‌تواند بسیار بزرگ، فراگیر و نماینده‌ی تمامی یک زبان و یا گونه‌ای از آن باشد؛ به شکل برگه‌های یادداشت یا پرونده‌های رایانه‌ای شامل متن‌های کامل یا گزیده‌ای از آنها، بخش‌های پیوسته‌ای از متون یا گزیده‌ای از نقل قول‌ها و نکات و حتی فهرست‌های واژگانی باشد. پیکره می‌تواند جهت بررسی در زمینه‌ای خاص، فراهم شود و یا دربرگیرنده‌ی مجموعه‌ی عظیم و بی‌ساختاری از متون مختلف باشد که برای منظورهای گوناگون بکار گرفته شود [۳].

با اهمیت یافتن پیکره‌های زبانی در عصر حاضر، بهره‌گیری از داده‌های واقعی زبانی به شدت گسترش یافته و جزو مهم‌ترین ملزومات بسیاری از مطالعات و پژوهش‌های حوزه زبان‌شناسی نظیر نظریه‌پردازی، توصیف ساختمان زبان، گویش‌شناسی، دستورنویسی، فرهنگ‌نگاری و همچنین ارائه ابزارهای کاربردی گوناگون نظیر خلاصه‌سازها، مترجم‌های ماشینی و موارد مشابه قرار گرفته است. امکاناتی که با ظهور رایانه در عرصه‌ی زبان‌شناسی فراهم گردید باعث پیدایش شاخه‌ای تخصصی در حوزه زبان‌شناسی رایانه‌ای به نام زبان‌شناسی پیکره‌ای[۹] گردید.

اصطلاح پیکره را به ویژه زبان‌شناسان ساخت‌گرا بکار می‌بردند و همواره تاکید می‌کردند که توصیف یک زبان یا گویش باید مبتنی بر داده‌های گردآوری شده و تحلیل این داده‌ها باشد و با آنکه فراگیری و بزرگی پیکره، عامل تعیین کننده‌ای در افزایش دقت و اعتبار نتایج به شمار می‌آمد؛ اما محدودیت‌های نیروی انسانی و زمان که در برابر این عامل، قرار داشت باعث می‌شد اکثرا به نمونه‌های برگزیده اکتفا شود [۳].

در آغاز دهه ۱۹۸۰ تعداد پیکره‌های الکترونیکی موجود در دنیا انگشت‌شمار بود. ولی اکنون شاید بیش از صدها پیکره‌ی بزرگ و کوچک برای بسیاری از زبان‌های جهان یافت شود. نخستین دلیل این گسترش شتابنده، راه یافتن ویژگی‌های منحصربفرد رایانه در این حوزه و به دنبال آن تغییرات اساسی در روش‌های پردازش متن، ذخیره‌سازی و دستیابی به آن است. "بدین سان، زبان‌شناسی پیکره‌ای اکنون ناگزیر با رایانه پیوند خورده است و همین امر، سرعت شگفت‌انگیز، شمارپذیری کامل، تکرار و روبرداری دقیق، صحت آماری و امکان بکارگیری حجم عظیم داده‌ها را به همراه آورده است." [1]

به کمک پیکره‌ها می‌توان به بررسی‌های آوایی، نحوی، اجتماعی یا سایر زمینه‌های یک زبان پرداخت. بدین ترتیب تکنیک‌های زبان‌شناسی پیکره‌ای را با موضوعات آوایی، نحوی و اجتماعی زبان پیوند زده‌ایم [4].

زبان‌شناسی پیکره‌ای با آمیختن سه روش، به فراهم آوردن دانش تجربی زبانی کمک می‌کند: استخراج خودکار داده‌های زبانی از پیکره‌ها، پردازش برونداد با روش‌های آماری، ارزیابی و تفسیر داده‌های پردازش شده.

موارد اول و دوم را می‌توان به طور کامل به صورت خودکار انجام داد، اما مورد سوم نیاز به تصمیم‌گیری و منطق انسانی دارد [5].

فناوری‌های پردازش هوشمند متون، در سه مرحله: تولید، فرآوری و عرضه‌ی محتوا موجب سرعت‌بخشی به کار شده و شناسایی الگوها، مدل‌ها و ارتباط میان عناصر مختلف در پایگاه داده‌ها را امکان‌پذیر می‌سازد تا دانش نهفته را به دانش کاربردی تبدیل نماید. پردازش هوشمند متون غالبا در دو مرحله‌ی کلی صورت می‌گیرد.

ابتدا در مرحله پایه‌ای "پردازش زبان طبیعی"، ابزارهایی نظیر موارد ذیل، عهده‌دار پردازش اولیه‌ی متن هستند:

- تشخیص واژه‌ها، عبارت‌ها و جملات از یکدیگر؛
- تجزیه صرفی واژگان؛
- ترکیب نحوی واژگان و عبارات؛
- حرکت‌گذاری واژگان؛
- تشخیص روابط منطقی بین کلمات جمله و بندهای متن.

در مرحله بعد، با استفاده از ابزارهای فوق، به کشف لایه‌های معنایی متون پرداخته می‌شود و با توجه به مورد کاربرد، ابزارهایی تولید می‌گردد. از جمله مواردی که می‌تواند جزو نتایج حاصل شده در این مرحله باشد، می‌توان به مواردی همچون ترجمه ماشینی، خلاصه‌سازی، رده‌بندی و دسته‌بندی موضوعی متون و غیره اشاره نمود.

با توجه به اعجازهای بی‌شمار ادبی و آماری در قرآن کریم و لزوم انجام پژوهش‌های گسترده و عمیق توسط رایانه جهت کشف اعجاز کلامی و معنایی قرآن کریم، تهیه‌ی بستری مناسب جهت پردازش رایانه‌ای و هوشمند متن آیات قرآن کریم ضروری به نظر می‌رسد. در ادامه، روش بکار رفته جهت تولید پیکره قرآنی "فرقان" که به عنوان بستری جهت پردازش هوشمند قرآن کریم تولید شده است، تشریح خواهد گردید.

۳- روش پیشنهادی

در روش بکار گرفته شده جهت تولید بستری مناسب جهت انجام روال‌ها و پردازش‌های معناگرایانه بر روی متن قرآن کریم، کلیه‌ی داده‌ها و اطلاعات قرآنی به صورت ساختارمند و در قالب RDF تولید گردید.

۳-۱- ضرورت انجام کار

با تولید پیکره‌ی الکترونیکی قرآن کریم در قالبی ساختارمند می‌توان با بهره‌گیری از فناوری‌های متن‌کاوی[۱۰]، ابزار پیشرفته هوش مصنوعی رایانه را در خدمت اکتشاف اطلاعات پنهان در متن قرآن قرار داد.

در همین راستا طراحی و پیاده‌سازی پیکره‌ی متنی قرآن کریم در قالب RDF و به شکل محتوای الکترونیکی حاوی اطلاعات کلی و آماری قرآن و اطلاعات صرفی و نحوی آیات و همچنین ترجمه‌های فارسی و انگلیسی آیات همراه با برچسب‌گذاری آنها، امکان استفاده و کاوش را برای هر گونه پژوهش و پردازش هوشمند ایجاد خواهد کرد. تبیین اعجاز زبانی قرآن کریم، در دستیابی به لایه‌های نامشهود زبانی این متن الهی می‌باشد. بدین ترتیب پس از فراهم شدن بستر مناسب جهت متن کاوی روی قرآن کریم می‌توان با طراحی و بهره گیری از ابزارهای متن کاوی به اکتشاف کلامی و معنایی متن قرآن کریم پرداخت.

دانش هوشمند متن کاوی به عنوان جدیدترین وجه همکاری بین دانش‌های فناوری اطلاعات، زبان‌شناسی و ادبیات در کاوش رایانه‌ای متون بشری قصد حصول این امر را دارد و چنین کاوشی مستلزم ایجاد پیکره‌های متنی برچسب‌گذاری شده و ساختارمند از آن متون است که به شکل محتوایی الکترونیکی حاوی اطلاعات آماری و اطلاعات صرفی و نحوی متون موردنظرند.

با تولید پیکره‌ای با ویژگی‌های مذکور، با متن‌کاوی بر روی محتوای الکترونیکی قرآن کریم می‌توان به تعیین ارتباط بین بخش‌های مختلف قرآن کریم و ارتباط پنهان سوره‌ها، آیه‌ها، کلمات، حروف، مفاهیم و غیره پرداخت.

۳-۲- داده‌های اولیه

در گام نخست، کلیه‌ی اسناد، اطلاعات، پایگاه داده‌ها و پیکره‌های قرآنی موجود، جمع‌آوری شده و به طور دقیق مورد بررسی قرار گرفتند و با توجه به نیازهای فعلی و اطلاعات لازم برای گام‌های آتی و انجام عملیات متن-کاوی بر روی قرآن، در نهایت، یک قالب RDF ثابت برای سوره، صفحات، حروف الفبایی، آیات قرآن کریم و صرف و نحو آنها در نظر گرفته شد. محتوای این RDFها با طراحی و پیاده‌سازی یک برنامه‌ی کامپیوتری تکمیل خواهد شد. بسیاری از داده‌های مورد نیاز نیز موجود نبودند که با بررسی و مطالعه‌ی اسناد موجود، تولید گردیده و سپس مورد استفاده قرار گرفتند.

۳-۳- قالب RDFهای در نظر گرفته شده

قالب RDFهای در نظر گرفته شده برای پروژه به صورت زیر می‌باشد:

۳-۳-۱- قالب RDF در نظر گرفته شده برای سوره:

- لینک‌های مربوط به توصیفات مفهوم سوره
- لینک‌های مربوط به مفهوم سوره در سایر آنتولوژی‌ها
- لینک فارسی و انگلیسی سوره درWikipedia
- لینک سوره درdbpedia
- لینک خلاصه توضیحات سوره در تفسیر نور
- لینک دانلود پی دی اف توضیحات کامل سوره در تفسیر نور
- پخش صوتی سوره
- شماره‌ی سوره
- جزءها و حزب‌های در بر گیرنده‌ی سوره
- برچسب سوره
- نام سوره به زبان‌های انگلیسی، عربی و فارسی
- مکی یا مدنی بودن سوره
- URI صفحه‌ی آغازین سوره در قرآن عثمان طه
- تعداد آیه‌های سوره
- تعداد سجده‌های سوره
- URI سوره‌ی قبلی و بعدی
- شماره ترتیب نزول سوره
- تعداد بخش‌ها و مباحث مجزای سوره برای قرائت در نماز (Rukus)
- آیا سوره با حروف مقطعه آغاز می‌شود؟
- شروع سوره با چندمین آیه از قرآن کریم است؟
- لینک HTML سوره
- لینک به URI آیات سوره

۳-۳-۲- قالب RDF در نظر گرفته شده برای آیه:

- لینک‌های مربوط به توصیفات مفهوم آیه
- لینک‌های مربوط به مفهوم آیه در سایر آنتولوژی‌ها
- شماره آیه
- چندمین آیه از قرآن کریم
- جزء و حزب در بر گیرنده‌ی آیه
- متن آیه به زبان عربی
- ترجمه‌ی آیه به زبان فارسی و انگلیسی
- ریشه‌یابی و برچسب‌زنی ترجمه‌ی آیه به فارسی و انگلیسی
- لینک به متن عربی آیه به همراه سه ترجمه فارسی و چهار ترجمه انگلیسی
- مفاهیم ذکر شده در آیه
- URI سوره‌ی در بر گیرنده آیه
- وضعیت سجده داشتن آیه
- URI صفحه در بر گیرنده‌ی آیه
- تعداد کلمات به کار رفته در آیه
- تعداد تکرار هر کدام از حروف الفبا در آیه (با اشاره به URI حرف)
- تعداد نقاط به کار رفته در حروف بکار رفته در آیه
- تعداد تکرار هر کدام از حرکات، تنوین‌ها، سکون و تشدید به کار رفته در آیه (با اشاره به URI نشانه)

- لینک به URI صفحه شامل متن صرف و نحو شده‌ی آیه با قالب مناسب
- لینک HTML آیه

۳-۳-۳- قالب RDF در نظر گرفته شده برای متن صرف و نحو شده‌ی آیه با قالب مناسب:

- لینک‌های مربوط به توصیفات مفهوم صرف و نحو آیه
- متن صرف و نحو شده‌ی آیه با قالب مناسب
- لینک HTML صرف و نحو آیه

۳-۳-۴- قالب RDF در نظر گرفته شده برای شماره صفحه:

- لینک‌های مربوط به توصیفات مفهوم صفحه
- شماره صفحه
- URI آیه‌های داخل صفحه به ترتیب
- لینک به URI صفحه‌های قبل و بعد
- لینک HTML صفحه

۳-۳-۵- قالب RDF در نظر گرفته شده برای حروف و اعراب و نشانه‌ها:

- لینک‌های مربوط به توصیفات مفهوم حروف الفبایی
- لینک‌های مربوط به مفهوم حروف الفبا در سایر آنتولوژی‌ها
- آدرس هر حرف یا علامت، مثلا: http://wtlab.um.ac.ir/linkdata/quran/alphabet/arabic/ba
- زبان، نام، صدا و اشکال مختلف حرف
- تعداد نقطه‌های هر نشانه
- لینک حرف در wikipedia و dbpedia
- اندیس حرف و عدد حرف در حروف ابجد
- لینک HTML حرف

۳-۳-۶- نمونه‌ای از RDF سوره

در شکل ۱ بخشی از RDF مربوط به سوره حمد آمده است.


```
- <rdf:RDF>
  - <rdf:Description rdf:about="http://wtlab.um.ac.ir/linkdata/quran/1.rdf">
      <rdf:type rdf:resource="http://dbpedia.org/class/yago/Sura106461830"/>
      <skos:subject rdf:resource="http://dbpedia.org/resource/Category:Sura"/>
      <wtlabprop:subject rdf:resource="http://wtlab.um.ac.ir/linkdata/resource/Category:Sura"/>
      <foaf:page rdf:resource="http://fa.wikipedia.org/wiki/سوره_فاتحه"/>
      <foaf:page rdf:resource="http://en.wikipedia.org/wiki/Al-Fatiha"/>
      <owl:sameAs rdf:resource="http://dbpedia.org/page/Al-Fatiha"/>
      <wtlabprop:tafseer rdf:resource="http://tafseerenoor.persiangig.com/1.htm"/>
      <wtlabprop:tafseernoor rdf:resource="http://gharaati.ir/pdf/tafsirnoor/taf1-s1.rar"/>
      <wtlabprop:tartilFile rdf:resource="http://www.iqna.ir/Sound/Quran/Tartil/Minshawi/001.wma"/>
      <dbpprop:suraNumber>1</dbpprop:suraNumber>
      <dbpprop:juz xml:lang="en">1</dbpprop:juz>
      <dbpprop:hizbNumber xml:lang="en">1</dbpprop:hizbNumber>
      <rdfs:label xml:lang="en">Al-Faatiha Sura</rdfs:label>
      <dbpprop:nameOfSurah>Al-Faatiha</dbpprop:nameOfSurah>
      <dbpprop:arabicName>الفاتحة</dbpprop:arabicName>
```


شکل ۱- بخشی از RDF سوره حمد

۳-۴- استخراج داده‌ها و تولید RDFها

جهت تولید داده‌های قرآنی در قالب RDF کلیه‌ی اسناد و اطلاعات متنی کاغذی و رایانه‌ای، پایگاه داده‌ها و پیکره‌های قرآنی جمع‌آوری شده شامل اطلاعات کلی قرآن کریم شامل تعداد جزءها، سوره‌ها و آیات، مکی و مدنی بودن سوره‌ها، متن قرآن، ترجمه‌های مختلف آن به زبان‌های فارسی و انگلیسی بایستی در روال‌های مشخصی مورد تجزیه و تحلیل قرار گرفته و با طراحی الگوریتم‌هایی مناسب جهت پردازش متن یا کار با پایگاه داده، اطلاعات مناسب را از داده‌های اولیه استخراج نموده و در فیلدهای مورد نظر در قالب RDF قرار داد.

هر کدام از آیات دارای پیوندی به صرف و نحو متن عربی آیه می‌باشد که بستر مناسبی برای کشف روابط بین کلمات با توجه به نقش و صرف و نحو آنها را فراهم آورده است. در مورد ترجمه فارسی و انگلیسی نیز به ازای ترجمه‌ی هر کدام از آیات، متن ریشه‌یابی شده و برچسب‌زده شده توسط ابزار پارسر [۶] و برچسب‌زن نحوی که در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد پیاده‌سازی شده، تولید گردید.

اطلاعات آماری هر کدام از آیات شامل تعداد حروف، تعداد اعراب، تعداد نقاط و سایر موارد مورد نیاز نیز با طراحی الگوریتم‌هایی محاسبه و مقداردهی شدند. بدین ترتیب کلیه فیلدهای در نظر گرفته شده برای هر کدام از قالب‌های RDF سوره، آیه، صفحه و سایر موارد به صورت خودکار مقداردهی شدند.

۳-۵- صرف و نحو متن عربی آیات

متن عربی هر کدام از آیات قرآن کریم در قالب زبان XML، حاوی اطلاعات صرفی و نحوی است که در کاربردهای پردازش متن و متن‌کاوی قرآن، قابل استفاده است. در پیکره متنی قرآن کریم کلیه واژه‌های قرآن کریم به واحدهای بنیادی‌تر شامل بن و اجزای قبل و بعد آن، تقطیع شده و در ابتدا اطلاعات جامع صرفی هر یک و سپس اطلاعات نحوی آنها با بر چسب‌هایی در ذیل هر واژه ارائه شده است. هر برچسب، حاوی دو بخش است؛ ابتدا خصیصه صرفی یا نحوی (مانند باب یا اعراب) و در پی آن محتوای این خصیصه (مانند استفعال یا منصوب). ارائه‌ی این اطلاعات صرفی و نحوی در ساختار XML این امکان را ایجاد کرده است تا هر نرم-افزار پردازشگری بتواند به راحتی از آنها استفاده کند.

۳-۶- خروجی نهایی و انتشار داده‌ها

در نهایت برای هر کدام از سوره‌ها، صفحات، حروف الفبایی، آیات و صرف و نحو آنها یک URI تخصیص داده شده و اطلاعات موجود پس از تکمیل، به صورت RDF و HTML بر روی سایت آزمایشگاه فناوری وب دانشگاه فردوسی مشهد[11] قرار گرفتند. برای تبدیل RDFها به فایل HTML هم تابعی طراحی شد که این عملیات را به صورت خودکار برای تمامی فایل-های RDF انجام دهد. تخصیص URI به عنوان مثال برای صفحه‌ی اول قرآن کریم، سوره‌ی حمد و همچنین اولین آیه از سوره‌ی حمد به شکل زیر صورت پذیرفت:

http://wtlab.um.ac.ir/linkdata/quran/page/1.html
http://wtlab.um.ac.ir/linkdata/quran/1.html
http://wtlab.um.ac.ir/linkdata/quran/1/1.html

با اطلاعات تولید شده در قالب RDF برای سوره، آیه، صفحه، صرف و نحو آیات و حروف الفبایی، دامنه‌ی وسیعی از اطلاعات مفید و سودمند برای انجام عملیات متن کاوی برای پژوهشگران و محققان قرآنی فراهم گردیده است. با بهره‌گیری از دانش هوشمند متن‌کاوی بر روی پیکره‌های متنی برچسب‌گذاری شده قرآن کریم و تهیه آنتولوژی یا هستان‌شناسی جامعی از مفاهیم موجود در قرآن کریم، در گام‌های آتی می‌توان در تبیین اعجاز زبانی قرآن کریم با دستیابی به لایه‌های نامشهود زبانی این متن الهی گام برداشت.

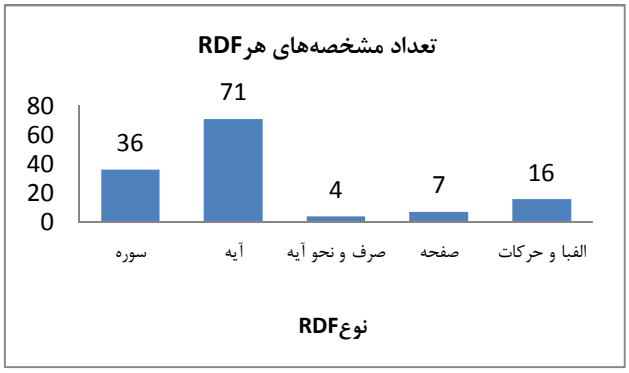
لازم به ذکر است که کلیه مفاهیم و موجودیت‌های موجود در پیکره‌ی جمع‌آوری شده، به آنتولوژی‌ها و مفاهیم مشابه با آنها در وب، لینک داده شده‌اند. بطوریکه پیکره‌ی موجود در حال حاضر شامل بیش از ۳۳۲,۵۸۹ پیوند می‌باشد که تعداد ۳۳,۸۵۴ مورد از آنها منحصر بفرد می‌باشد. در کل پیکره‌ی تهیه شده با حجم داده‌ای نزدیک به ۵۸۷ مگابایت، بیش از ۱۳,۲۹۸ RDF وجود دارد. همچنین ۱۳,۲۹۹ فایل HTML (صفحه وب) برای بازنمایی اطلاعات RDFها وجود دارد.

برخی از اطلاعات آماری مربوط به پیکره تولید شده در جدول ۱ قابل مشاهده است.

جدول ۱- برخی از اطلاعات آماری پیکره "فرقان"

مشخصه	مقدار
حجم کل پیکره	۵۸۷ مگا بایت
تعداد فایل‌های پیکره	۲۶۵۹۸
تعداد RDFهای پیکره	۱۳۲۹۸
تعداد HTMLهای پیکره	۱۳۲۹۸
تعداد RDFهای مربوط به سوره‌ها	۱۱۴
تعداد RDFهای مربوط به آیه‌ها	۶۲۳۶
تعداد RDFهای مربوط به صرف و نحو آیه‌ها	۶۲۳۶
تعداد RDFهای مربوط به صفحات	۶۰۴
تعداد RDFهای مربوط به الفبا و حرکات	۱۰۵

تعداد مشخصه‌ها، سه‌گانه‌ها یا attribute های هر نوع RDF نیز مطابق با نمودار شکل ۲ می‌باشد.

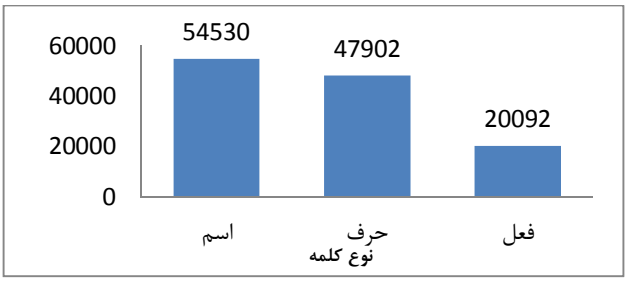


شکل ۲- تعداد مشخصه‌های انواع RDFهای پیکره

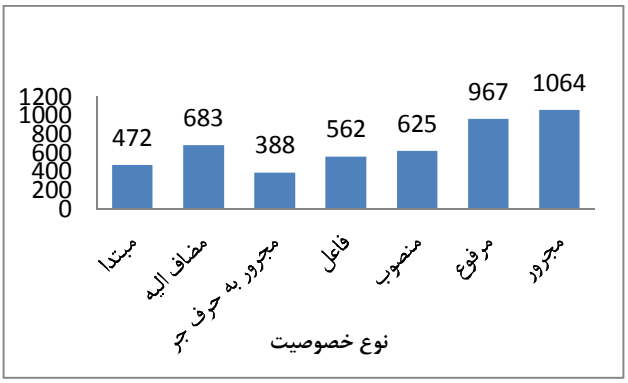
لازم به ذکر است ابزاری جهت تجزیه و تحلیل RDF مربوط به صرف و نحو آیات، طراحی و پیاده‌سازی گردید که در ادامه برخی از نتایجی که از بخش صرف و نحو آیات در پیکره فرقان، قابل حصول است، ارائه خواهد گردید.

- لیست اسم، فعل و حرف در کل قرآن، سوره یا آیه‌ای خاص.
- لیست اسم های مضاف الیه، مجرور به حرف جر، فاعل، مبتدا، مفعول به در کل قرآن، سوره یا آیه‌ای خاص.
- لیست اسم های منصوب، مرفوع، مجرور و تهی.
- لیست کلمات اشتقاق گرفته از یک ریشه‌ی خاص با ذکر سوره‌ها و آیات آنها.
- لیست سوره‌ها و آیاتی که به عنوان مثال کلمه "الله"، دارای ویژگی خاصی است.

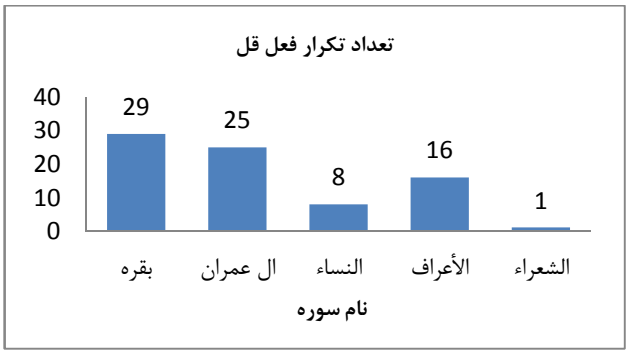
نمودار بیانگر تعداد مواردی که به ازای یک ویژگی خاص توسط ابزار حاصل گردیده است در ادامه در شکل‌های ۳ تا ۶ قابل مشاهده است.



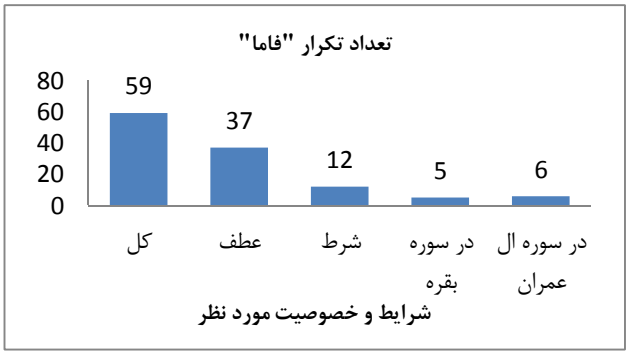
شکل ۳- تعداد تکرار انواع کلمه در بخش صرف و نحو آیات



شکل ۴- تعداد تکرار کلمه جلاله "الله" با بعضی خصوصیات



شکل ۵- تعداد نتایج به ازای فعل "قل"



شکل ۶- نتایج ابزار به ازای شرایط مورد نظر

همچنین ابزاری جهت SPARQL زدن بر روی RDFهای پیکره، تولید گردید که برخی نتایج قابل حصول توسط آن نیز در ادامه ذکر می‌گردد.

- تعداد تکرار یک حرف، اعراب یا نقطه در هر آیه یا هر سوره یا جزء.
- لیست سوره‌های مکی یا مدنی.
- لیست آیات و لیست سوره‌های دربرگیرنده‌ی یک مفهوم خاص مثلا توحید، معاد و غیره.
- لیست آیات یا سوره‌های دربردارنده یک ویژگی یا خصوصیت خاص.
- دسته‌بندی آیات مرتبط با هم از لحاظ ویژگی‌های آماری یا تشابه مفهومی.

۳-۷- گام‌های آتی

در حال حاضر ابزار پارس کردن اطلاعات صرف و نحوی آیات، طراحی گردیده است و کار برای طراحی ابزاری جهت SPARQL زدن بر روی داده‌های RDF قرآنی ادامه دارد. از جمله اقداماتی که می‌توان در گام‌های آتی جهت غنی‌تر شدن خروجی کار و همچنین تولید دانش از پیکره‌ی موجود به آنها پرداخت، می‌توان به موارد ذیل اشاره نمود:

- مشخص کردن موضوع و مفهوم غالب در هر سوره با بررسی و پردازش مفاهیم مندرج در آیات سوره.
- پردازش متن و دسته‌بندی موضوعات سوره‌ها و آیات و مشخص کردن رابطه بین آنها.
- ساخت آنتولوژی موضوعات و مفاهیم قرآن.
- منتسب کردن آیات به مفاهیم.
- به دنبال آن، تعیین ارتباط بین لغات، آیه‌ها، سوره‌ها، جزءها و غیره با پیوند دادن آنها با اطلاعات موجود در وب.
- تکمیل و گسترش آنتولوژی مفاهیم قرآنی با روند یادگیری ماشینی.
- پرسش و استنتاج بر روی پیکره‌ی تولید شده با SPARQL زدن بر روی داده‌های RDF و پارس کردن فایل XML حاوی اطلاعات صرف و نحو متن عربی آیات و تجزیه - تحلیل انسانی و یا هوشمند نتایج جهت کشف روابط کلامی و معنایی پنهان در متن قرآن کریم.

۴- جمع‌بندی

قرآن کریم به عنوان آخرین کتاب آسمانی، هدایتگر بشر در طول تاریخ خواهد بود. در برداشت از روایاتی همچون کلام امام علی (ع) که می‌فرمایند: "قرآن را زمانه تفسیر می‌کند." است که علامه طباطبایی هر دهه را نیازمند تفسیری جدید از قرآن دانسته‌اند و بدیهی است این مهم مستلزم بهره‌مندی از فناوری‌های معاصر است. زبان‌شناسی رایانه‌ای از جمله علومی است که جهت پردازش و تحلیل متن قرآن کریم جهت کشف اعجاز کلامی و معنایی آن می‌تواند به خدمت گرفته شود. پردازش متن قرآن کریم مستلزم در اختیار داشتن پیکره‌ای استاندارد، ساختارمند و قابل فهم برای رایانه است که حاوی تمامی اطلاعات مورد نظر جهت کشف روابط و دانش پنهان در قرآن باشد.

پیکره‌ی قرآنی "فرقان" که در آزمایشگاه فناوری وب دانشگاه فردوسی مشهد، با بهره‌گیری از سامانه‌ای هوشمند، طراحی و پیاده‌سازی شده است، با دربرداشتن حجم عظیمی از داده‌های قرآنی در قالب RDF و HTML، بستر مناسبی را برای پژوهشگران و علاقه‌مندان به پردازش و تجزیه و تحلیل متون قرآنی فراهم آورده است.

سپاسگزاری

در این قسمت، لازم می‌دانیم از زحمات زنده یاد مهندس فرزاد فرخزاده، مهندس دادخواه و اعضای آزمایشگاه فناوری وب دانشگاه فردوسی مشهد و همکاری سرکار خانم زهره استیری سپاسگزاری نماییم. همچنین از گروه پژوهشی پردازش رایانه‌ای قرآن کریم دانشگاه نبی اکرم (ص) در تبریز جهت تهیه‌ی فایل صرف و نحو آیات و همچنین واحد پژوهش بیت القرآن امام علی (ع) شهرستان قم و کلیه‌ی نهادها و سازمان‌هایی که داده‌های خود را در اختیار ما قرار دادند، تشکر می‌کنیم.

مراجع

[1] Kennedy, Graeme , "*An Introduction to Corpus Linguistics*", London, Longman, 1998.

[2] Atkins, B.T.S., Clear, J., and Ostler, N. , "*Corpus Design Criteria*", Journal of Literary and Linguistic Computing, 7, pp. 1-16, 1992.

[۳] عاصی. مصطفی.، **"از پیکره‌ی زبانی تا زبان‌شناسی پیکره‌ای"**، پنجمین کنفرانس زبان‌شناسی، تهران، دانشگاه علامه طباطبایی، ۱۳۸۲.

[4] Leech, Geoffry, "*Corpora and theories of linguistic performance*" in: Svartvik , pp. 105-122, 1992.

[5] Teubert, Wolfgang, "*Corpus linguistics: A partisan view*" in: International Journal of Corpus Linguistics, Vol.4, No.1, 1999.

[۶] استیری. احمد.، کاهانی. محسن.، سعیدی. رضا.، عسگریان. احسان.، **"طراحی ابزار پارسر زبان فارسی"**، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، دانشکده مهندسی برق و کامپیوتر دانشگاه سمنان، شهریور ۱۳۹۱.

زیرنویس‌ها

[1] - Linguistic Corpus

[2] - Linguistic Corpus
[3] - Knowledgebase
[4] - Monolingual
[5] - Multilingual
[6] - Annotated
[7] - Part-of-Speech - POS
[8] - Lemma
[9] - Corpus Linguistics
[10] - Text Mining
[11] - http://wtlab.um.ac.ir