به نام خداوند بخشنده مهربان

دانشکده ادبیات و علوم انسانی دکتر علی شریعتی

پایان‌نامه‌ی کارشناسی ارشد

گروه آموزشی زبان‌شناسی

گرایش زبان‌شناسی همگانی

تشخیص مرزهای تکواژی در زبان فارسی بر اساس اطلاعات واجی

(با هدف کاربرد در برنامه‌های رایانه‌ای پردازش زبان)

استاد راهنما:

دکتر اعظم استاجی

استاد مشاور:

دکتر سعید راحتی قوچانی

نگارش:

مریم شمایلی

تابستان ۱۳۹۰

تأیید به هیأت داوران

پایان نامه‌ی کارشناسی ارشد/دکتری

گروه آموزشی زبان‌شناسی

امضاکنندگان زیر، اعضای هیأت داوران پایان نامه‌ی آقای/خانم مریم شیرعلی با عنوان:

تشخیص مرزهای گروه‌های نحوی در زبان فارسی بر اساس اطلاعات نوایی با هدف کاربرد رایانه‌ای

در رشته‌ی زبان‌شناسی دانشکده‌ی ادبیات و علوم انسانی دانشگاه فردوسی، در جلسه‌ی دفاع حاضر شدند و پس از بررسی کامل، برابر آیین نامه‌ی مربوط، آن را با نمره‌ی ۱۹٫۳۱ و درجه‌ی عالی برای دریافت درجه‌ی ~~دکتری~~/ کارشناسی ارشد تأیید کردند.

- اعضای هیأت داوران:

استاد(ان) راهنما: دکتر اعظم استاجی — امضا

استاد مشاور ۱: دکتر سعید [illegible] — امضا

استاد مشاور ۲: دکتر — امضا

استاد داور ۱ (از دانشگاه): دکتر علی ایزانلو — امضا

استاد داور ۲ (از دانشگاه): دکتر شهلا شریفی — امضا

استاد داور ۳ (از دانشگاه): دکتر — امضا

- نماینده‌ی تحصیلات تکمیلی دانشگاه: دکتر شریفی — امضا

تاریخ: ۱۳۹۰/۷/۲۰
پیوست: دارد

دانشکدهٔ ادبیات و علوم انسانی دکتر علی شریعتی

باسمه تعالی

صورت جلسهٔ دفاع پایان‌نامهٔ گروه زبان‌شناسی

جلسهٔ دفاعیهٔ پایان‌نامهٔ تحصیلی خانم مریم شمایلی، دانشجوی کارشناسی ارشد، تحت عنوان «تشخیص مرزهای تکواژی در زبان فارسی بر اساس اطلاعات واجی (با هدف کاربرد در برنامه‌های رایانه‌ای پردازش زبان)» که استاد راهنمای آن سرکار خانم دکتر استاجی و استاد مشاور آقای دکتر راحتی و استادان منتخب (استاد مدعو) آقای دکتر ایزانلو و خانم دکتر شهلا شریفی می‌باشند، روز سه‌شنبه مورخ۱۳۹۰/۶/۲۲ ساعت ۱۲– ۱۰ در آزمایشگاه آواشناسی برگزار گردید. ابتدا خانم مریم شمایلی، خطابهٔ دفاعیهٔ خود را خواند و سپس هیأت داوران تذکّرات و راهنماییهای علمی خود را دربارهٔ محتوای پایان‌نامه ابراز نمودند. آن‌گاه پس از استماع توضیحات و دفاعیات دانشجو، داوران نظر خود را به شرح زیر اعلام داشتند:

پایان‌نامهٔ کارشناسی ارشد خانم مریم شمایلی، پس از بحث و تبادل‌نظر دربارهٔ محتوای آن با نمرهٔ (۱۹/۳۱) ارزیابی و پذیرفته اعلام گردید و نامبرده فارغ‌التحصیل شناخته شد.

استاد راهنما

دکتر استاجی

استاد مشاور

دکتر سعید راحتی

استاد منتخب۱ و نمایندهٔ تحصیلات تکمیلی

خانم دکتر شهلا شریفی

استاد منتخب۲

آقای دکتر علی ایزانلو

اظهارنامه

اینجانب مریم شمایلی دانشجوی دوره کارشناسی ارشد رشته‌زبان شناسی همگانی دانشکده ادبیات و علوم انسانی دکتر علی شریعتی دانشگاه فردوسی مشهد نویسنده پایان‌نامه **تشخیص مرزهای تکواژی در زبان فارسی بر اساس اطلاعات واجی (با هدف کاربرد در برنامه‌های رایانه‌ای پردازش زبان)** تحت راهنمایی خانم دکتر اعظم استاجی می‌باشم:

- تحقیقات در این پایان‌نامه، توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می‌باشد و مقالات مستخرج با نام «دانشگاه فردوسی مشهد» و یا «Ferdowsi University of Mashhad» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آوردن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ: امضای دانشجو:

مالکیت نتایج و حق نشر

- کلیه‌ی حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه فردوسی مشهد می‌باشد.

این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

تقدیم:

به پدرم

و مادرم،

برای مهربانی شان

تقدیر و تشکر

بر خود می دانم تا صادقانه از زحمات و لطف تمامی اساتید و عزیزانی که در این پژوهش مرا یاری نمودند تشکر نمایم:

استاد گرامی، سرکار خانم دکتر استاجی، قطعاً بدون راهنمایی های ارزنده ایشان این پایان نامه به انجام نمی رسید.

جناب آقای دکتر راحتی که زحمت مطالعه و مشاوره این پایان نامه را تقبل نمودند، و با همکاری صمیمانه شان مرا مورد لطف قرار دادند.

اساتید محترم گروه زبانشناسی دانشگاه فردوسی، که در تمام طول تحصیل در محضرشان درس گرفتم.

پدر و مادر عزیزم، همسر همیشه پشتیبانم، خواهران خوبم و دوستان همیشه همراهم که بودنشان همواره سبب دلگرمی من است.

بسمه تعالی

مشخصات رساله/پایان‌نامه تحصیلی دانشجویان

دانشگاه فردوسی مشهد

عنوان رساله/پایان‌نامه: تشخیص مرزهای تکواژی در زبان فارسی بر اساس اطلاعات واجی (با هدف کاربرد در برنامه‌های رایانه‌ای پردازش زبان)

نام نویسنده: مریم شمایلی **نام استاد راهنما:** دکتر اعظم استاجی **نام استاد مشاور:** دکتر سعید راحتی قوچانی

دانشکده: ادبیات و علوم انسانی	**گروه:** زبان‌شناسی	**رشته تحصیلی:** زبان‌شناسی همگانی
تاریخ تصویب:		**تاریخ دفاع:** ۱۳۹۱/۰۶/۲۲
مقطع تحصیلی: **کارشناسی ارشد** ☑	**دکتری** ☐	**تعداد صفحات: ۱۰۹**

چکیده

این پایان‌نامه با عنوان «تشخیص مرزهای تکواژی در زبان فارسی بر اساس اطلاعات واجی **(با هدف کاربرد در برنامه‌های رایانه‌ای پردازش زبان)**» به بررسی روش تجزیه ساختواژی پیشنهادی هریس (۱۹۹۵) و میزان کارآمدی این روش بر روی زبان فارسی می‌پردازد. هدف از انجام این تحقیق این است که با آزمودن مدل تجزیه ساختواژی هریس بر روی داده‌های نوشتاری زبان فارسی، امکان استفاده از آن را در برنامه‌های پردازش زبان فارسی برای تجزیه تکواژی مشخص کند.

برای آزمودن روش هریس از پنجاه جمله فارسی استخراج شده از پایگاه داده‌های زبان فارسی برای جامعه نمونه استفاده کردیم. به دلیل عدم وجود پیکره زبان فارسی مناسب برای آزمودن فرضیه این پژوهش از سه گویشور زبان فارسی برای جمع‌آوری سایر داده‌ها کمک گرفتیم. سپس داده‌های جمع‌آوری شده را به صورت دستی، واج‌نویسی کردیم و به شمارش تعداد متغیر همنشینی واج‌ها پس از هر واج پاره‌گفتارها پرداختیم. تقطیع پاره‌گفتارها در نقاط اوج تعداد متغیر همنشینی، تکواژهای به دست آمده از این روش تقطیع را مشخص نمود. نتایج این آزمون بر روی جامعه نمونه نشان می‌دهد که تقطیع ساختواژی با دقت ۹۷٪ و بازیابی ۷۵٪ صورت گرفته است. میزان دقت ۰/۹۷ این روش برای تعیین مرزهای تکواژی، نتیجه بسیار خوبی است. نکته مهم این است که این نتایج از پردازش یک پیکره برچسب‌گذاری نشده، به دست آمده است.

با اینکه در آزمودن روش هریس تکواژهای فارسی با دقت خوبی به دست می‌آیند به نظر می‌رسد این روش نمی‌تواند به عنوان مدل رایانه‌ای برای پردازش ساختواژی متون فارسی به کار گرفته شود و نتایج قابل قبولی را برای کاربرد در موتورهای جستجو، ماشین‌های ترجمه و یا دیگر برنامه‌های پردازش زبانی ارائه دهد. با این وجود از آن‌جا که در برنامه‌های تبدیل گفتار به متن، آواها به صورت خام به دست می‌آیند، به نظر می‌رسد این مدل بتواند متون آوانویسی شده فارسی را با دقت خوبی به تکواژها تجزیه کند و ازاین‌رو در برنامه‌های پردازش گفتار به کار رود.

کلید واژه‌ها:

۱. مرزهای تکواژی

۲. پردازش زبان

۳. ریشه‌یابی

٤. تقطیع ساختواژی

۵. الگوریتم

امضای استاد راهنما:

تاریخ:

فهرست مطالب

فصل اول: کلیات

فصل دوم: پیشینه تحقیق

فصل سوم: روش پژوهش و تحلیل داده‌ها

فصل چهارم: بحث و نتیجه‌گیری

فهرست جداول و نمودارها

فصل اول

کلیات

مقدمه

پیشرفت علم و گسترش آن در سراسر جوامع انسانی ثمره راه‌هایی است که انسان برای به دست آوردن اطلاعات و انتقال آن می‌پیماید. یکی از سریع‌ترین راه‌های انتقال اطلاعات استفاده از متون نوشتاری الکترونیکی است. اخبار، یافته‌های جدید علمی، مستندات تاریخی و هزاران مطلب دیگر به صورت متون نوشتاری در هر لحظه در دسترس میلیون‌ها کاربر اینترنت و دیگر شبکه‌های اطلاعاتی قرار می‌گیرند. گرچه به نظر می‌رسد اینترنت مشکل دسترسی به اطلاعات را حل می‌کند اما در واقع این طور نیست و همین حجم افزاینده اطلاعات، کاربر را در یافتن اطلاعات مورد نظرش دچار مشکل می‌کند. به این معنا که کاربر نیاز به برنامه‌های کمک کننده‌ای دارد که بتواند اطلاعات مورد نیازش را از میان میلیاردها صفحه اطلاعات غیر مرتبط بیابد. به این فرآیند بازیابی اطلاعات[1] می‌گویند. بازیابی اطلاعات از اطلاعات زبانی استفاده می‌کند و سعی در شناسایی و پردازش متون مربوط به موضوع مورد جستجو کاربر می‌کند (Han, 2011). یکی از مراحل پردازش متن، پردازش ساختواژی است. برنامه بازیابی اطلاعات باید این توانایی را داشته باشد که واژه مورد جستجو را با صورت‌های ساختواژی دیگر آن مرتبط کند. برای این منظور به تجزیه ساختواژی واژه می‌پردازد و وندها را از ریشه متمایز می‌سازد.

رشد روز افزون مستندات الکترونیکی به زبان فارسی، نیاز به برنامه‌های بازیابی اطلاعات که مختص زبان فارسی طراحی شده باشند را نمایان‌تر می‌سازد. در این پژوهش تلاش بر این است تا با آزمودن روش تجزیه ساختواژی پیشنهادی هریس[2] (Harris, 1955) بر روی داده‌های زبان فارسی، میزان کارآمدی این روش را برای زبان فارسی بسنجیم و در صورت امکان با اعمال تغییراتی این مدل را با در نظر گرفتن خصوصیات ساختواژی فارسی برای زبان فارسی مناسب‌سازی کنیم.

---

[1] Information Retrieval
[2] Harris

انتظار داریم که با استفاده از مدل هریس بتوان بسیاری از جملات فارسی را به تکواژهایشان تجزیه کرد. در این پژوهش سطح بررسی را پاره‌گفتار[3] قرار دادیم. به دلیل رسم‌الخط چسبان فارسی، در پژوهش‌های پیشین تشخیص واژه‌ها برای رایانه همیشه مسئله‌ای چالش برانگیز بوده‌است (بی‌جن‌خان، ۱۳۸۳). با در نظر گرفتن هر پاره‌گفتار برای تجزیه و تحلیل، رایانه نیازی به شناسایی واژه‌ها از یکدیگر ندارد و برای پردازش اجزاء بین فاصله دو نقطه در متن را انتخاب می‌کند. هدف از تجزیه یک پاره‌گفتار دست‌یابی به تمام تکواژهای آزاد و مقید آن است. در این روش تجزیه، واژه‌های بسیط بدون تغییر باقی می‌مانند و واژه‌های غیربسیط از محل مرزهای تکواژی‌شان[4] تقطیع می‌شوند. در این روش تنها از اطلاعات واجی برای تعیین مرزهای ساختواژی استفاده می‌شود به این معنا که تجزیه و تحلیل‌ها بر روی داده‌های خام که دربردارنده هیچ‌گونه اطلاعات زبانی نیستند، صورت می‌گیرد.

لازم به ذکر است که رویکرد این پژوهش هم‌زمانی است نه درزمانی. از همین رو، واژه‌هایی مانند «چوپان» و «داور» را بسیط به شمار می‌آوریم، هر چند که هر دو از دیدگاه درزمانی غیربسیط محسوب می‌شوند.

در این فصل به بیان سؤال‌هایی که در این تحقیق به آن‌ها پاسخ خواهیم داد، هدف از انجام این تحقیق و اهمیت موضوع تحقیق در زبان‌فارسی خواهیم پرداخت. روش جمع‌آوری داده‌ها و مشکلاتی که در طول کار با آن مواجه شدیم نیز در ادامه شرح داده می‌شوند. سپس برای روشن‌تر شدن برخی از اصطلاحات استفاده شده در این تحقیق، به تعریفی مختصر از هر یک می‌پردازیم و در انتهای فصل ساختار تحقیق و نحوه فصل‌بندی کل را بیان می‌کنیم.

---

[3] Utterance
[4] Morpheme Boundaries

۱-۱ سؤال تحقیق

سؤال‌های نظری:

۱- تا چه حد مدل هریس برای تشخیص مرز تکواژها در زبان فارسی کارآمدی دارد؟

۲- نقاط ضعف و قوت مدل هریس در تشخیص مرز تکواژها در زبان فارسی چیست؟

سؤال‌های پایه کاربردی:

۱- تعیین متغیر همنشینی حروف[5] در تکواژها تا چه حد بر تقطیع تکواژها تأثیر دارد؟

۲- تنوع حروف در مرز تکواژها چگونه است؟

۱-۲ هدف تحقیق

هدف از انجام این تحقیق این است که با آزمودن مدل تجزیه ساختواژی هریس بر روی داده‌های نوشتاری زبان فارسی، میزان کارآمدی این مدل برای زبان فارسی مشخص گردد. همچنین ویژگی‌های نوشتاری و ساختواژی زبان فارسی در کارکرد این مدل قابل توصیف هستند. به این معنی که در پی آزمودن این مدل، می‌توانیم ویژگی‌های نوشتاری و ساختواژی زبان فارسی که در این مدل تطابق نمی‌یابند را تبیین کنیم. هدف دیگر این است که نقاط ضعف و قوت فرضیه در تشخیص مرز تکواژها در زبان فارسی را مشخص کنیم.

[5] Letter Successor Variety

۱-۳ اهمیت موضوع

اهمیت تحقیقات مربوط به زبان فارسی، این میراث گرانقدر ایرانیان بر کسی پوشیده نیست. تا کنون پژوهش‌های فراوانی چه از جهت نظری و چه علمی‌بر روی این زبان صورت گرفته است. اما تغییرات زبانی و گسترش جنبه‌های کاربردی زبان، انجام پژوهش‌های تازه را ضروری می‌سازد. انتقال اطلاعات از طریق متون نوشتاری رایانه‌ای یکی از کاربردهای نو زبان است.

هدف از این پژوهش، بررسی روشی است که رایانه بتواند توسط آن واژه‌ها را به عنوان اجزای تشکیل دهنده متون الکترونیکی، تجزیه و شناسایی کند. در سال‌های اخیر تحقیقات دیگری نیز رویکردهای متفاوت ریشه‌یابی را بر روی زبان فارسی مورد بررسی قرار داده‌اند که در فصل دو همراه با نقاط قوت و ضعف‌شان معرفی خواهند شد. نگاهی بر پژوهش‌های پیشین، مشخص می‌کند که اغلب آن‌ها از روش‌های مبتنی بر قاعده[6] استفاده نموده‌اند. از مهم‌ترین نقاط ضعف این پژوهش‌ها این است که در ریشه‌یابی صورت‌های بی‌قاعده زبان دچار مشکل می‌شوند زیرا این صورت‌ها در قواعد برنامه طراحی‌شده تعریف نشده‌اند (Tashakori and others, 2002). دیگر مشکلاتی که در پردازش زبان فارسی وجود دارد، مشکلاتی است که رسم‌الخط فارسی به وجود می‌آورد. در روش‌های ریشه‌یابی که تا کنون برای فارسی پیشنهاد شده است، پردازش بر روی تک تک واژه‌ها صورت می‌گیرد و این مورد باعث بروز مشکلاتی در شناسایی واژه‌ها می‌شود. تنوع شیوه‌های نوشتاری در زبان فارسی نیز بر این مشکلات دامن می‌زند (بی‌جن‌خان،۱۳۸۳). مشکلاتی مانند تنوع نحوه به کار بردن بعضی پیشوندها و پسوندها از جمله نحوه استفاده از «می» چسبان و غیر چسبان، در واژه‌های «می‌تواند» و «میتواند»، نحوه به کار بردن «ها» چسبان و غیر چسبان، مثل «آن‌ها»، «آنها» و «آن‌ها»، تنوع در نگارش واژه‌های مرکب مثل «همین که» و «همینکه» یا «راه گشا» و «راهگشا». از نقاط ضعف دیگر این روش‌ها این است که قواعد به کار رفته در آن‌ها تأثیر بسیار زیادی بر نتایج دارند و در واقع نتایج کار وابسته به تعداد

---

[6] Rule Based Approaches

محدودی قاعده است. در صورتیکه این قواعد از دقت کـافی برخـودار نباشـند، باعـث اشـتباه در بخش نسبتاً زیادی از نتایج می‌شوند.

امروزه برخی از مهم‌ترین رویکردهای بازیابی اطلاعات از روش‌هایی استفاده مـی‌کننـد کـه در آن‌ها قواعد جداسازی وندها[۷] توسط رایانه و به صورت خودکار از متن استخراج مـی‌شـوند. ایـن روش‌ها را جزء دسته روش‌های بدون نظارت[۸] پردازش زبان بـه حسـاب مـی‌آورنـد. توانـایی در پردازش صورت‌های بی‌قاعده و کم کاربرد زبان از جمله مزایای به کارگیری این روش‌هاست. تـا آنجایی که تحقیقات نگارنده نشان می‌دهد، تا کنون روش‌های بدون نظارت پردازش ساختواژی و یا حداقل روش به کار رفته در این پژوهش، برای ریشـه‌یـابی متـون فارسـی مـورد بررسـی قـرار نگرفته‌اند. از این رو در این تحقیق به بررسی کارکرد روش تقطیع سـاختواژی بـدون نظـارت بـر زبان فارسی می‌پردازیم. روشی که در این پژوهش به کار می‌گیریم بر اساس رویکردی اسـت کـه هریس برای تجزیه ساختواژی متون نوشتاری زبان به صورت دستی معرفـی کـرده اسـت. سـطح پردازش زبانی در این روش، پاره‌گفتار می‌باشد. از این رو مشکلات روش‌های قبل را در تشخیص واژگان متن نخواهد داشت. این روش همچنین در سال‌های اخیر برای ریشه‌یابی زبان‌های اروپایی به کار رفته و نتایج خوبی را نشـان داده اسـت. در فصـول آینـده مـوارد کـاربرد ایـن روش را در زبان‌های دیگر بررسی خواهیم کرد. امید است که این رویکرد بتواند قدمی هر چند کوچک در راه طراحی برنامه‌های پردازش زبان فارسی باشد.

[7] Affix Stripping
[8] Unsupervised Methods

۱-٤ روش جمع‌آوری داده‌ها

برای آزمایش فرضیه تحقیق نیاز به پیکره[۹] بسیار بزرگ و متنوعی از متون نوشتاری زبان فارسی داریم که در بردارنده تمام و یا اغلب رشته حروف پاره‌گفتارهای محتمل زبان باشد. این پیکره همچنین باید قابلیت جستجو واج‌ها و رشته‌های واجی[۱۰] را داشته باشد. همچنین داده‌های پیکره باید به صورت واج‌نویسی شده قابل دسترسی و جستجو باشند. پیکره‌های زبانی که تا کنون برای زبان فارسی ساخته شده‌اند، هیچکدام خصوصیات مورد نظر برای انجام این پژوهش را ندارند. از این رو برای جمع‌آوری داده‌های این تحقیق، به مانند روش جمع‌آوری داده‌ها در پژوهش اولیه‌ای که هریس انجام داده بود، از گویشوران زبان کمک گرفتیم. در طول انجام کار از سه گویشور زبان فارسی کمک گرفتیم. از گویشوران خواستیم که برای هر رشته واجی ارائه شده پاره‌گفتارهایی از زبان را بنویسند که با رشته واجی مورد نظر آغاز شده باشند.

تعداد ۵۰ جمله فارسی را از پایگاه داده‌های زبان فارسی گردآوری کردیم. برای انتخاب این جملات محدودیت‌هایی را قرار دادیم. اول اینکه همگی جملات را از متون فارسی امروز انتخاب کردیم. دوم اینکه طولی برابر ۱ تا ٦ واژه برای جملات در نظر گرفتیم و جملاتی را که بیش از ٦ واژه طول داشت، حذف کردیم. این محدودیت به این دلیل ایجاد شد که شمارش متغیر همنشینی حروف در دنباله‌های واجی بلندتر به صورت دستی کار مشکلی است و احتمال خطا در شمارش‌های اولیه را بالا می‌برد.

شیوه پرسش از گویشوران از روایی داده‌ها نمی‌کاهد؛ زیرا در جمع‌آوری داده‌ها هیچ‌گونه اطلاعات ساختواژی از گویشور خواسته نمی‌شود و تمام پاره‌گفتارها به طور یکسان و هر کدام به صورت یک رشته واجی ممتد نوشته می‌شوند. از سوی دیگر گویشور تنها پاره‌گفتارهایی را تولید می‌کند که به طور معمول در زبان اتفاق می‌افتد. ذهن انسان مانند یک پیکره بزرگ و قابل جستجو

---

[9] Corpus
[10] Phoneme Sequence

عمل می‌کند (بی‌جن‌خان، ۱۳۸۳). در بخش بعدی به مشکلاتی که در جمع‌آوری داده‌ها با آن‌ها مواجه شدیم و دیگر دشواری‌های تحقیق خواهیم پرداخت.

۱-۵ مشکلات تحقیق

مهم‌ترین مشکل در انجام این پژوهش نبود پیکره نوشتاری مناسبی در زبان فارسی برای جمع‌آوری داده‌هاست. همان طور که پیش‌تر هم گفتیم هیچکدام از پیکره‌های موجود برای زبان فارسی قابلیت جستجو در متون را ندارند. برای حل مشکل جمع‌آوری داده‌ها از گویشوران زبان کمک گرفتیم. گرچه این روش تأثیری بر نتایج اولیه نداشت، اما دشواری‌هایی را در طول کار به وجود آورد. گویشوران برای یادآوری پاره‌گفتارهای مورد نظر نیاز به تمرکز و وقت کافی داشتند. همچنین تمامی پاره‌گفتارهای جمع‌آوری شده برای هر یک از واج‌ها در مرحله اول باید ثبت و سپس واج‌نویسی می‌شد. بدیهی است که استفاده از یک پیکره بزرگ و قابل جستجو بسیاری از این مشکلات را رفع می‌کرد.

برخی دیگر از مشکلات، چالش‌هایی است که شیوه رسم‌الخط فارسی به وجود می‌آورد. عدم ظاهر شدن واکه‌های کوتاه در نوشتار به معنای حذف صورت برخی از پرکاربردترین واج‌ها در نوشتار است. این مسئله سبب تفاوت زیاد بین صورت نوشتار عادی و صورت واج‌نویسی شده‌این متون می‌شود و عملاً استفاده از متون عادی -با کمی چشم پوشی از تفاوت‌ها- را غیرممکن می‌سازد. اگر این تفاوت چشم‌گیر نبود و می‌توانستیم از متون عادی برای جمع‌آوری داده‌ها استفاده کنیم، آن‌گاه موتورهای جستجوی اینترنتی می‌توانستند در جمع‌آوری داده‌ها کمک کننده باشند زیرا این موتورها قابلیت جستجو در متون عادی موجود بر روی وب را دارند. برای غلبه بر این مشکل تمام پاره‌گفتارهای تولید شده گویشوران، توسط نگارنده واج‌نویسی شد.

نبود فرهنگ بین رشته‌ای زبان‌شناسی رایانه‌ای برای یافتن معادل‌های فارسی واژگان تخصصی که در متون انگلیسی به کار رفته‌اند، نیز یکی دیگر از مشکلاتی است که در طول کار با آن مواجه