دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال پانزدهم، شماره ۵۷ و۵۸ ، پاییز و زمستان ۱۴۰۲
صفحات:۱۵۳ الی ۱۶۴

Noor-Ghateh: A Benchmark Dataset for Evaluating Arabic Word Segmenters in Hadith Domain

Hoda Ashoheib[*], Behrooz Minaei Bidgoli[**], Mohammad Ebrahim Shenasa[***], Seyyed Ali Hosseini[****]

[*]Ph.D. student, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[**]Professor, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[***]Faculty member, Electrical and Computer Engineering, Islamic Azad University, North Tehran Branch, Tehran, Iran
[****] Artificial intelligence laboratory, Digital humanities and Islamic sciences research institute (noor), Qom, Iran

Abstract

The Arabic language has a very rich and complex morphology, which is very useful for the analysis of the Arabic language, especially in traditional Arabic texts such as historical and religious texts, and helps in understanding the meaning of the texts.

In the morphological data set, the variety of labels and the number of data samples helps to evaluate the morphological methods, in this research, the morphological dataset that we present includes about 22, 3690 words from the book of Sharia al-Islam, which have been labeled by experts, and this dataset is the largest in terms of volume and The variety of labels is superior to other data provided for Arabic morphological analysis. To evaluate the data, we applied the Farasa system to the texts and we report the annotation quality through four evaluation on the Farasa system.

Keywords: Morphology, Arabic Language, Annotation, Dataset, Morphological Analysis.

نور-قطعه: یک دادگان معیار برای ارزیابی روش‌های جداساز واژگان عربی در دامنه‌ی متون فقهی

هدی الشهیب*، بهروز مینایی بیدگلی[1]**، محمدابراهیم شناسا***، سیدعلی حسینی****

*دانشجوی دکتری دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

**استاد گروه مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

***عضو هیئت علمی دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران‌شمال، تهران، ایران

****آزمایشگاه هوش مصنوعی پژوهشکده علوم اسلامی و انسانی دیجیتال (نور)، قم، ایران

تاریخ دریافت: ۱۴۰۱/۰۸/۰۴ تاریخ پذیرش:۱۴۰۱/۱۲/۱۰

نوع مقاله: پژوهشی

چکیده

زبان عربی ریخت‌شناسی بسیار غنی و پیچیده‌ای دارد که برای تحلیل زبان عربی و به ویژه در متون عربی سنتی مانند متون تاریخی و مذهبی بسیار مفید است و در فهم معنای متون کمک می‌کند. جداسازی واژگان به معنای تفکیک واژه به بخش‌های مختلف مانند هسته و وندها می‌باشد.

در مجموعه داده‌های ریخت‌شناسی تنوع برچسب و تعداد نمونه‌های دادگان به ارزیابی روش‌های ریخت‌شناسی کمک بیشتری می‌کند، در این پژوهش مجموعه‌ی داده محکی برای ارزیابی روش‌های جداساز واژگان عربی ارائه می‌کنیم که شامل حدود ۲۲۳۶۹۰ کلمه از کتاب شرائع الاسلام در ۵۲ باب فقهی است و توسط متخصصین برچسب‌گذاری شده است. این مجموعه دادگان با داشتن از نظر حجم و تنوع کلمات نسبت به سایر دادگان‌های موجود برتر می‌باشد و تا جایی که می‌دانیم هیچ دادگانی از متون فقهی عربی در این زمینه وجود ندارد. برای ارزیابی دادگان، سامانه فراسه را بر روی متون اعمال کردیم و کیفیت جداسازی واژه‌ها را از طریق چهار معیار بر روی سامانه فراسه گزارش کردیم.

واژگان کلیدی: جداسازی واژگان، زبان عربی، حاشیه‌نویسی، دادگان، برچسب‌گذاری صرفی

۱. مقدمه

در پژوهش‌های اخیر زبان عربی بسیار مورد توجه پژوهش‌گران قرار گرفته است، زیرا زبان اصلی در بسیاری از کشورها است و با پیاده‌سازی سامانه‌های مختلف عربی، خواسته‌های کاربران در این کشورها برآورده می‌شود. زبان عربی ساختارهای معنایی و آوایی پیچیده‌ای دارد و در مقایسه با دیگر زبان‌ها از ریخت‌شناسی[3] غنی برخوردار است و این موضوع فرآیند تجزیه و تحلیل را دشوار می‌کند. تجزیه و تحلیل ریخت‌شناسی در بسیاری از وظایف پردازش زبان طبیعی مانند بازیابی اطلاعات، چک کردن املا و زبان‌شناسی

[1] نویسنده مسئول: بهروز مینائی‌بیدگلی b_minaei@iust.ac.ir

[3] Morphology

کاربرد دارد و مرحله اولیه هر تحلیل نحوی است. ساختار کلمات عربی بر ریشه تکیه دارد، ریشه اصلی‌ترین قسمت در ساختار کلمه است که با ترکیب پیش‌وندها یا پس‌وندها به مشتقات ریشه به دست می‌آیند و افعال، صفت‌ها و اسم‌ها را ارائه دهند. اشتقاق یک کلمه از ریشه معمولاً در دو مرحله صورت می‌گیرد. در ابتدا با توجه به الگوی کلمه، ریشه آن بدست می‌آید، سپس ضمائم به ریشه اضافه می‌شوند. معمولا برای کاهش ساختار کلمه به ریشه، از الگوریتم‌های ریشه‌یابی استفاده می‌شود. در زبان انگلیسی، معمولاً یا فقط پیش‌وند یا فقط پس‌وند وجود دارد. اما در زبان عربی کلمه عموما از پیش‌وند، هسته و پس‌وند تشکیل می‌شود. بنابراین استخراج ریشه در این حالت دارای چالش بیشتری نسبت به سایر زبانها می‌باشد.

یکی از مشکلات پردازش زبان طبیعی عربی، تنوع شکل کلمات به دلیل غنای ریخت‌شناسی و فقدان قوانین املایی است. مشکل دیگر ضعف در حل ابهاماتی مانند ناهنجاری کلمات است زیرا بیشتر ریشه‌ها از سه حرف تشکیل شده اند و همچنین چالش‌های تغییرات املایی و ریخت‌شناسی بسیار غنی پیچیده است.

ریخت‌شناسی ستون اصلی وظایف پردازش زبان طبیعی است. بنابراین، تجزیه و تحلیل ریخت‌شناسی، نقطه شروع بیشتر سامانه های پردازش زبان برای تعامل بین افراد است. تحلیل‌گرهای ریخت‌شناسی در مراحل اولیه کار خود از منابعی مانند فرهنگ لغت که تمام تحلیل‌های ممکن را برای یک کلمه مشخص ارائه می‌دهند، استفاده می‌کنند.

از تحلیل‌گرهای ریخت‌شناسی انتظار می‌رود که همه جداسازی‌های مطلوب یک کلمه معین را همراه با در نظر گرفتن همه حالت‌های مختلف انطباق یک بن[1]کلمه برگردانند. تحلیل‌گرهای قدرتمندتر با استفاده از فرهنگ لغت‌های زبانی موجود به دقت ساخته شده و به صورت دستی بررسی می‌شوند. استاندارد دامنه‌های تحلیل‌گرهای ریخت‌شناسی از روش‌هایی که برای ایجاد آن‌ها مرسوم است، پشتیبانی می‌کند.

فرهنگ‌لغت‌های عربی دردسترس زیادی وجود ندارد. لذا، محققان سعی کردند با جمع‌آوری مجموعه داده نور-قطعه، به پژوهش‌گران کمک کنند تا سامانه‌های خود را بر روی این مجموعه داده آزمایش کرده و میزان موفقیت سامانه‌های خود را بسنجند. در این راستا، تصمیم گرفته شده است تا مجموعه داده‌های کتاب شرایع را تحلیل و بررسی کنیم. کار با متون عربی کتاب‌های حدیث، برای محققانی که به طور کلی در زمینه‌های پردازش زبان عربی کار می‌کنند، مهم است. تا آنجا که می‌دانیم، هیچ مجموعه فقهی عربی در حال حاضر در دسترس نیست. مجموعه ریخت‌شناسی نور-قطعه را می‌توان در بسیاری از ابزارهای پردازش زبان عربی استفاده کرد.

۲. کارهای انجام شده

در طول چند دهه گذشته، چندین روش برای تجزیه و تحلیل ریخت‌شناسی عربی پیشنهاد شده است. هنوز چالش‌های زیادی وجود دارد که موتورهای تحلیل ریخت‌شناسی با آن مواجه هستند. این بخش تحقیقات اخیر در مورد تجزیه و تحلیل ریخت‌شناسی عربی را مورد بحث قرار می‌دهد.

یکی از موتورهای اولیه تجزیه و تحلیل ریخت‌شناسی باکوالتر [۱] است که برای برچسب گذاری اجزای کلام متون عربی طراحی شده است. داده‌ها اساساً از سه فایل واژگان عربی ـــ انگلیسی تشکیل شده‌اند و هدف آن حمایت از محققانی است که به واژگان عربی و تجزیه و تحلیل ریخت‌شناسی نیاز دارند. آن‌ها شش جدول طراحی کردند: سه تای آن‌ها جدول واژگانی و سه تای آن‌ها جدول سازگاری هستند. سامانه از ریخت‌شناسی الگوی ریشه پشتیبانی می‌کند و بنابراین اولین نسخه تولید شده توسط LDC[2] است. که نسخه ۲.۰ آن [۲] با نام باما در سال ۲۰۰۴ منتشر شد. آخرین نسخه این سامانه [۳] که سما نام دارد تعدادی از ناسازگاری‌های نسخه باما را اصلاح کرد و موارد جدیدی را هنگام بازسازی ورودی‌های موجود برای مطابقت با استانداردهای جدید اضافه کرد، به علاوه ناهماهنگی‌های اضافی را هنگام وارد کردن ورودی‌های جدید اضافه کرد. مدل لاما [۴] اساس این پژوهش‌ها است که در آن LDC عربی پن[3] را منتشر کردند و در بسیاری از مدل‌سازی‌های ریخت‌شناسی عربی از آن استفاده کردند. پژوهش [۵] ریشه‌های کلمات عربی را با یادگیری باناظر تشخیص می‌دهد. این الگوریتم بر روی مجموعه‌ای از متون عربی استخراج شده از وب‌سایت الجزیره آزمایش شد و آن‌ها کار خود را روی ۲۷۰۰ کلمه منحصر به فرد آزمایش کردند و به دقت ۹۲ درصد رسیدند.

یکی دیگر از موتورهای تحلیل ریخت‌شناسی ماجید [۶] است و می‌تواند کلمات گویشی و کلمات نوشتاری را به ریشه/الگو مرتبط تجزیه و تحلیل کند. این تجزیه و تحلیل اینترنتی که برای به دست آوردن ریشه+ الگو+ ویژگی است، شکل‌های واجی و املایی را جدا می‌کند و امکان ادغام تک‌واژهای گویش‌های مختلف را فراهم می‌کند که شامل ریشه‌ها، با درجات مختلف دقت است. در پژوهش [۷] تنها از آمار وابستگی، ریخت‌شناسی ریشه سامی و بدون فرهنگ لغت برای ریخت‌شناسی عربی استفاده شده است. آن‌ها الگوریتمی را برای ریخت‌شناسی الحاقی عربی معرفی کرده‌اند و در

[1] Lemma
[2] Linguistic Data Consortium
[3] Penn Arabic Treebank

الگوریتم شناسایی ریشه، الگوی مصوت را به عنوان کاراکترهای ریشه ممکن حذف می‌کنند.

موتور تجزیه وتحلیل ریخت‌شناسی المورگینا [۸] در ابتدای کار اجزای مختلف مورد نیاز سامانه را پیش‌پردازش می‌کند. پایگاه داده آن مقادیری را که هر ویژگی می‌تواند داشته باشد و همچنین مقادیر ویژگی پیش‌فرض را برای هر کلمه پوشش می‌دهد و اشکال واجی را پوشش نمی‌دهد و به طور خودکار پر شده است و به طور کامل بررسی نشده است. الاکسیرفم [۹] از واژگان باکوالتر استفاده کرده است و جنسیت، عدد، حالت‌ها و مدل‌سازی کامل و فرم‌های واجی را پوشش داده است. این تحلیلگر ریخت‌شناسی به زبان هسکل نوشته شده است که شامل یک کتابخانه برنامه نویسی چندمنظوره و لغت‌نامه ریخت شناسی زبانی است. مزیتش این است که چهار حالت مختلف عملکرد را برای تجزیه و تحلیل یک کلمه یا متن عربی در اختیار کاربر قرار می‌دهد، اما پوشش محدودی دارد زیرا فقط کلمات را در عربی نوشتاری مدرن تحلیل می‌کند.

در پژوهش [۱۰]، محققان چندین تکنیک یادگیری ماشین را برای پژوهش [۱۱] اعمال کردند و قابلیت طبقه‌بندی چند برچسبی[1] را دارد که برای یادگیری در حوزه‌های تخصصی می‌تواند قابل استفاده باشد. مقاله [۱۲] از الگوهای متقابل زبانی و بدون نظارت برای تجزیه و تحلیل تقسیم‌بندی[2] ریخت‌شناسی استفاده می‌کند. آن‌ها همچنین شواهدی مبنی بر اینکه توجه به زبان‌های نزدیک به هم سودمندتر است ارائه کردند، البته اگر این مدل بتواند به صراحت ساختار زبان مشترک را نشان دهد. پژوهش [۱۳] یک مدل لاگ‌خطی[3] برای تقسیم‌بندی ریخت‌شناسی بدون‌نظارت ارائه کرده و الگوریتم‌هایی را برای یادگیری و نتیجه‌گیری ارائه کرده است. موتور آن‌ها فقط مشخصات تک زبانه را پشتیبانی می‌کند.

پژوهش [۱۴] یک سامانه تجزیه و تحلیل ریخت‌شناسی را برای جملات عربی بی‌صدا ارائه کرده است. در ابتدا، آن‌ها کارآمدی مولفه اولیه را برای جستجوی پایه‌ای که توسط حاشیه‌نویس‌های احتمالی در بین ریشه‌های دیگر تعیین شده بود آزمایش کردند و متوجه شدند که مقدار ریشه‌های تولید شده توسط سامانه از ۱ تا ۱۲ ریشه است. سپس با معرفی رویکردی که مدل‌های مارکوف پنهان را پشتیبانی می‌کند، ریشه مناسب هر کلمه را انتخاب کردند. پژوهش [۱۵] از یادگیری ماشینی برای پیش‌بینی ویژگی‌های ریخت‌شناسی استفاده کرده است. آن‌ها از فاصله لونشتاین[4] و تطبیق الگوی کلمه عربی برای استخراج جمع مکسر استفاده کردند.

رویکرد [۱۶] مبتنی بر یادگیری واژگان بیزی[5] است که مدلی از یادگیری تکواژواژه‌نامه ارائه کرده‌اند که قادر به مدیریت ریخت‌شناسی پیوسته و غیرهمبسته تا سطح دو تکواژ می‌باشد. پژوهش [۱۷] یک رویکرد بدون‌نظارت برای یادگیری ریخت‌شناسی ارائه کرده است، آن‌ها ریشه‌های ثلاثی[6] و قالب‌های الگو[7] را یاد می‌گیرند و مبتنی بر یادگیری ماشینی عمل می‌کنند. کلیم‌گلف[8] [۱۸] یک تحلیل‌گر ریخت‌شناسی برای گویش کشورهای حاشیه خلیج فارس است که یک تحلیل‌گر از ورودی‌های فرهنگ لغت آوایی ایجاد کرده است و سپس پارادایم‌های املایی و واژگان و انواع املایی مرتبط را تولید می‌کند.

سامانه دیگری برای گویش خلیجی [۱۹] وجود دارد که با استفاده از ترکیب‌های مختلف تحلیل‌گرهای ریخت‌شناسی، مدل‌های ابهام‌زدایی و اندازه داده‌های آموزشی آزمایش شده است. آن‌ها از مجموعه داده‌های حاشیه‌نویسی شده گومار[9] [۲۰] استفاده کرده‌اند که بخشی از زیرمجموعه عربی اماراتی گومار [۲۱] است، و از سه تحلیل‌گر ریخت‌شناسی سما [۲۱]، کالیما[10] [۲۳] و گلف ـــ مپ[11] برای عربی خلیجی استفاده کرده‌اند که به طور خودکار انجام می‌شد. پژوهش [۲۴] تجزیه و تحلیل ریخت‌شناسی دیگری را برای عربی استاندارد نوشتاری توسعه داد. سامانه آن‌ها به فناوری خودکار ریخت‌شناسی عربی متکی است که ریشه را پیدا می‌کند و سپس بلافاصله برای تولید و تجزیه و تحلیل ریخت‌شناسی عربی مورد استفاده قرار می‌گیرد. پژوهش [۲۵] بر اساس شبکه‌های عصبی بازگشتی[12] است.

توسعه‌دهندگان واحدهای حافظه طولانی کوتاه‌مدت[13] را در اشکال مختلف و مجموعه‌های تعبیه‌شده برای مدل‌سازی ویژگی‌های ریخت‌شناسی مختلف بکار گرفتند. الخلیل [۲۶,۲۷] موتور تحلیل ریخت‌شناسی دیگری است که برای کلمات استاندارد عربی دو نسخه دارد. در الخلیل ریخت‌شناسی مبتنی بر ساقه[14] است، شامل ویژگی‌های ریشه‌ای و نحوی است و قادر است متون غیرآوایی را پردازش کند و از مدل سازی مجموعه بسیار وسیعی از قوانین ریخت‌شناسی عربی پشتیبانی می‌کند. این سامانه از پیکره تشکیله[15] و نملار[16] استفاده می‌کند و با متن غیرآوایی، جزئی یا کاملاً صدادار کار می‌کند. پژوهش [۲۸] تحلیل‌گر نوشتاری الخلیل را به تونسی تغییر می‌دهد، در این پژوهش الگوهای مشتق را اصلاح

[1] multi-label classifier
[2] Segmentation
[3] log-linear
[4] Levenshtein
[5] Bayesian
[6] tripartite roots
[7] pattern templates
[8] CALIMAGLF
[9] Annotated Gumar Corpus
[10] CALIMA
[11] GLF-MAPC
[12] Recurrent Neural Network
[13] Long Short Term Memory
[14] Stem
[15] Tashkeela
[16] Nemlar

می‌کنند و ریشه‌ها و الگوهای خاص گویش تونسی را اضافه می‌کنند. مقاله [۲۹] بخشی از عربی پن[1] را افزایش می‌دهد تا جنسیت و عدد عملکردی و منطق واژگانی را در بر بگیرد، اما کل پایگاه داده مورد استفاده باما یا سما را پوشش نداد. مادامیرا [۳۰] موتور دیگری است که مادا [۳۱-۳۳] و امیره [۳۴] را برای تجزیه و تحلیل ریخت‌شناسی و ابهام‌زدایی از کلمات عربی ترکیب می‌کند. در ابتدا، سامانه با استفاده از تحلیل‌گر سما، کلمات جمله را خارج از متن تجزیه و تحلیل می‌کند، سپس از روش SVM[2] برای رفع ابهام بین کلمات به دست آمده از مرحله قبل استفاده می‌کند.

بسما[3] [۳۵] از باما برای فرآیند ابهام‌زدایی ریخت‌شناسی استفاده کرده است. آن‌ها تعداد، جنسیت و معرف‌ها[4] را مطابق با ویژگی‌های ریخت‌شناسی خود تغییر داده‌اند. در این پژوهش برخی از برچسب‌ها، لغات و واژه نامه‌ها را اصلاح کرده‌اند و به واژگان ICA[5] اضافه کرده‌اند. همچنین تجزیه و تحلیل و واجد شرایطین را به عنوان ریشه، الگوی ریشه و موجودیت‌های نامدار[6] اضافه کرده‌اند. پس از انتخاب موثرترین راه‌حل، اجزای کلمات برای هر کلمه، لم‌ها، ریشه‌ها، الگوهای ریشه، تعداد، جنسیت، قطعیت، حروف و در نهایت صدای هر کلمه را تشخیص می‌دهد. همچنین پیش‌وندها و پس‌وندها را از ریشه جدا کرده‌اند و کلمه ورودی هر کلمه را نمایش داده‌اند و پاسخ باما را فقط با کلماتی که اصلاً راه‌حلی نداشتند نشان داده‌اند. یماما[7] [۳۶] مانند مادامیرا در تحلیل و ابهام‌زدایی کار کرده است. مؤلفه اصلی آن مدل کلمه و محتمل‌ترین تحلیل ریخت‌شناسی کلمه است و برای ابهام‌زدایی از مدل حداکثر درست‌نمایی استفاده می‌کند. آن‌ها از عربی پن بخش‌های ۱ تا ۳ برای گویش مصری از ARZ Treebank[8] استفاده کرده‌اند. کالیما [۲۳]، ایکال[9] [۳۷] را گسترش داد و یک تحلیل‌گر ریخت‌شناسی برای گویش مصری به نام کالیما طراحی کرد که از نظر زبانی دقیق و در مقیاس بزرگ ارائه شده است. سامانه آن‌ها چندین گونه املایی را می‌پذیرد و آن‌ها را به یک املای سنتی تبدیل می‌کند. کالیما دارای ۱۰۰ هزار ریشه است که با ۳۶ هزار بن مطابقت دارد. در این روش ۲۴۲۱ پیش‌وند پیچیده و ۱۱۷۹ پس‌وند پیچیده وجود دارد. تعداد کامل کلمات قابل تحلیل توسط کالیما ۴۸ میلیون کلمه است. سامانه کالیماستار [۲۰] ویژگی‌های ریخت‌شناسی عملکردی و مبتنی بر فرم را به‌عنوان نشانه‌ساز داخلی، نمایش واج‌شناختی و واژگانی پیاده‌سازی می‌کند. پژوهش [۳۸] مجموعه‌ای از پیکره‌های حاشیه‌نویسی صرفی را که مجموعا بیش از ۲۰۰۰۰۰ کلمه را برای هفت گویش عربی ارائه کرده است که این کار به صورت دستی در مجموعه استانداردهای مشترکی برای املا، بن‌های تشریحی، نشانه‌گذاری، واحدهای ریخت‌شناسی و لغات انگلیسی حاشیه نویسی شدند.

۳. دادگان‌های موجود

جهت آشنایی با ویژگی‌های دادگان‌های موجود و مقایسه آن‌ها با دادگان این مقاله، به معرفی و بررسی هر یک از آن‌ها می‌پردازیم.

۱.۳ بانک وابستگی عربی پراگ[39]

شامل حاشیه‌نویسی‌های چند سطحی نوشتاری از جمله سطح ریخت شناسی و تحلیلی بازنمایی زبانی است که برای استفاده عمومی در پردازش زبان طبیعی طراحی شده است. در این مجموعه با اضافه کردن یک عدد به عنوان نشانگر معنای بن، بن‌هایی را با نمایش متنی یکسان رفع می‌کند [۳۹].

۲.۳ کلارا

یک مجموعه آموزشی با مرزهای ریخت‌شناسی مشخصی است که شامل ۱۰۰۰۰۰ کلمه، یک پایگاه داده از رشته‌ها با مرزهای ریخت‌شناسی مشخص شده، و یک مجموعه آموزشی دیگر با حاشیه‌نویسی قسمت‌هایی از گفتار است. اندازه تحلیل شده این مجموعه حدود ۱۵۰۰۰ کلمه است [۴۰].

۳.۳ بانک عربی پن

حاوی بیش از نیم میلیون (۵۴۲۵۴۳) کلمه عربی است که از روزنامه‌های آژانس فرانسه پرس و الحیات و النهار جمع‌آوری شده است که شامل برچسب گذاری اجزای کلمات و تجزیه[10] است [۴].

۴.۳ مجموعه قرآنی

حاوی محتوای متنی عربی قرآنی است که در آن همه عبارات با اطلاعات صرف نحوی حاشیه‌نویسی شده اند، این مجموعه شامل ریشه، الگوی ریشه، بن، الگوی بن و ریشه است و از رویکردی نیمه خودکار استفاده می‌کند که از سامانه ریخت‌شناسی الخلیل استفاده کرده است [۴۱].

۵.۳ مجموعه قرآنی عربی

یک مجموعه مشروح اینترنتی چند لایه حاشیه‌نویسی شده است که شامل ارزیابی نحوی، تقسیم‌بندی ریخت‌شناسی، برچسب گذاری اجزای کلمات و هستی‌شناسی معنایی[11] است که با استفاده از قواعد وابستگی حاصل شده‌اند [۴۲].

۶.۳ المصحف

1 Penn Arabic TreeBank
2 Support Vector Machine
3 Basma
4 definiteness
5 International Corpus of Arabic
6 Name Entities
7 YAMAMA
8 LDC2012E{93,98,89,99,107,125}, LDC2013E{12,21}
9 ECAL
10 parsing
11 Semantic ontology

مجموعه قرآنی است که با استفاده از سامانه ریخت‌شناسی الخلیل و رویکرد نیمه نظارت شده، برچسب گذاری صرفی شده است. این مجموعه یک رویکرد سلسله مراتبی خاص از کلاس‌های برچسب عربی، و همچنین عبارت‌های ترجمه شده به انگلیسی، اسپانیایی و فرانسوی است [۴۳].

۷.۳ مجموعه قرآنی حیفا

شامل تحلیل‌های چندگانه برای کلمات است. گردآورندگان این مجموعه واحدهای برچسب اجزای کلمات که برای حاشیه‌نویسی بدنه استفاده شده است را به خوبی ارائه داده‌اند، این مجموعه کامل نیست و به صورت دستی تأیید نشده است و دارای چندین تجزیه و تحلیل ممکن برای هر کلمه در مجموعه داده‌های منتشر شده نهایی است [۴۴].

۸.۳ مجموعه عربی قرآن

یک ارزیابی کاملا منحصر به فرد برای هر عبارت در آیه متنی خود ارائه می‌دهد و ارزیابی خودکار به صورت دستی در این مجموعه بررسی شده است و از مجموعه برچسب اجزای کلمات استفاده می‌کند، حتی اگر چند برچسب اسمی وجود داشته باشد. این مجموعه از محتوای متنی عربی همراه با رونویسی آوایی، ترجمه عبارت به عبارت و تابع مرجع استفاده می‌کند [۴۵].

۹.۳ مجموعه قرآنی مصحف

تاکنون به صورت آنلاین در دسترس نیست و فقط در [۴۶] با چند نمونه ذکر شده است، این مجموعه غنی شده با اطلاعات ریخت‌شناسی است. فرآیند ساخت این پیکره شامل یک تکنیک نیمه خودکار با استفاده از الخلیل و سپس فرآیندهای دستی است. این پیکره دارای ۱۷۷۰ ریشه، الگوهای علامت‌گذاری شده است.

۱۰.۳ مجموعه داده‌های قرآن ریزسطحی[1]

مجموعه‌ای از متون قرآنی است و به دانش ساختاریافته چند سطحی شامل سطح فصل، سطح کلمه و سپس به سطح شخصیت منتقل می‌شود و همه اینها با توضیحات و تعاریف، تجزیه، ترجمه‌ها، ریشه‌های لحن و ریشه کلمات ذکر شده است. مجموعه داده نهایی در قالب اکسل و به صورت اسناد پایگاه داده در اختیار قرار گرفته شده است [۴۷].

۴. معرفی دادگان نور-قطعه

تجزیه و تحلیل ریخت‌شناسی پایه بسیاری از عملیات پردازش زبان طبیعی است و افزایش دقت آن سبب افزایش عملیات می‌شود. هر چه برچسب‌های ریخت‌شناسی یک کلمه متنوع‌تر باشد، دقیق‌تر بوده و می‌تواند ملاک ارزیابی ابزارهای تحلیل صرفی بیشتری قرار بگیرد. در زبان عربی، یک کلمه معمولا از یک هسته اصلی و مجموعه‌ای از تک‌واژها تشکیل می‌شود. در این مقاله دادگان نور-قطعه را برای ارزیابی روش‌های جداسازی واژگان ارائه می‌کنیم. توصیف آماری این دادگان در جدول ۱ مشاهده می‌گردد.

جدول ۱. توصیف آماری دادگان نور-قطعه

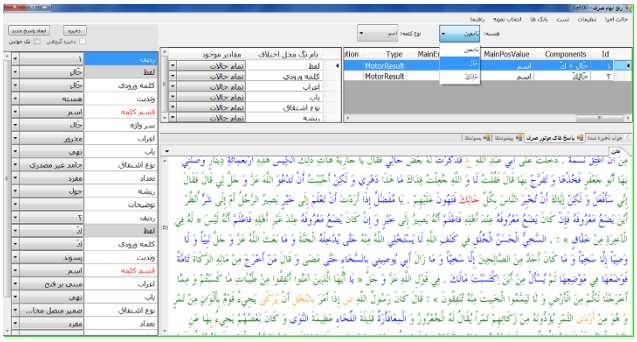
ویژگی	فراوانی
تعداد ابواب فقهی	۵۲
تعداد کلمات	۲۲۳۶۹۰
تعداد جملات	۱۰۱۶۰
تعداد اسم	۱۲۰۴۳۲
تعداد فعل	۴۰۰۲۹
تعداد حرف	۱۲۴۷۲۴
تعداد هسته	۲۶۴۰۹۷
تعداد پیش‌وند	۷۴۲۴۲
تعداد پس‌وند	۱۸۶۱۷

برای تهیه این دادگان، نسخه دیجیتالی با کیفیت و فرآوری شده حاوی محدوده محتوای مناسب از کتاب شرایع‌الاسلام انتخاب شد. سپس عملیات نرمال‌سازی و پالایش متن با استفاده از ابزار فراسه [۴۸] بر روی کلمات موجود انجام گردید. در ادامه یک تیم ۵ نفره برچسب‌های جداسازی واژگان و مقادیر معتبر برای هر یک از آنها را استخراج کردند و البته هر واژه لزوما توسط هر ۵ نفر بررسی نمی‌گردید و در مواردی حداکثر آرا لحاظ می‌شد. برای جداسازی واژگان ابزاری بومی با استفاده از ابزار مادامیرا [۳۰] ساخته شد که امکان بازبینی، رفع ابهام و اصلاح پاسخ های ماشینی توسط انسان خبره در حداقل زمان ممکن را فراهم می‌آورد. نمایی از این ابزار در شکل ۱ نمایش داده شده است. در نهایت خروجی آن‌ها به صورت دسته‌بندی شده توسط یک ناظر، بررسی و تایید شد و به قالب XML تبدیل گردید. در ادامه به بررسی و معرفی ویژگی‌های این دادگان می‌پردازیم.

۱.۴ برچسب‌های ریخت‌شناسی

در این دادگان، مطابق تمامی تحلیل‌های جداسازی واژگان، کلمه عربی به سه بخش تک‌واژ پیش‌وند، هسته و پس‌وند تقسیم و تحلیل شده است. در این مجموعه، برچسب‌های اجزای کلمات اسمی برای تجزیه و تحلیل نحوی و ریخت‌شناسی معرفی شده‌اند و هر برچسب معنا و مفهوم خود را دارد. در این دادگان ۵ گروه مختلف از برچسب‌ها موجود است که در جدول ۲ مشاهده می‌گردد.

[1] Fine-Grained Quran Dataset



شکل ۱. نمایی از ابزار بومی اصلاح و بازبینی واژه‌ها

جدول ۲. شرح هر یک از برچسب‌ها

برچسب	معنی/ مقادیر
Seq	ترتیب تک‌واژ/ اعداد طبیعی
Slice	محتوای تک‌واژ / کلمه
Entry	تکواژی که ممکن است شامل تکواژهای دیگر باشد
Affix	نوع تک‌واژ / پیشوند، پسوند، هسته
Pos	نوع کلمه / اسم، فعل، حرف

۲.۴ قالب مجموعه داده

فرمت مجموعه داده نور-قطعه، XML اســت. شــکل ۲ نمونه‌ای از مجموعه کدگذاری شده در قالب XML را نشان می‌دهد.

پژوهش‌گران برای ســهولت اســتفاده از این پیکره، یک فایل XML تولید کردند و از نمادهایی برای تســهیل خولندن پیکره اســتفاده کردند: کلمات با برچسب <Root> شروع و با برچسب </Root> پایان می‌یابند. ابتدا کلمه را به صــورت یک دنباله می‌نویســد، به عنوان مثال، "لطلوع" به دو قسمت "ل" + "طلوع"، بنابراین شماره ۱ به دنباله "ل" و شــماره ۲ به دنباله "طلوع" تعلق می‌گیرد. و سپس برچسب‌های مختلف کلمه با توجه به نوع آن ذکر می‌شود.

```
<Base>
 <Root>
  <word Seq="1" Slice="إِذَا" Entry="إِذَا" Affix="اسم"pos= "هسته" />
 </Root>
 <Root>
  <word Seq="1" Slice="بَقِي" Entry="بَقِي" Affix="فعل"pos="هسته"/>
 </Root>
 <Root>
  <word Seq="1" Slice="لِ" Entry="لِ" Affix="پیشوند" pos="حرف" />
  <word Seq="2" Slice="طُلُوعِ" Entry="طُلُوعِ" Affix="اسم"pos= "هسته" />
 </Root>
 <Root>
  <word Seq="1" Slice="الْ" Entry="الْ" Affix="پیشوند" pos="حرف"/>
  <word Seq="2" Slice="فَجْرِ" Entry="فَجْرِ" Affix="اسم"pos= "هسته" />
 </Root>
 <Root>
  <word Seq="1" Slice="مِنْ" Entry="مِنْ" Affix="هسته"pos="حرف" />
 </Root>
 <Root>
  <word Seq="1" Slice="يَوْمِ" Entry="يَوْمِ" Affix="هسته"pos="اسم" />
 </Root>
 <Root>
  <word Seq="1" Slice="يَجِبُ" Entry="يَجِبُ" Affix="هسته"pos="فعل" />
 </Root>
 <Root>
  <word Seq="1" Slice="صَوْمُ" Entry="صَوْمُ" Affix="هسته"pos="اسم" />
  <word Seq="2" Slice="هُ" Entry="هُ" Affix="پس‌وند"pos="اسم" />
 </Root>
</Base>
```

شکل ۲. نمونه‌ای از قالب فایل XML

۵. مقایسه دادگان‌های ریخت‌شناسی عربی

جدول شماره ۳ مقایسه‌ی بین دادگان‌های ریخت‌شناسی عربی را نشان می‌دهد. در این مقایسه تنوع زبانی و موضوعی هر مجموعه دادگان و تعداد کلمات هر مجموعه و کاربردهای آن‌ها را ذکر کرده‌ایم که بتوان دیدگاه شاملی از آن‌ها را ارائه کنیم.

جدول ۳. مقایسه میان دادگان‌های ریخت‌شناسی عربی

دادگان	تنوع زبانی	تعداد کلمات	کاربرد	موضوع
نور-قطعه	رسمی	۲۲۳۶۹۰	جداسازی و برچسب‌گذاری	فقهی
بانک وابستگی عربی پراگ [۳۹]	رسمی	۱۱۴ هزار	برای استفاده عمومی	خبری
کلارا [۴۰]	رسمی	۳۷ ملیون	برچسب‌گذاری اجزای کلمات	مطالب نوشتاری متفرقه
بانک عربی پن [۴]	گویشی+ رسمی	بیش از ۱.۳ ملیون کلمه	جداسازی واژه، شرح انگلیسی	خبری
مجموعه قرآنی [۴۱]	رسمی	۱۷۴۵۵ کلمات متمایز	ساقه، بن و ریشه	قرآن
مجموعه عربی	رسمی	-	جداسازی و برچسب‌گذاری	قرآن

قرآنی [۴۲]				
مجموعه المصحف [۴۳]	رسمی	۱۷۴۵۵ کلمات متمایز	کلمات سطحی	قرآن
مجموعه قرآنی حیفا [۴۴]	رسمی	۷۷ هزار	واژگان، جنسیت، عدد	قرآن
مجموعه عربی قرآنی [۴۵]	رسمی	۷۷ هزار	برچسب‌گذاری بخشی از گفتار	قرآن

۶. روش‌های ارزیابی

برای اینکه این مجموعه داده به عنوان یک مجموعه معیار قابل قبول باشد، لازم است توسط روش‌های ارزیابی آن را محک بزنیم و نتایج خود را در ارتباط با دادگان‌های دیگر مورد بررسی و تحلیل قرار دهیم. بنابراین به روش ارزیابی می‌پردازیم.

۶. ۱ سامانه فراسه

فراسه[1] [۴۸] یک تقسیم‌ساز سریع و دقیق عربی است. رویکردش بر اساس رتبه SVM و با استفاده از بانک عربی پن آموزش داده شده است و با استفاده از کرنل‌های خطی است. برای ارزیابی فراسه، پژوهش‌گران آن را با دو تقسیم‌کننده مادامیرا [۳۰] و بخش عربی استانفورد [۴۹] مقایسه کردند. فراسه از هر دو تقسیم‌کننده برای وظایف بازیابی اطلاعات عملکرد بهتری دارد و با مادامیرا برای وظایف ترجمه ماشینی برابری می‌کند.

۷. آزمایش‌ها و نتایج

برای ارزیابی دقت تفکیک اجزای کلمه، خروجی‌های سامانه فراسه را بر روی تمامی کلمات موجود در دادگان نور-قطعه که به صورت دستی توسط متخصصان انسانی تهیه شده است مقایسه کردیم تا بتوانیم صحت[2]، دقت[3]، فراخوانی[4] و سنجه‌ی اف [5] را اندازه‌گیری کنیم. رابطه‌ی کلی برای محاسبه این سه معیار در ادامه نمایش داده شده است.

$$Precision = \frac{TP}{TP+FP} \quad (۱)$$

$$Recall = \frac{TP}{TP+FN} \quad (۲)$$

سنجه‌ی اف معیاری از دقت روش را ارائه می‌دهد که مجموعه‌ای از دقت و یادآوری را ارائه می‌دهد.

$$F-Measure = \frac{۲ * Precision * Recall}{precision + Recall} \quad (۳)$$

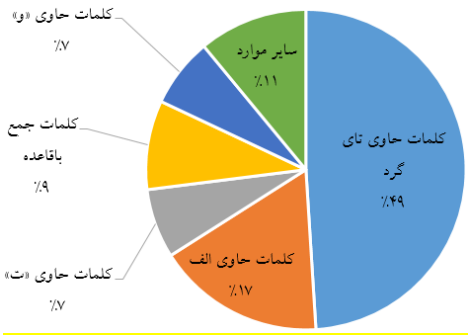
در این ارزیابی، دقت به صورت نسبت جداسازی‌هایی که توسط ماشین به درستی انجام شده‌اند، به کل جداسازی‌های موجود در دادگان تعریف می‌شود. معیار آنکه یک جداسازی به درستی انجام شده است یا خیر، تطبیق تمامی اجزای جدا شده با تمامی اجزای جداسازی انسانی می‌باشد. برای مثال جداسازی ماشینی عبارت «سَنُقَتِّلُهُم» زمانی به درستی انجام شده است که تمامی اجزای آن مانند «سَ»، «نُقَتِّلُ» و «هُم» به درستی جدا شده باشند. معیار فراخوانی نیز نسبت واژه‌های جداسازی شده به کل جداسازی‌های موجود در دادگان می‌باشد.

جدول شماره ۴ دقت تفکیک اجزای کلمه را برای سامانه فراسه نشان می‌دهد.

جدول ۴. دقت روش‌ها روی دادگان نور-قطعه

Segmenter	Precision	Recall	F-Score	Accuracy
Farasa	۰.۸۱	۰.۹۹	۰.۸۹	۰.۸۱

همانطور که مشاهده می‌شود، سامانه ۹۹ درصد واژه‌های موجود در دادگان را جداسازی کرده است، که از این میان ۸۱ درصد آنها به درستی جداسازی شده‌اند و سامانه ۱۹ درصد خطا داشته است. شکل ۳ دسته‌بندی و تحلیل خطای سامانه را نشان می‌دهد. مطابق شکل بیشترین خطا مربوط به دسته کلمات حاوی تای گرد می‌باشد.



شکل ۳. دسته‌بندی و تحلیل خطای سامانه‌ی فراسه

در جدول ۵ نمونه هر یک از دسته خطاها مشاهده می‌گردد.

[1] Farasa
[2] Accuracy
[3] Precision
[4] Recall
[5] F-Score

جدول ۵. نمونه هر یک از دسته خطاها

دسته خطا	جداسازی درست	خروجی فراسه
کلمات حاوی تای گرد	ال+طهارۀ	ال+طهار+ۀ
کلمات حاوی الف	موقفا	موقف+ا
کلمات حاوی «ت»	صلات+ه	صلا+ت+ه
کلمات جمع باقاعده	ال+مسجدین	ال+مسجد+ین
کلمات حاوی «و»	وجب	و+جب

۷. نتیجه‌گیری و کارهای آینده

در این پژوهش مجموعه داده نحوی و ریخت‌شناسی از کتاب شرایع را ارائه کردیم که یک کتاب تاریخی در زمینه احادیث است. حاشیه‌نویسی‌های صرفی دستی کتاب را با کمک متخصصین جمع‌آوری کردیم. نمونه‌ای از این مجموعه داده را برای حمایت از محققان علاقه‌مند به پردازش زبان طبیعی عربی در دسترس عموم قرار خواهیم داد. برای اینکه دادگان نور-قطعه را به عنوان یک دادگان استاندارد معرفی کنیم، یکی از روش‌های شاخص و استاندارد عربی یعنی فراسه را بر روی آن تست کردیم. نتیجه نشان می‌دهد دادگان ما رفتار مشابهی نسبت به سایر دادگان‌ها داشته است. در آینده، ما قصد داریم روشی نوین برای جداسازی واژگانی عربی ارائه کنیم و از این مجموعه‌ی داده به عنوان محکی برای ارزیابی روش خود و مقایسه با سایر روش‌های لبه‌ی دانش استفاده کنیم. همچنین نسخه‌های بعدی این مجموعه دادگان با تعداد برچسب‌های بیشتری را در اختیار دانش‌پژوهان قرار دهیم.

مراجع

[1] Buckwalter, T., Buckwalter Arabic morphological analyzer version 1.0. Linguistic Data Consortium, University of Pennsylvania, 2002.

[2] Buckwalter, T., Buckwalter Arabic morphological analyzer version 2.0. Linguistic data consortium, university of Pennsylvania, 2002. LDC cat alog no. 2004, Ldc2004l02. Technical report.

[3] Graff D, Maamouri M, Bouziri B, Krouna S, Kulick S, Buckwalter T. Standard arabic morphological analyzer (SAMA). Linguistic Data Consortium LDC2009E73, 2010.

[4] Maamouri, M., et al. The penn Arabic treebank: Building a large-scale annotated Arabic corpus. In NEMLAR conference on Arabic language resources and tools. 2004. Cairo.

[5] Elghamry, K. A constraint-based algorithm for the identification of Arabic roots. In Proceedings of the 1st Midwest Computational Linguistics Colloquium. 2004. Indiana Univ. Bloomington.

[6] Habash, N. and O. Rambow. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. 2006.

[7] Rodrigues, P. and D. Cavar, Learning Arabic morphology using statistical constraint-satisfaction models. Amsterdam studies in the theory and history of linguistic science series 4, 2007. 289: p. 63.

[8] Habash, N., Arabic morphological representations for machine translation, in Arabic computational morphology. 2007, Springer. p. 263-285.

[9] Smrz, O. ElixirFM–implementation of functional Arabic morphology. In Proceedings of the 2007 workshop on computational approaches to Semitic languages: common issues and resources. 2007.

[10] Daya, E., D. Roth, and S. Wintner, Identifying Semitic roots: Machine learning with linguistic constraints. Computational Linguistics, 2008. 34(3): p. 429-448.

[11] Roth, D. Learning to resolve natural language ambiguities: A unified approach. In AAAI/IAAI. 1998.

[12] Snyder, B. and R. Barzilay. Unsupervised multilingual learning for morphological segmentation. In Proceedings of acl-08: hlt. 2008.

[22] Graff, D., et al., Standard Arabic morphological analyzer (SAMA) version 3.1. Linguistic Data Consortium LDC2009E73, 2009: p. 53-56.

[23] Habash, N., R. Eskander, and A. Hawwari. A morphological analyzer for Egyptian Arabic. In Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology. 2012.

[24] Gridach, M. and N. Chenfour. Developing a new system for Arabic morphological analysis and generation. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP). 2011.

[25] Zalmout, N. and N. Habash. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[26] Boudlal, A., et al. Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts. In International Arab conference on information technology. 2010. Elsevier Science Inc New York, NY.

[27] Boudchiche, M., et al., AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. Journal of King Saud University-Computer and Information Sciences, 2017. **29**(2): p. 141-146.

[28] Zribi, I., M.E. Khemekhem, and L.H. Belguith. Morphological analysis of Tunisian dialect. In Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013.

[29] Alkuhlani, S. and N. Habash. A corpus for modeling morpho-syntactic agreement in Arabic: gender, number and rationality. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.

[30] Pasha, A., et al. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In LREC. 2014. Citeseer.

[13] Poon, H., C. Cherry, and K. Toutanova. Unsupervised morphological segmentation with log-linear models. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009.

[14] Boudlal, A., et al., A Markovian approach for Arabic root extraction. Int. Arab J. Inf. Technol., 2011. 8(1): p. 91-98.

[15] Attia, M., et al. An open-source finite state morphological transducer for modern standard Arabic. In Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing. 2011.

[16] Fullwood, M. and T. O'Donnell. Learning non-concatenative morphology. In Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL). 2013.

[17] Khaliq, B. and J.A. Carroll. Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. In Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013.

[18] Khalifa, S., S. Hassan, and N. Habash. A morphological analyzer for Gulf Arabic verbs. In Proceedings of the Third Arabic Natural Language Processing Workshop. 2017.

[19] Khalifa, S., N. Zalmout, and N. Habash. Morphological analysis and disambiguation for Gulf Arabic: The interplay between resources and methods. In Proceedings of the 12th Language Resources and Evaluation Conference. 2020.

[20] Taji, D., et al. An Arabic morphological analyzer and generator with copious features. In Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology. 2018.

[21] Khalifa, S., et al., A large scale corpus of Gulf Arabic. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.

[31] Habash, N. and O. Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05). 2005.

[32] Habash, N., O. Rambow, and R. Roth. MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt. 2009.

[33] Habash, N., et al. Morphological analysis and disambiguation for dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.

[34] Diab, M., K. Hacioglu, and D. Jurafsky, Automated methods for processing Arabic text: from tokenization to base phrase chunking. Arabic computational morphology: Knowledge-based and empirical methods. Kluwer/Springer, 2007.

[35] Alansary, S., Basma: Bibalex standard Arabic morphological analyzer. The Egyptian Journal of Language Engineering, 2016. **3**(1): p. 24-33.

[36] Khalifa, S., N. Zalmout, and N. Habash. Yamama: Yet another multi-dialect Arabic morphological analyzer. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. 2016.

[37] Kilany, H., et al., Egyptian colloquial Arabic lexicon. LDC catalog number LDC99L22, 2002.

[38] Alshargi, F., et al. Morphologicaly annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, Jordanian, Syrian, Iraqi and Moroccan. In Proceedings of the Fourth Arabic Natural Language Processing Workshop. 2019.

[39] Hajic, J., et al. Prague Arabic dependency treebank: Development in data and tools. In Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools. 2004.

[40] Zemánek, P. CLARA (Corpus Linguae Arabicae): An Overview. In Proceedings of ACL/EACL Workshop on Arabic Language. 2001.

[41] Zeroual, I. and A. Lakhouaja, A new Quranic Corpus rich in morphosyntactical information. International Journal of Speech Technology, 2016. 19(2): p. 339-346.

[42] Dukes, K. and N. Habash. Morphological Annotation of Quranic Arabic. In Lrec. 2010. Citeseer.

[43] Imad, Z. and L. Abdelhak, Al-Mus' haf Corpus: A New Quranic Corpus rich in Morphosyntactical Information and accurate Part of Speech tagging.

[44] Dror, J., et al., Morphological Analysis of the Qur'an. Literary and linguistic computing, 2004. 19(4): p. 431-452.

[45] Eric A., Corpus resources for learning Arabic to understand the Quran. Higher Education Academy workshop on "The Role of Corpora in LSP (Language for Specific Purposes) Learning and Teaching", 2012.

[46] Zeroual, I. and A. Lakhouaja. Clitiques-Stemmer: nouveau stemmer pour la langue Arabe. In The First National Doctoral Symposium on Arabic Language Engineering (JDILA'2014). 2014.

[47] Hegazi, M., A. Hilal, and M. Alhawarat, Fine-Grained Quran Dataset. International Journal of Advanced Computer Science and Applications, 2015. 6.

[48] Abdelali, A., et al. Farasa: A fast and furious segmenter for Arabic. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations. 2016.

[49] Monroe, W., S. Green, and C.D. Manning. Word segmentation of informal Arabic with domain adaptation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014.