



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم انسانی و مطالعات فرهنگی

پژوهشکده زبان شناسی

پایان نامه

برای دریافت مدرک کارشناسی ارشد زبان شناسی همگانی

عنوان پژوهش :

بن واژه سازی اسم در پردازش خودکار متن فارسی نوشتاری امروز
(با استفاده از معیارهای زبان شناختی)

استاد راهنما: جناب آقای دکتر سید مصطفی عاصی

استاد مشاور: جناب آقای دکتر یحیی مدرسی

نگارش: افسانه شاه آبادی



تقدیر:

"به پایان آمد این دفتر حکایت همچنان باقی"ست
همه دردی کشان مدهوش هشیوار همچنان ساقیست

بضاعت سخت ناکافی و طوفان سخت هائل بود
از این رزق فزون دانم که بالا دست رزاقیست

هر آن کس دست یاری داد برای خلق این مکتوب
حسابش با خدا باشد که او دلباز خلاقیست

بدینوسیله مراتب قدردانی بی ریای خود را از تمامی کسانی که به دور از کلیه وابستگی های مادی صادقانه
یاریم کردند بعمل می آورم

۱-جناب آقای دکتر عاصی: استاد محترم راهنما که با سعه صدر و بی کم و کاست برایم وقت اختصاص
دادند؛

۲- جناب آقای دکتر مدرسی : استاد محترم مشاور که عالمانه مشاورم بودند؛

۳- سرکار خانم دکتر کلباسی :استاد محترم داور که اغماض را چاشنی داوری مدقانه خویش نمودند؛

۴-دخترم دنیز که با لبخندهای صبورانه شیرینش و گاه با پاره کردن اوراق تحقیق مادر وقتی را که کاملاً مال
او بود از همان اوان زندگیش وقف علم کرد؛

۵-سرکار خانم سلیقه که با تمام وجود سعی در جبران تاخیر چند ساله من با شب بیداری ها و تایپ متن
حاضر داشتند؛

۶ - و همسر من که اگر نبود همراهی و مهربانی او، مرا هرگز توان به پایان رساندن این رساله میسر
نمی شد .

چکیده:

امروزه اهمیت و کارایی فرایند بن واژه‌سازی چه بر افرادی که با تهیه و تدوین فرهنگ‌ها، واژه‌نامه‌های بسامدی، فهرست‌های راهنما یا واژه‌نما، ترجمه ماشینی، پیکره‌های زبانی، پایگاه داده‌های زبانی و... سر و کار دارند و چه بر آنهایی که به عنوان کاربر این حوزه‌ها از بسته نرم‌افزارهای حاوی برنامه بن واژه‌سازی استفاده می‌کنند، دیگر پوشیده نیست.

سؤال این است که با این همه کاربردهایی که فرایند بن واژه‌سازی در حوزه های مختلف زبان‌ها می‌تواند داشته باشد و با توجه به حجم کارهایی که در مورد زبان‌های انگلیسی، فرانسوی، آلمانی و جز آن در این زمینه انجام گرفته است، آیا برای زبان فارسی هم در این خصوص اهتمام بایسته صورت گرفته است؟ با نگاهی به محدودیت کارهایی که از نظر گستره و تعداد برای زبان فارسی انجام گرفته است به خصوص در مقایسه با زبان‌هایی که از آنها یاد شد، و نیز با توجه به گسترش روزافزون پردازش اطلاعات با رایانه، سیستم‌های ترجمه ماشینی و سیستم‌های هوشمند، کاربردهای هوش مصنوعی و برنامه‌های واژه‌پرداز رایانه‌ای، بازیابی اطلاعات از متون به کمک رایانه و هزاران مورد مشابه، اهمیت این قبیل کارهای تحقیقاتی تا حد قابل توجهی آشکار می‌گردد.

در این رساله که موضوع آن بن واژه‌سازی اسم در پردازش خودکار متن فارسی نوشتاری امروز است، هدف ارائه انگاره‌ای نظری برای برنامه رایانه‌ای است که بتواند با استفاده از معیارهای زبان شناختی (صرفی و نحوی)، ضمن شناسایی اسم‌ها، مشخصات آنها از جمله مفرد و جمع بودن و یا ساده و مرکب و مشتق بودن و ... را به کاربر بدهد و با مد نظر قرار دادن هر دو رویکرد و نظریه رایج در بن واژه‌سازی که در یکی از آنها بن واژه را مساوی ستاک و در دیگری آن را معادل ریشه واژه‌ها در نظر می‌گیرند، در سه لایه اقدام به استخراج بن واژه اسم‌ها نماید. بر همین اساس پس از شناسایی ابتدایی آنها از طریق جستجو در فهرست اسم‌ها (پیوست شماره ۲) و ابهام زدایی بین برخی اسم‌های هم‌نویسه با واژه‌ها و وندهای پرسامدی چون پیشوندهای «هم» و «می»، در لایه اول به جداسازی وندهای تصریفی و هم‌چنین عناصری همت گمارده شد که گاه ممکن است در نتیجه سرهم نویسی خط فارسی به اسم‌ها متصل شوند (مانند که، را و...) و در لایه دوم نیز سعی شد تا اسم‌های مشتق از صورت اشتقاقی خود تهی شده و به ریشه‌شان تقلیل پیدا کنند و بالاخره در لایه سوم که در واقع مکمل دو لایه قبل به شمار می‌رود به شناسایی اسم‌های مرکب پرداخته شد. در پژوهش حاضر سعی بر این بوده است که «دستورالعمل‌ها» در نمودارهای گردش کار طوری تنظیم شوند که اولاً نیاز به تعامل و محاوره با کاربر به حداقل ممکن برسد و ثانیاً غالب مشخصات مربوط به صورت‌ها و حالات مختلف اسم‌ها در آنها لحاظ گردد. ضمن این که بتوان برنامه بن واژه‌سازی اسم را

مستقل و بدون نیاز به یک برنامه تقطیع و برچسب‌دهی پیش برد و البته چنانچه در آینده برنامه تقطیع و برچسب‌دهی کارآمدی برای زبان فارسی تهیه شود امکان تلفیق با آن را هم داشته باشد.

در واقع این انگاره طوری طراحی شده است که یک برنامه نویس رایانه بتواند آن را به راحتی به یک زبان برنامه نویسی رایانه‌ای تبدیل نماید. اصولاً هدف افزودن جزء مکمل دیگری به مجموعه کارهایی است که در راستای تحلیل و پردازش خودکار متن فارسی صورت گرفته است، و یا پایه و مبنایی برای کارهای آتی در این زمینه است. بی‌گمان نیاز به توضیح نیست که مقوله اسم نیز تا چه اندازه از گستردگی و پیچیدگی و تنوع در مقایسه با سایر مقولات برخوردار می‌باشد.

این پایان نامه از چهار فصل، فهرست منابع و چهار پیوست تشکیل شده است، ضمناً فایل برنامه به صورت دیسک فشرده نیز قابل دسترسی است.

کلید واژه ها:

- ۱- اسم ۲- بن واژه ۳- بن واژه سازی ۴- پردازش خودکار ۵- فارسی معیار

فهرست مطالب

عنوان	صفحه
تقديم-----	پ
تقدير-----	ت
چكیده-----	ث
فهرست مطالب-----	ج
فصل اول: کلیات پژوهش-----	۱
۱-۱- مقدمه-----	۲
۱-۲- پیشگفتار-----	۲
۱-۳- بیان مسئله، اهداف و ضرورت تحقیق-----	۵
۱-۴- تعریف کلیدواژه ها-----	۷
۱-۴-۱- اسم-----	۷
۱-۴-۲- بن واژه-----	۷
۱-۴-۳- بن واژه سازی-----	۸
۱-۴-۴- پردازش خودکار-----	۸
۱-۴-۵- فارسی معیار-----	۸
۱-۵- سئوالات اصلی تحقیق-----	۹
۱-۶- فرضیات پژوهش-----	۹
۱-۷- شیوه تحقیق-----	۹
۱-۷-۱- بخش اول-----	۹
۱-۷-۲- بخش دوم-----	۱۰
۱-۷-۳- بخش سوم-----	۱۰
۱-۷-۴- بخش چهارم-----	۱۰
۱-۸- ساختار رساله-----	۱۰

عنوان	صفحه
فصل دوم: زمینه تحقیق-----	۱۲
۱-۲ کاربرد رایانه در زبانشناسی-----	۱۳
۱-۲-۱-۱ مقدمه-----	۱۳
۱-۲-۲ رایانه و ترجمه ماشینی-----	۱۴
۱-۲-۳ نقش رایانه ها از پیکره زبانی تا پایگاه داده های زبان	۱۹
۱-۲-۴ نقش رایانه ها در فرهنگ نگاری-----	۲۵
۱-۲-۴-۱ Concordances (فهرست های واژه نما)-----	۳۲
۱-۲-۲ پیشینه تحقیق-----	۳۵
۱-۲-۲-۱ مقدمه-----	۳۵
۱-۲-۲-۲ پژوهشهای غیر ایرانیان در مورد اسم و گروههای اسمی	
فارسی-----	۳۵
۱-۲-۲-۲-۱ لمبتون (۱۹۵۳)-----	۳۵
۱-۲-۲-۲-۲ لازار (۱۹۵۷)-----	۳۶
۱-۲-۲-۲-۳ بویل (۱۹۶۶)-----	۳۸
۱-۲-۲-۲-۴ لر (۱۹۸۰)-----	۳۹
۱-۲-۲-۳ پژوهش های ایرانیان در مورد اسم و گروههای اسمی	۴۱
فارسی-----	
۱-۲-۲-۳-۱ باطنی (۱۳۴۸)-----	۴۱
۱-۲-۲-۳-۲ وزین پور (۱۳۶۶)-----	۴۶
۱-۲-۲-۳-۳ حسینیان (۱۳۶۸)-----	۴۶
۱-۲-۲-۳-۴ کشانی (۱۳۷۱)-----	۴۹
۱-۲-۲-۳-۵ کلباسی (۱۳۷۱)-----	۴۹
۱-۲-۲-۳-۶ عماد افشار (۱۳۷۲)-----	۵۱
۱-۲-۲-۳-۷ جعفر بگلو (۱۳۷۳)-----	۵۳

- ۵۶ -----۲-۲-۳-۸- خانلری (۱۳۷۷)
- ۵۷ -----۲-۲-۳-۹- مشکوه الدینی (۱۳۷۳)
- ۶۱ -----۲-۲-۳-۱۰- ارزشنگ (۱۳۷۴)
- ۶۳ -----۲-۲-۳-۱۱- طباطبایی (۱۳۸۲)
- ۶۸ -----۲-۲-۴- بررسی پایاننامه ها
- ۶۸ -----۲-۲-۴-۱- حاجی عبدالحسینی (۱۳۷۵)
- ۷۰ -----۲-۲-۴-۲- سمایی (۱۳۷۷)
- ۷۲ -----۲-۲-۴-۳- فرخ (۱۳۸۱)
- ۷۳ -----۲-۲-۴-۴- دانش کار آراسته (۱۳۸۱)
- ۷۴ -----۲-۲-۴-۵- ذوالفقاری (۱۳۸۱)
- ۷۶ -----۲-۲-۵- سایرین

عنوان	صفحه
فصل سوم: گستره پژوهش	۷۹
۳- ۱- مقدمه	۸۰
۳- ۲- چارچوب نظری	۸۰
۳- ۲- ۱- برخی مفاهیم پایه در ساختواژه	۸۰
۳- ۲- ۱- ۱- ریشه	۸۰
۳- ۲- ۱- ۲- ستاک	۸۰
۳- ۲- ۱- ۳- پایه	۸۱
۳- ۲- ۱- ۴- وند و شبه وند	۸۱
۳- ۲- ۱- ۴- ۱- وند	۸۱
۳- ۲- ۱- ۴- ۲- شبه وند	۸۲
۳- ۲- ۱- ۵- واژه بست و ادات	۸۳
۳- ۲- ۱- ۵- ۱- واژه بست	۸۳
۳- ۲- ۱- ۵- ۲- ادات	۸۳
۳- ۲- ۲- دستگاه صرف یا ساختواژه	۸۳
۳- ۲- ۲- ۱- تصریف	۸۴
۳- ۲- ۲- ۱- ۱- پسوندها یا نشانه های جمع	۸۴
۳- ۲- ۲- ۱- ۲- علامت نکره	۸۶
۳- ۲- ۲- ۱- ۳- الف ندا	۸۷
۳- ۲- ۲- ۱- ۴- نشانه های تانیث عربی	۸۷
۳- ۲- ۲- ۱- ۵- سایر وندهای تصریفی	۸۷
۳- ۲- ۲- ۲- واژه سازی	۸۹
۳- ۲- ۲- ۲- ۱- اسم بسیط (ساده)	۸۹
۳- ۲- ۲- ۲- ۲- اسم مشتق	۹۰
۳- ۲- ۲- ۲- ۳- وندافزایی	۹۲

- ۹۲ ----- پیشوندهای اشتقاقی ۱-۱-۲-۲-۲-۲-۳
- ۹۳ ----- پسوندهای اشتقاقی ۲-۱-۲-۲-۲-۲-۳
- ۹۴ ----- تبدیل یا اشتقاق صفر-۲-۲-۲-۲-۲-۳
- ۹۵ ----- اشتقاق پسرو یا پس ساز-۳-۲-۲-۲-۲-۳
- ۹۶ ----- اسم مرکب-۳-۲-۲-۲-۲-۳
- ۹۸ ----- اسم های مرکب اضافه ای و اضافه مقلوبی-۱-۳-۲-۲-۲-۳
- ۹۸ ----- اسم های مرکب پیش وابسته ای-۲-۳-۲-۲-۲-۳
- ۹۹ ----- اسم های مرکب پیوندی-۳-۳-۲-۲-۲-۳
- ۱۰۲ ----- اسم های مرکب حرف اضافه دار-۴-۳-۲-۲-۲-۳
- ۱۰۳ ----- اسم های مرکب شبه جمله ای و جمله گونه -۵-۳-۲-۲-۲-۳
- ۱۰۳ ----- اسم های مرکب هم بافتی و فعلی-۶-۳-۲-۲-۲-۳
- ۱۰۴ ----- اسم مرکب- مشتق-۴-۲-۲-۲-۲-۳
- ۱۰۷ ----- تفاوت اسم های مرکب و گروههای نحوی-۵-۲-۲-۲-۲-۳
- ۱۰۸ ----- توضیح ساختار نظری برنامه-۳-۳-۲-۲-۲-۳
- ۱۱۰ ----- وارد کردن متن بعنوان درونداد-۱-۳-۳-۲-۲-۲-۳
- ۱۱۰ ----- شناسایی واژه ها-۲-۳-۳-۲-۲-۲-۳
- ۳-۳-۳- لایه اول: بررسی ابتدایی واژه ها و جدا کردن وندهای
- ۱۱۰ ----- تصریفی
- ۳-۳-۴- لایه دوم: جداسازی وندهای اشتقاقی و اجزای چسبیده
- ۱۲۵ ----- اسم های مرکب
- ۱۲۷ ----- لایه سوم: پیدا کردن اسم های مرکب-۵-۳-۳-۲-۲-۲-۳

صفحه	عنوان
۱۳۴	فصل چهارم: نتایج و یافته ها و پیشنهادی برای پژوهش های آتی ----
۱۳۵	۴- ۱- نتایج و یافته ها-----
۱۳۸	۴- ۲- پیشنهادی برای پژوهش های آتی -----
۱۳۹	فهرست منابع-----
۱۴۴	پیوستها:-----
۱۴۵	پیوست ۱: نمودارهای گردش کار-----
۱۹۱	پیوست ۲: فهرست اسم ها -----
۳۴۷	پیوست ۳: جداول -----
۳۷۰	پیوست ۴: واژه نامه -----
۳۷۱	چکیده انگلیسی-----

فصل اول

کلیات پژوهش

۱-۲-۱- مقدمه

در این فصل سعی بر آنست که پس از ارائه پیشگفتار (ن. ک ۲-۱) به بیان مسئله، اهداف و ضرورت انجام تحقیق پرداخته شود (ن. ک ۳-۱)، تعریفی از کلید واژه‌ها ارائه (ن. ک ۴-۱)، سؤالات اصلی پژوهش طرح (ن. ک ۵-۱) و فرضیات آن تبیین گردد (ن. ک ۶-۱) و نهایتاً روش تحقیق شرح داده شود (ن. ک ۷-۱) و چارچوب ساختار اصلی رساله بیان گردد (ن. ک ۸-۱).

۲-۱- پیشگفتار

نیافتن مدخلی جداگانه در توصیف اصطلاح «بن واژه» یا «بن واژه‌سازی»^۱ در دایره‌المعارف ۱۱ جلدی آشر^۲ (۱۹۹۴) و بسنده کردن به تعریفی کوتاه از این اصطلاح در ضمیمه آن دایره‌المعارف، شاید دلیلی محکم بر این ادعا باشد که دیر زمانی و چه بسا بیشتر از دو دهه از ساخته شدن و ورود این اصطلاح به معنای امروزی آن به عرصه مطالعات زبان و زبان‌شناسی نمی‌گذرد.

با این حال با توجه به پیشرفت‌های روزافزون در حوزه زبان و زبان‌شناسی در عصر حاضر که عمدتاً مرهون بکارگیری رایانه و پیشرفت در سایر علوم بخصوص علم کامپیوتر می‌باشد، جای تعجب نیست که در اندک زمانی با حجم زیادی از کارها و پژوهش‌ها در خصوص «بن واژه‌سازی» و بکارگیری هر چه بیشتر آن در زیر شاخه‌های زبان‌شناسی از جمله فرهنگ نگاری، ترجمه ماشینی، زبان‌شناسی پیکره‌ای و مهم‌تر از همه تنوع و تعدد نرم‌افزارهای «بن واژه‌سازی» در مورد زبان‌های مختلف مواجه شویم.

البته، از مطالب مقاله جان میلز^۴ (۱۹۹۶) با عنوان «نقش رایانه‌ها در زبان‌شناسی کاربردی» که در آن از ثبت داده‌ها بر روی برگیزه‌ها و تنظیم آنها بر اساس «بن واژه» به عنوان شیوه سنتی و معمول تنظیم محتوا و اختار فرهنگ‌ها یاد شده است، چنین استنباط می‌شود که اولاً این فرایند پیشتر نیز وجود داشته ولیکن به صورت دستی انجام می‌گرفته است و ثانیاً آغاز کاربرد آن عمدتاً در فرهنگ نگاری بوده است (جان میلز ۱۹۹۶، ص ۲۶۳).

در همان مقاله به نقش رایانه‌ها و پردازش الکترونیکی^۵ در تسهیل بخشیدن به فرایند «بن واژه‌سازی» اشاره شده است.

به گفته او رایانه‌ها این عمل را به سه شیوه انجام می‌دهند:

الف) ممکن است «بن واژه» به طور خودکار و مطابق قواعدی به یک صورت واژگانی پیوند بخورد،

-
1. Lemma
 2. Lemmatization
 3. Asher
 4. Jon Mill
 5. Electronic Processing

ب) یا متن پیش - ویرایش گردد به طوری که در کنار هر واژه، بن واژه آن نیز آورده شود،
ج) و یا از جستجو در یک واژه‌نامه قابل خوانش برای رایانه، در بررسی و تطابق استفاده گردد (جان میلز ۱۹۹۶، ص ۲۶۳).

وی معتقد است که بن واژه آن واژه پایه است که در فرهنگ‌ها به عنوان مدخل یا سرواژه آورده می‌شود و بن واژه‌سازی فرایندی است که به کمک آن بتوان صورت‌های صرف شده را زیر صورت پایه برد به عبارتی دیگر بن واژه‌سازی فرایندی است برای پیدا کردن صورت پایه از تمام صورت‌های صرفی که به هم مرتبطند (جان میلز ۱۹۹۶، ص ۲۶۳ و ۲۵۳). مثلاً در انگلیسی فعل‌ها را زیر صورت ساده فعل و مصدر بدون to می‌برند. یعنی به فرض «bring»، «brought»، «bringing»، «brings» را زیر «bring» می‌برند و یا می‌توان گفت که بن واژه همه این‌ها ((bring)) می‌باشد.

افراد دیگری نیز از جمله سوزان هاکی^۱ (۱۹۹۸) تعاریفی کمابیش مشابه با آنچه که در بالا گفته شد از «بن واژه‌سازی» ارائه کرده‌اند. در واقع آنها «بن واژه» را برابر با «ستاک»^۲ فرض می‌کنند.

هر چند برخی نیز از جمله دیواین^۳ (۱۹۸۷، ص ۵۸) معتقدند که برای رسیدن به «بن واژه»، واژه‌ها نه تنها باید از صورت‌ها و وندهای تصریفی عاری شوند بلکه این مسئله صورت‌ها و وندهای اشتقاقی را نیز شامل می‌شود. در حقیقت این گروه «بن واژه» را معادل «ریشه»^۴ می‌گیرند.

رساله حاضر که موضوع آن «بن واژه‌سازی اسم در پردازش خودکار متن فارسی نوشتاری امروز با استفاده از معیارهای زبان شناختی» است، سعی دارد هر دو رویکرد را مد نظر قرار داده و در دو لایه اقدام به استخراج بن واژه‌ها نماید. بدین نحو که در نخستین لایه فقط وندهای تصریفی (مانند پسوندهای جمع، یاء نکره، الف منادا و ...) از اسم‌ها جدا می‌شوند. البته بی‌گمان جداسازی عناصری هم که بی‌هیچ دلیل قانع کننده و صرفاً از رهگذر بی‌قاعدگی‌های خط فارسی و تمایل برخی به سرهم نویسی و عدم رعایت قواعد ساختوازی به اسم‌ها یا به طور کلی واژه‌ها می‌چسبند، مانند (که) و (را) در این لایه صورت می‌گیرد.

لایه دوم که پس از لایه اول یعنی پس از جدا شدن وندهای تصریفی وارد عمل می‌شود. به جدا کردن وندهای اشتقاقی و یافتن ریشه واژه‌های مشتق، مرکب و مشتق □ مرکب اختصاص می‌یابد.

و اما چرا اسم و آن هم زبان فارسی؟

-
1. Susan Hockey
 2. Stem
 3. Devine
 4. Root

اسم، چون در میان سایر اجزاء کلام در هر زبانی اسم‌ها از گستردگی و تنوع بیشتری برخوردارند و البته بی‌قاعدگی‌ها در ساخت اسم‌ها، تعداد بسیار زیاد آنها در مقایسه با فعل، صفت، قید و ترکیبات بی‌شمار اسمی را نیز باید به ویژگی‌های فوق افزود و زبان فارسی، چون زبان رسمی ماست و کوشش در اعتلای آن بر همه ما بخصوص زبان‌شناسان ایرانی واجب است و این امر میسر نمی‌شود مگر به مدد هموار کردن راه‌های پیشرفت آن.

به نظر نگارنده بایسته نیست زبان فارسی را از یافته‌ها و دستاوردهای نوین زبان‌شناسی رایانه‌ای صرفاً به بهانه مشکلات خط و ... بی‌نصیب سازیم، بلکه بهتر آن است که تا جای ممکن راهگشای مشکلات موجود در این مسیر باشیم و اگر به آن هدفی که منظور نظرمان بوده هم نرسیم، این را مایه دلگرمی خود سازیم که دست کم تجربیات ما، دیگران را یک گام در این مسیر فراتر خواهد برد.

همانطور که بر سر راه این پایان نامه نیز موانع و محدودیت‌های زیادی وجود داشت که اگر این باور نبود، شاید بارها نگارنده پر وسواس اما کم حوصله و پرمشغله در رویارویی با این محدودیت‌ها از ادامه راه بازمانده بود.

استثنائات و بی‌قاعدگی‌های فراوان که در مقوله اسم وجود دارد تفکر در این باب را بسیار زمانبر و البته گاه جذاب می‌نماید. از جمله این مسائل می‌توان به سر هم نویسی اسم‌ها و مشکلات هم نویسه‌ها اشاره کرد که کار تفکیک و شناسایی را بسیار دشوار و در پاره‌ای موارد دچار ابهامات پایه‌ای می‌کند.

عدم وجود یک برنامه رایانه‌ای کارآمد و دقیق برای برچسب دهی متن‌ها به زبان فارسی خلاء دیگری است که ممکن است در برخی از قسمت‌های تحقیق به طور کامل احساس شود و این امر نگارنده را بر آن می‌دارد که تا حد ممکن انگاره را بدون اتکاء به برچسب‌دهی دستوری پیش برد.

همچنین در این پژوهش سعی بر این است که «دستورالعمل‌ها» در نمودارهای گردش کار طوری تنظیم شوند که اولاً نیاز به تعامل و محاوره با کاربر به حداقل ممکن برسد و ثانیاً غالب مشخصات مربوط به صورت‌ها و حالات مختلف اسم‌ها در آن‌ها لحاظ گردد.

ساختار پژوهش حاضر به جز این فصل که پیشگفتار و طرح کلیات پژوهش است شامل ۳ فصل دیگر نیز به شرح زیر می‌باشد:

فصل دوم به نقش رایانه در زبان‌شناسی کاربردی و پیشینه تحقیق اختصاص می‌یابد. در فصل سوم چارچوب نظری کار و ساختار کلی برنامه توضیح داده می‌شود و بالاخره در فصل چهارم به نتایج و یافته‌های رساله پرداخته و به پیامدهای نظری و کاربردی هر یک اشاره می‌شود. ضمن اینکه پیشنهادهای نیز برای پژوهش‌های آتی ارائه می‌گردد.

۳-۱- بیان مسئله، اهداف و ضرورت انجام تحقیق

بی‌گمان افرادی که با تهیه و تدوین فرهنگ‌ها، فهرست‌های راهنما، واژه‌نامه‌های بسامدی از جمله فهرست‌های واژه‌نما، ترجمه ماشینی، پیکره‌های زبانی، پایگاه داده‌های زبانی و... سرو کار دارند، از اهمیت و کارایی فرایند بن‌واژه‌سازی در این حوزه‌ها بخوبی آگاهند. گرچه مزایای آن بر افرادی هم که به عنوان کاربر این حوزه‌ها از بسته نرم‌افزارهایی استفاده می‌کنند که در آنها برنامه بن‌واژه‌سازی نیز گنجانده شد، است هم پوشیده نیست.

یکی از کارایی‌های عمده این فرایند آنست که به هنگام تهیه فهرست راهنمای واژه‌ها، فهرست واژه‌نما و یا فرهنگ از یک متن، در صورتی که آن متن قبلاً بن‌واژه‌سازی شده باشد می‌توان اطمینان حاصل نمود که تمام صورت‌های یک واژه خاص در آن متن صرفاً با جستجوی بن‌واژه‌اش قابل بازیابی است. فرایند بن‌واژه‌سازی در متن‌های الکترونیکی که هر کدام به تنهایی یک فرهنگ، فهرست بسامدی یا فهرست راهنمای بالقوه به شمار می‌روند، صحت روندهای جستجوی ساده و پیشرفته را تضمین می‌کند. همچنین چنانچه این فرایند با فرایند تقطیع و برچسب دهی متن همراه شود می‌توان به نتایج دقیق‌تر و مطلوب‌تری در تحلیل‌های متنی دست یازید.

بنابراین «بن‌واژه‌سازی» نقش مؤثری در بازیابی اطلاعات (سبک شناختی، واژگان و...) از متن ایفا می‌کند. به فرض می‌توان پی برد که فلان نویسنده کدام واژه‌ها را در آثار خود یا یک اثر بخصوص بکار برده است. سؤال این است که با این همه کاربردهایی که فرایند بن‌واژه‌سازی در حوزه‌های مختلف زبان‌ها می‌تواند داشته باشد و با توجه به حجم کارهایی که در مورد زبان‌های انگلیسی، فرانسوی، آلمانی، و جزآن در این زمینه انجام گرفته است، آیا برای زبان فارسی هم در این خصوص اهتمام بایسته صورت گرفته است؟ تا آنجایی که نگارنده اطلاع دارد دو پروژه در حلد رساله کارشناسی ارشد (ماندانا فرخ ۱۳۸۱)، پویان‌دانش کارآراسته ۱۳۸۱) به فرایند بن‌واژه‌سازی فعل پرداخته‌اند.

اما در مورد اسم که با هیچ یک از مقولات زبانی دیگر قابل قیاس نیست (چه از نظر تعداد و چه از نظر تنوع ساختی) ظاهراً تا کنون کسی از این باب وارد مسئله نشده است. با نگاهی گذرا به فهرست واژه‌های اسمی در ضمیمه رساله (بالغ بر ۳۳۸۰۷ سرواژه) می‌توان به حجم حوصله و کار لازم به انجام در باب این مقوله پی برد. گذشته از آن سایر اجزای کلام بخصوص صفات را که تغییر مقوله داده و یا جانشین اسم‌ها می‌شوند نیز به این حجم و تعداد باید اضافه نمود.

براین اساس نگارنده پس از مرور و بررسی انبوه نظریاتی که در خصوص اسم و گروه اسمی فارسی یا مطالبی بسته گریخته در خصوص فرایند بن‌واژه‌سازی در لابلای برخی مقالات آمده، قصد دارد با استفاده

از معیارهای زبان شناختی، ضمن طبقه‌بندی و تعریفی تازه از اسم‌ها، انگاره‌ای نظری برای برنامه نرم‌افزاری ارائه دهد که به توان با آن ضمن شناسایی اسم‌ها، مشخصات آن‌ها (از جمله جمع و مفرد بودن، پیشوندی، پسوندی، مرکب، مشتق، نکره و معرفه و...) را نیز به کاربر داد.

پس ابتدا ارائه تعریفی جامع و البته ساختاری از اسم با توجه به ماهیت تحقیق که انجام پروژه‌ای برای کار با رایانه‌هاست، ضروری می‌نماید و این تعریف نه تنها باید در برگیرنده مشخصات صوری و تغییرات ساختارهای اسم باشد، بلکه جایگاه نحوی آن را در جمله نسبت به سایر عناصر و مقولات دستوری زبان مشخص کند. لذا رساله حاضر عمدتاً با مبنا قرار دادن نظر کلباسی (۱۳۷۱) و طباطبائی (۱۳۸۲) درخصوص ساخت اشتقاقی واژه‌ها و ترکیب و دیدگاه باطنی (۱۳۴۸) و مشکوٰالدینی (۱۳۷۳) در مورد جایگاه اسم و وابسته‌های پیشین و پسین آن به این حوزه وارد خواهد شد و با استفاده از همین معیارهای زبان شناختی (صرفی و نحوی) تلاش خواهد کرد تا موارد ابهام میان اسم و مقولات مختلف زبانی و یا حتی ابهامات موجود بین خود اسم‌ها که تقریباً در تمام زبانها به خصوص هنگام بن واژه‌سازی با رایانه، فرد با آن مواجه می‌شود، تا حدودی مرتفع سازد.

بنابراین هدف افزودن جزء مکمل دیگری به مجموعه کارهایی است که در راستای تحلیل و پردازش خودکار متن فارسی صورت گرفته است، و یا پایه و مبنایی برای کارهای آتی در این زمینه است. با نگاهی به محدودیت کارهایی که از نظر گستره و تعداد بخصوص درمقایسه با زبانهایی چون انگلیسی، آلمانی و فرانسوی، برای زبان فارسی انجام گرفته است و توجه به گسترش روزافزون پردازش اطلاعات با رایانه، سیستم‌های ترجمه ماشینی و سیستم‌های هوشمند، کاربردهای هوش مصنوعی و برنامه‌های واژه‌پرداز رایانه‌ای، بازیابی اطلاعات از متون به کمک رایانه و هزاران مورد مشابه، اهمیت این قبیل کارهای تحقیقاتی تا حد قابل توجهی آشکار می‌گردد.

بی‌گمان نیاز به توضیح نیست که سرعت بخشیدن به تحقیقات زبانی موجب تسریع روند تحقیقات در حوزه سایر علوم می‌گردد چرا که هر پژوهشگری در هر شاخه از علوم و در مراحل از تحقیق خود به زبان نیاز پیدا خواهد کرد.

۴-۱- تعریف کلید واژه‌ها

در این بخش از رساله به تعریف مفاهیمی پرداخته می‌شود که به عنوان واژه‌های کلیدی از اهمیت ویژه‌ای برخوردار هستند و درک صحیح رساله مبتنی بر شناخت کافی و وافی از آنهاست:

۱-۴-۱- اسم^{۱۰}

تعاریف مختلفی از اسم ارائه شده است. در پاره‌ای از این تعاریف صرفاً از جنبه معنایی به اسم پرداخته شده است: «اسم کلمه‌ای است که برای نامیدن شخصی یا حیوانی یا چیزی و بیان حالتی بکار می‌رود» (عماد افشار ۱۳۷۲، ص ۷).

دسته‌ای نیز مشخصه‌های ساختاری و یا نقش آنرا در جمله و نسبت به فعل مورد توجه قرار داده‌اند: «از لحاظ تجزیه روستختی اسم واژه‌ای است که به عنوان هسته در جایگاه اصلی گروه اسمی به کار می‌رود. اسم با نشانه‌های صرفی خاصی همراه می‌شود. نشانه‌های صرفی اسم به مفاهیم دستوری جنس و نوع، معرفه و نکره و نیز مفرد و جمع اشاره می‌کند.» (مشکوئه‌الدینی، ۱۳۷۳، ص ۱۶۷). یا «اسم گروهی از کلمات است که می‌تواند نقش‌های نهاد، مفعول، متمم، مسند، بدل، مضاف‌الیه و ... بپذیرد» (جعفر بگلو ۱۳۷۳، ص ۴۷).

و بالاخره در دسته‌ای دیگر از تعاریف نیز تمامی یا ترکیبی از مؤلفه‌ها (معنایی، ساختاری، نقشی و ...) لحاظ شده است.

برخی از تعاریف اسم در بخش دوم فصل بعدی بطور کامل ارائه شده است. پر واضح است که با توجه به ماهیت موضوع رساله تعاریفی از اسم که به جنبه‌های ساختاری آن پرداخته‌اند در اینجا حائز اهمیت می‌باشد.

۱-۴-۲- بن واژه^{۱۱}

واژه Lemma و جمع آن Lemmata، اصالتاً واژه‌ای لاتینی بوده و ریشه یونانی دارد که در فارسی به بن واژه و گاه سر واژه ترجمه شده است.

همان‌گونه که در فصل قبلی اشاره شد، بنا به تعریف «بن واژه، آن واژه پایه‌آست که در فرهنگ‌ها به عنوان مدخل یا سرواژه آورده می‌شود.» (جان میلز ۱۹۹۶، ص ۲۵۳).

عده کثیری این واژه پایه را معادل «ستاک» می‌گیرند که از تمام صورتهای صرفی یک واژه بدست می‌آید. (جان میلز ۱۹۹۶، سوزان هاکی ۱۹۹۸) و عده‌ای نیز آن را معادل «ریشه» فرض می‌کنند که تمام صورتهای صرفی و اشتقاقی یک واژه را در برمی‌گیرد (دیواین ۱۹۸۷).

توضیح بیشتر در خصوص سه اصطلاح پایه، ریشه و ستاک در فصل ۳ ارائه شده است.

-
1. Noun
 2. Lemma
 3. Base

۱-۴-۳- بن واژه‌سازی^{۱۳}

بن واژه‌سازی فرایندی است که با کمک آن می‌توان تمام صورت‌های صرف شده را زیر صورت پایه آنها برد یا به عبارت دیگر روشی است برای یافتن صورت پایه از تمام صورت‌های صرفی که به هم مرتبطند (جان میلز ۱۹۹۶، ص ۲۶۳).

در این جا نیز هم چنان که پیشتر اشاره شد، برخی بر این باورند که در طی فرایند بن واژه‌سازی هم صورت‌های صرفی و هم صورت‌های اشتقاقی زیر بن واژه (= ریشه) ای که از آن مشتق یا صرف شده‌اند برده می‌شوند.

این فرایند در گذشته بصورت دستی انجام می‌گرفته است بدین صورت که بن واژه‌ها به عنوان مدخل برای فرهنگ‌ها روی برگیزه نوشته شده و سپس به ترتیب الفبایی مرتب می‌گردیدند. امروزه این عمل به مدد رایانه‌ها سرعت و دقت بسیار بیشتری یافته است (ن.ک: بخش ۱-۱).

۱-۴-۴- پردازش خودکار

پردازش، هر گونه تغییر کیفی، کمی یا صوری است که در ساختار و روابط داده‌ها (اطلاعات خام که با روش‌های گوناگون به درون رایانه وارد می‌شود) و آرایش، ترتیب، حجم و اندازه و یا شکل آنها بوجود آید (عاصی ۱۳۷۳^(۳)، ص ۱۴).

۱-۴-۵- فارسی معیار

منظور از فارسی معیار، نثر فارسی سالم و نوشتاری امروز یعنی زبان رسمی فارسی می‌باشد که در واقع همان زبان معمولی در تهیه گزارش‌های خبری و سیاسی و علمی صدا و سیما و زبان کتاب‌های درسی و کتاب‌ها و مقالات علمی و سیاسی و سخنرانی‌ها است (ارژنگ ۱۳۷۴، ص ۳).

۱-۵- سؤالات اصلی تحقیق

- ۱- آیا می‌توان با استفاده از معیارهای زبان شناختی (صرفی و نحوی) راهکارهای نظری برای طراحی نرم‌افزارهایی ارائه نمود که اسم را در جملات متون فارسی معیار تشخیص دهد و لایه‌های صرفی و اشتقاقی آن (از قبیل علامات جمع، پیشوندها و...) را جدا سازد و به بن واژه آن اسم برسد؟
- ۲- آیا می‌توان با استفاده از معیارهای زبان شناختی، مشکل شناسایی هم نویسه‌هایی را که به مقولات مختلف زبانی و یا حتی یک مقوله واحد تعلق دارند، برطرف نمود؟

- ۳- آیا اسم‌هایی هستند که تعیین بن واژه آنها بدون برچسب دهی کاری دشوار یا غیرممکن باشد؟
- ۴- آیا معیاری برای تشخیص وندهای هم نویسه (به خصوص اگر یکی تصریفی و دیگری اشتقاقی باشد و یا در ساخت واژه‌هایی با مقولات دستوری مختلف شرکت کنند) وجود دارد؟
- ۵- آیا می‌توان معیاری برای تشخیص اسم‌های مرکب از گروه‌های اسمی در نظر گرفت؟

۱-۶- فرضیات پژوهش

- ۱- با استفاده از معیارهای صرفی و نحوی می‌توان اسم‌ها را در متون فارسی معیار شناسایی و لایه‌ها (وندها) ی آن را جدا کرد و به بن واژه آن دست یافت.
- ۲- فرایند بن واژه‌سازی اسم را در زبان فارسی می‌توان تا حدی بدون نیاز به برچسب دهی هم پیش برد.
- ۳- ارائه انگاره‌ای نظری برای شناسایی و بن واژه‌سازی اسم فارسی به کمک رایانه امکان پذیر است.

۱-۷- شیوه تحقیق

همانگونه که در بخش ۲-۲ بیان مسأله گفته شد. در این رساله سعی بر این است تا با مبنا قرار دادن دیدگاه‌های کلباسی (۱۳۷۱) ، طباطبائی (۱۳۸۲) ، باطنی (۱۳۴۸) ، مشکوٰالدینی (۱۳۷۳) و کاتمبا (۱۹۹۳) در خصوص اسم، شیوه‌های اسم‌سازی و گروه‌های اسمی، ابتدا تمامی مشخصات صوری اسم فارسی در تمام حالات بررسی و تعیین گردد و سپس انگاره‌ای نظری برای نرم‌افزار برنامه تشخیص اسم و بن واژه سازی آن در زبان فارسی ارائه شود.

در راستای نیل به این هدف از شیوه‌های مختلفی بهره جسته شده است به طوری که از نظر شیوه تحقیق این رساله را می‌توان به چهار بخش عمده تقسیم نمود:

۱-۷-۱- بخش اول:

این بخش شامل گردآوری مطالب و اطلاعات زبان شناختی در خصوص اسم فارسی و بررسی پایان‌نامه‌های دانشجویی با زمینه‌های پژوهشی مشابه است که به شیوه کتابخانه‌ای صورت گرفته است.

۱-۷-۲- بخش دوم

در این بخش به کاربرد رایانه‌ها در زبان شناسی کاربردی و درواقع حوزه‌هایی پرداخته شده است که فرایند بن واژه‌سازی می‌تواند در آنها نقش مؤثر داشته باشد این قسمت نیز به شیوه کتابخانه‌ای و جستجوی اینترنتی انجام گرفته است.

۱-۷-۳- بخش سوم

تهیه فهرست اسم‌ها به کمک برنامه Excel که از فرهنگ فشرده سخن استخراج گردیده‌اند، شیوه دیگری است که در این رساله بکار گرفته شده است. در این فهرست بیش از ۳۷ هزار سرواژه اسمی آورده شده است که با توجه به قابلیت‌های نرم افزار Excel، امکان جستجوی متنوع در آن میسر می‌باشد.

۱-۷-۴- بخش چهارم

و بالاخره در این بخش پس از انتخاب واژه‌هایی از فهرست مذکور، رفتار آنها در جایگاه‌های مختلف نحوی در پیکره‌های متنی پایگاه داده‌های زبان فارسی پژوهشگاه علوم انسانی و مطالعات فرهنگی بررسی و بر اساس نتایج به دست آمده مسیرهای نمودار گردش کار برنامه و انگاره نظری ارائه گردیده است.

۱-۸- ساختار رساله

رساله حاضر مشتمل بر ۴ فصل و ۴ پیوست می‌باشد.

فصل اول شامل دو بخش پیشگفتار و کلیات پژوهش است که در آن موضوع تحقیق تا حدی تبیین می‌شود و سپس به بیان مسأله، اهمیت و ضرورت تحقیق ارائه تعریفی از کلید واژه‌ها، روش پژوهش و سؤالات اصلی تحقیق، پرداخته می‌شود.

فصل دوم نیز در دو بخش ارائه می‌شود: بخش اول که به نقش رایانه‌ها در زبان‌شناسی کاربردی اختصاص یافته است به ویژه حوزه‌هایی از آن که فرایند بن واژه‌سازی در آنها کاربرد وسیعی یافته است.

بخش دوم شامل پیشینه تحقیق است که در سه قسمت (۱) پژوهش‌های غیرایرانیان درباره اسم / گروه اسمی فارسی (۲) پژوهش‌های ایرانیان درباره اسم / گروه اسمی فارسی و (۳) پایان نامه‌ها به آن پرداخته می‌شود.

در فصل سوم که گستره و در واقع پیکره اصلی پژوهش است ابتدا چارچوب نظری کار و سپس انگاره نظری و نمودار گردش کار برنامه ارائه می‌شود.

فصل چهارم که به ارائه نتایج و یافته‌های پژوهش و همچنین پیشنهاداتی برای طرح‌های آتی اختصاص یافته است.

فهرست منابع تحقیق در دو بخش منابع فارسی و منابع انگلیسی ذکر شده است.

پیوست‌ها شامل:

(۱) نمودارهای گردش کار و

(۲) فهرست اسم‌ها

می باشد که هر دو با نرم افزار Excel تهیه شده است.

۳) جداول شامل:

۱-۳) معروف ترین جمع های مکسر،

۲-۳) پسوند های جمع،

۳-۳) پیشوندها،

۴-۳) پسوندها،

۵-۳) پسوند واره ها،

۶-۳) الگوهای اسم مرکب و

۷-۳) پیوندها) که در آخر رساله گنجانده شده اند.

۴) واژه نامه ای از واژه ها و اصطلاحات به کار رفته در رساله نیز بدان ضمیمه شده است.

ضمناً ۱۱ جدول رفع ابهام در لابلای فصل سوم قابل مشاهده است.