گزارش عملکرد

گروه پژوهشی

زبانشناسی رایانه ای

۱۳۹۱

فهرست مندرجات

مقدمه :

نخستین مطلبی که در بحث هر علم پیش می آید ، حوزه های مختلف آن می باشد . حوزه های متداول و مختلف علم زبانشناسی را نیز می توان به طور اجمالی نام برد .

- حوزه مقدماتی : شامل آوا شناسی ، واج شناسی ، نحو ، واژه شناسی ، معنا شناسی و سایر موارد .
- حوزه دستور زبان : که شامل توصیف ساختارهای زبان به شیوه سنتی ، ساختگرایی ، نقش گرایی ، گشتاری و نظایر آن .
- تجزیه و تحلیل گفتمان
- جامعه شناسی زبان : شامل گونه های طبقاتی ، لهجه های جغرافیایی ، زبانهای میانجی ، توانش تکلمی ، دو زبانگی ، دو گویشی ، دو لهجه ای و مانند آن .
- روانشناسی زبان : در برگیرنده مسایل فراگیری زبان اول ، آموزش زبان های دوم و خارجی ، آموزش زبان برای اهداف ویژه ، تحلیل تقابلی ، تحلیل خطاهای گویش و نظریه های یادگیری و از این قبیل .
- تدریس مهارتهای زبان : نظیر شنیدن ، صحبت کردن ، خواندن و نوشتن و تکنیک های مختلف آموزش .
- شیوه ها و روشهای تدریس زبان خارجی : در انواع مختلف درستور ، ترجمه مستقیم ، گفتاری ، شنیداری ، خواندن ، رمز شناختی ، مفهومی ، نقشی ، تکلمی ، القایی شامل می شوند .
- آموزش زبان و آمار پایه .
- زبانشناسی رایانه ای که اخیراً تمامی حوزه های دیگر را نیز شامل شده و از دستاوردهای غرب است و این حوزه یک حوزه بکر و دست نخورده ای است و حیطه ای است میان رشته ای که از تعامل رشته های مختلف چون رایانه ، علوم اطلاع رسانی و زبانشناسی ایجاد شده است لذا روز به روز به موارد استفاده و زیر مجموعه آن افزوده می گردد .

زبانشناسی رایانه ای یا زبانشناسی محاسباتی (Coputational Lingustics) حوزه ای میان رشته ای است که سعی دارد با بهره گیری از روش های آماری و مبتنی بر قاعده (rul – based) ، از منظر محاسباتی و مدل سازی زبان طبیعی بپردازد . به طور کلی ، زبانشناسی محاسباتی از همکاری دانشمندان و کارشناسان رشته های زبان شناسی ، علوم رایانه ای ، هوش مصنوعی ، ریاضی ، منطق ، علوم شناختی ، روانشناسی شناختی ، روان – زبانشناسی ، مردم شناسی ، عصب شناسی و برخی دیگر از رشته ها استفاده می کند .

رایانه اگرچه از دستاوردهای غرب است اما د دو دهه اخیر آنچنان در جوامع مختلف از جمله جامعه ما رخنه کرده که امروز در بعضی از مراکز حتی تصور اینکه روزی رایانه در آن مرکز نباشد ، بسیار مشکل است . با وجود این هنوز تنها تعداد محدودی از زبانها با رایانه تطبیق داده شده و در آن قابل استفاده می باشند و زبان فارسی از این جهت در ابتدای راه می باشد .

زبانشناسی رایانه ای به قسمتی از زبانشناسی گفته می شود که مسائل مربوط به زبان و رایانه را به طور عمودی مورد بحث و بررسی قرار می دهد و نشان می دهد که زبانشناسی چگونه می تواند به حل مسائل موجود در مورد تطبیق خط و زبان در رایانه کمک نماید .

اعضاء این گروه سعی دارند که با اجرای طرحها و نشر کتابها و مقالات و سخنرانیها و ... بتواند در برطرف کردن مشکلات موجود در خصوص هم آهنگی خط و زبان فارسی با رایانه کمک کرده و در ایجاد یک نگرش علمی نسبت به آن موثر واقع شوند .

در زیر به تعدادی از موارد کاربرد علم زبانشناسی رایانه ای اشاره می شود :

۱-۱ ساخت ریشه ساز زبان فارسی : ریشه سازی یکی از ابزارهایی است که در بازیابی اطلاعات در برخورد با مسئله عدم انطباق واژگان مورد استفاده قرار می گیرد . (یعنی عدم تطبیق واژه های پرس و جو با واژه های مدرک) به فرآیند حذف هر پسوند از واژه ها و تبدیل این واژه ها به ریشه های آنها ریشه سازی گفته می شود . برای مثال ریشه سازی واژه "ورزش" را به شکل "ورز" (ریشه مضارع فعل ورزیدن) به وجود می آورد . ریشه سازها عموماً متناسب با هر زبان خاص تهیه می شوند . طراحی ریشه سازها مستلزم خبرگی زبانشناسی در زبان و درک نیازهای مرتبط با بازیابی اطلاعات می باشد .

۱-۲ فرا یافت Concept و تجزیه زبانهای برنامه نویسی : نوآم چامسکی توانست تشابه زبانهای طبیعی و زبانهای برنامه نویسی را به اثبات برساند . یعنی یک زبان رایانه ای مانند زبانهای طبیعی دارای یک

دستور زبان و یک فرهنگ می باشد . تفسیر یک متن از تجزیه واژها Lexicon آغاز سپس با تجزیه نحو Syntax و در آخر با تجزیه مفهوم Semantic آن پایان می یابد .

۳-۱ ترجمه ماشینی : این شاخه اززبان شناسی رایانه ای زمان درازی کم اهمیت جلوه می کرد . اما امروزه یکی از موارد مورد علاقه پژوهشگران این رشته می باشد . پس از مرحله تجزیه مفهوم و سپس پرگماتیک را نیز افزود . در واقع این دو سعی در شناخت مفهوم خاص یک واژه در مکانی که ظاهر می شود را دارد .

۴-۱ پرسش و پاسخ با زبانهای طبیعی : این ایده مدتی به عنوان پاسخی قانع کننده به مسئله ارتباط انسان و ماشین تلقی می شد . این دید در واقع جنبه وسیعتری از دستور زایشی چامسکی است .

۵-۱ صرف محاسباتی : به مطالعات مربوطه به ساختارهای درونی کلمات صرف گفته می شود . اغلب دست آوردها و نتایج تحقیقات در صرف محاسباتی در سایه تلاشهای علمی انسان به منظور ایجاد و ساخت سیستم های پردازش زبانهای طبیعی انسانی فراهم آمده است .

۲- **ساخت ماشین های ترجمه** : ترجمه ماشینی یکی از نخستین استفاده هایی است که از پردازش زبان طبیعی به عمل آمده است . ساخت این گونه ماشینها استفاده از تئوریهای نحوی و معنایی موجود در زبانشناسی ضروری است . در این زمینه زبانشناسان می توانند در حیطه های : ایجاد سیستم های ترجمه تماماً خودکار ، ترجمه ماشینی نیازمند به انسان ، ترجمه با کمک ماشین ، ساخت ویرایش گرهای املایی ، ایجاد سیستم های ترجمه روی خط ، ایجاد اصطلاحنامه های روی خط و سرانجام ماشینهای همگرا و واگرا مفید باشند .

نکته حائز اهمیت آنکه برای ساخت ماشین های ترجمه از نظریه اطلاعات و رمزنگاری نیز استفاده می شود که این ارتباط حیطه های اطلاع رسانی با زبانشناسی رایانه ای را نشان می دهد . بدیهی است مراکز اطلاع رسانی برخوردار از این مزیت قادر خواهند بود اطلاعات مورد نیاز کاربران را به بیش از یک زبان ارائه نمایند .

۲-۲- **ترجمه ماشینی (Machine Translation – MT)** زیر شاخه ای از زبان شناسی محاسباتی می باشد که عبارت است از ترجمه متنی از یک زبان طبیعی به زبانی دیگر ، توسط کامپیوتر . در سطح مقدماتی، ترجمه ماشینی یک جایگزینی ساده برای کلمات از زبان طبیعی به زبان دیگری است . با استفاده از تکنیک های زبان شناسی پیکره ای، ترجمه های پیچیده بیشتری قابل دستیابی هستند . همچنین این تکنیک ها کنترل بهتر تفاوتهای گونه شناسی در زبان، تشخیص عبارات و ترجمه اصطلاحات را به خوبی و درستی جدا کردن عبارات نامتعارف در متن، مقدور می سازند .

نرم افزارهای ترجمه ماشینی کنونی اغلب به کاربر اجازه تغییر دلخواه بر اساس حوزه کاری یا حرفه ای دلخواه را می دهد (مانند : گزارش آب و هوا) در واقع ارتقاء کیفیت خورجی با استفاده از محدود کردن کلمات جایگزین شونده، انجام می شود . این تکنیک بطور خاص در حوزه رسمی با زبانهای فرموله شده استفاده می شود. این بدین معنی است که ترجمه ماشینی از اسناد قانونی و دولتی آسانتر از تولید خروجی قابل استفاده از مکالمات یا متون غیر چهارچوب بندی شده دیگر است . همچینین کیفیت خروجی بهبود یافته می تواند با استفاده از دخالت انسان بدست آید . برای مثال سیستم هایی موجودند که اگر کاربر بطور کاملاً واضحی کلماتی که اسامی خاص هستند را معین کرده باشد، قادر به ترجمه دقیقتری هستند. با کمک گرفتن از این تکنیک ها ترجمه ماشینی بعنوان یک ابزار برای کمک کردن به مترجمان (انسانها) و بسیاری از موضوعهای محدود، قادر به تولید خروجی قابل استفاده و نهایی است.

در ترجمه ماشینی ویژگی هایی وجود دارد که نه تنها از نظر جاذبه و کشش علمی، بلکه از دیدگاه اقتصادی و دیگر ضرورت ها و اقتضاهای عصر، انجام آن را کاملاً توجیه می کند . به عنوان مثال، در مقر سازمان ناتو در بروکسل و جامعه اروپا علیرغم آنکه حدود ۱۲۰۰ مترجم ورزیده به کار اشتغال دارند، در حال حاضر از ترجمه ماشینی نیز استفاده می شود. دلیل این امر سرعت و هزینه است. میزان کاری که مترجمی ورزیده در خلال چندین روز انجام می دهد، توسط کامپیوتر در عرض چند دقیقه انجام می شود. حتی اگر کیفیت و دقت ترجمه ماشینی کمتر از حاصل کار مترجم باشد، باز هم از جهات گوناگون اهمیت و ارزش خاص آن چشمگیر است.

۳– **بازیابی اطلاعات** : اساس کار مراکز و پایگاههای اطلاع رسانی بر جستجو و بازیابی اطلاعات بنا شده است . زیر بخشی از زبانشناسی رایانه ای که به معناشناسی رایان های موسوم است تاکنون توانسته کیفیت این گونه سیستم ها را به نحو مطلوبی ارتقاء دهد . از معناشناسی منطقی نیز استفاده های فراوانی به عمل می آید .

به اختصار می توان گفت که اساس کار سیستم های بازیابی اطلاعات اسناد موجود نیست، بلکه بازنمودهای معنایی است و ایجاد این بازنمودها کاری است در حیطه معناشناسی رایانه ای . در ایجاد زمنیه زبانشناسان در ایجاد بازیابی غیر نظام مند، بازیابی نظام مند، ماشین های پرسش و پاسخ و ابهام زدایی از مفهوم واژه در متن نقش ایفا می نماید . دو مورد از مشکلات عمده موجود در سیستم های بازیابی اطلاعات به ابهام و تشابه واژگانی مربوط است که رفع این مشکلات نیز در حیطه زبانشناسی رایانه ای است.

منطق توصیفی، شبکه های معنایی و صورت گرایی قالب- محور مبنای کار سیستم های اطلاع رسانی در حیطه معناست.

۴– **بازشناسی گفتار و سیستم های گفتار به متن** : بازشناسی گفتار بر اساس مدلهای زبانشناختی بنا شده که از آن جمله می توان به مدل مارکو اشاره نمود . دستور حالتهای محدود چامسکی نیز کاربرد فراوان دارد . در سیستم های بازشناسی گفتار شناخت امواج صوتی، چگونگی تعبیر و تفسیر آنها، رمزگذاری و رمز گشایی و چگونگی تغییر مشخصه های آوایی مسائلی عمده و اساسی می باشند که یک آواشناس یا واج شناس از عهده آن بر می آید.(آواشناسی و واج شناسی یکی از زیر شاخه های علم زبانشاسی است). تاکنون در زمینه تلفظ و املا الگوهای مختلفی ارائه گردیده است.

۵- **برچسب زنی کلمات** : این حیطه نیز از حیطه های مطرح در علم اطلاع رسانی است و به یکی دیگر از زیر شاخه های علم زبانشناسی که نحو نام دارد مربوط است . شناخت مقوله های مختلف دستوری و چگونگی ایجاد ارتباط و تمایز بین دسته های مختلفی از کلمات به اطلاعات تخصصی در حیطه نحو نیازمند است .

۶- **سیستم های متن به گفتار** : حیطه رایج دیگری در علم اطلاع رسانی است که از واج شناسی و آواشناسی رایانه ای استفاده های زیادی به عمل می آورد . شناخت اندام های صوتی، آوانویسی آواهای زبانی، استفاده از الگوها و قواعد واجی و واج شناسی در سیستم های متن به گفتار همگی نیازمند به فردی است که حیطه واج شناسی به مطالعه و تحقیق پرداخته باشد .

در ادامه برای رعایت اختصار کاربردهای عمده دیگر این حیطه فهرست وار ارائه می گردد .

۷- **واژه سازی** : جستجو و پیشنهاد معادلهای مناسب برای واژه های فاقد معادل فارسی (البته برای این کار لازم است با هماهنگی و یا همکاری فرهنگستان زبان و ادب فارسی اقدام گردد) .

۸- **نمودارهای** N : از این نمودارها برای شمارش اعدا در متن و ... استفاده می شود و مطالب موجود تا حد زیادی به ریاضیات و علم رایانه نیز مربوط است .

۹- **دستورهای بافت وابسته و مستقل از بافت** : سیستم های بازیابی با استفاده از این گونه الگوها می توانند به جستجو یا تجزیه متن دست بزنند که از آن جمله می توان به تجزیه های از کل به جز و از جز به کل اشاره نمود .

۱۰- **تجزیه و تحلیل معنایی**

۱۱- **گفتمان** : که در آن خصوص چگونگی تعبیر و تفسیر ضمایر و یافتن مرجع آنها و نکات مرتبط دیگر بحث می شود که در ماشینهای ترجمه و ... کاربرد فراوان دارد .

۱۲- **تبدیل متن به گفتار**

۱۳- **تبدیل گفتار به متن**

۱۴- **خلاصه سازی خودکار**

۱۵- **موتورهای کاوش هوشمند**

اهداف مهم گروه پژوهشی زبانشناسی رایانه ای به شرح زیر می باشد :

- مطالعه ساختار و واژگان های علمی و پژوهشی جهت دریافت توصیفگرهای مناسب
- مطالعه و بررسی علوم اطلاع رسانی و کتابداری و زبانشناسی و افزایش بهره وری آنها در سطح کشور و منطقه
- مکانیزه کردن نمایه سازی و چکیده نویسی مدارک علمی کشور و منطقه
- مطالعه در ساختار زبان فارسی و ارتباط منطقی واژه ها و پراکنش آماری آنها
- مطالعه در ساختار ترجمه در زبانهای مختلف با استفاده از رایانه و منطق زبانشناسی
- افزایش بهره وری اطلاع رسانی با استفاده از علم زبانشناسی درسطح کشور منطقه
- تعیین استراتژیهای جستجو با استفاده از اصلاحنامه ها و تزاروسها و واژه نامه های کامپیوتری
- بررسی چگونگی تثبیت اصطلاحات فارسی و کاربرد صحیح آن زبان بعنوان یک زبان علمی
- بررسی و رفع مشکلات جستجو و بازیابی اطلاعات با الگوهای منطقی و علم زبانشناسی
- بررسی املای هوشمند از طریق ارتباط علم زبانشناسی و رایانه
- بررسی ارتباط زبانشناسی و بازیابی اطلاعات با الگوهای منطقی زبانشاسی
- مطالعه ساختار و صرف و نحو زبانهای طبیعی با استفاده از علم منطق زبانشناسی
- بررسی چگونگی ساختار (میانجی) واسط های زبانهای طبیعی برای سیستم های کامپیوتری
- بررسی استراتژی های تبدیل گفتار به نوشتار و بالعکس با استفاده از علم و منطق زبانشناسی
- مطالعه خلاصه سازی خودکار در رایانه با استفاده از علم و منطق زبانشناسی

- بررسی چگونگی رو در رویی (ارتباط) زبان طبیعی و سیستم های کامپیوتری
- برچسب زنی کلمات جهت بازشناسی رایانه ای، با استفاده از منطق زبانشناسی
- بررسی ساختار دستورهای بافت وابسته و مستقل از بافت جهت تعیین الگوهایی برای بازیابی اطلاعات

طرحهای تحقیقاتی

پایان یافته در سال ۱۳۹۱

واژه نامه برگردان نام و نام خانوادگی نویسندگان خارجی (نوشته شده با حروف انگلیسی) به فارسی با استفاده از تحلیل رخداد

مجری : دکتر محمدرضا فلاحتی فومنی قدیمی

تاریخ شروع : ۸۹/۸/۱۲

تاریخ پایان : ۹۰/۱۰/۱۴

چکیده :

هدف اصلی از انجام پژوهش حاضر بررسی امکان استفاده از روش رخداد محور برای تعیین گونه غالب املایی اسامی نویسندگان خارجی (نوشته شده با حروف انگلیسی) بود . در مجموع ۹۶۸ نام و نام خانوادگی از کتاب " فهرست مستند اسامی مشاهیر و مولفان ، ویراست دوم) " استخراج گردید . با استفاده از توانش زبانی محقق ، تحلیل های واژه سازی و جستجوی وبی گونه های مختلف املایی اسامی پیش بینی و ثبت گردید . سپس فراوانی هر نام (تعداد سندهای حاوی آن نام) در موتور جستجوی گوگل تعیین و ثبت شد . اطلاعات پس از گردآوری به ترتیب فراوانی ذیل هر مدخل ثبت گردید و هر مدخل چهار نوع اطلاعات شامل نام ، نام خانوادگی ، نام و نام خانوادگی ، سرانجام نام خانوادگی و نام تهیه و در قالب یک یک واژه نامه ارائه گردید . یافته ها حاکی از آن بود که چنانچه جستجوی وبی را مدنظر قرار دهیم صورت های املایی منتخب در کتاب فهرست مستند اسامی تنها در ۴۹٪ از موارد با صورت های املایی منتخب در روش رخداد محور یکسان است . در موارد فراوان صورت های املایی منتخب در فهرست مربوطه به کاربر در امر بازیابی اطلاعات کمک بایسته ارائه نمی دهد و این در حالی است که در فهرست مربوطه خطاهای املایی فراوانی چه در نگارش اسامی فارسی و چه در نگارش اسامی انگلیسی وجود داشت که به خاطر عدم ارتباط مستقیم با اهداف تحقیق از ورود به آن بحث خودداری گردید .

استخراج و تحلیل خودکار حروف اضافه و ربط بسیط در متون فارسی

(سامانه تحلیل حروف ربط، نشانه و اضافه ی زبان فارسی)

مجری : دکتر محمدرضا فلاحتی فومنی قدیمی

تاریخ شروع : ۱۳۹۰/۱۰/۱۴

تاریخ پایان : ۱۳۹۱/۵/۱۴

چکیده :

امروزه انجام پردازش ها و تحلیل های زبانی با استفاده از نرم افزار شیوع بسیاری پیدا کرده است. هدف پژوهش حاضر آن بود تا نرم افزاری تهیه گردد تا بتوان به کمک آن، واژگان زبان فارسی متعلق به شش گروه حروف اضافه ی ساده، حروف اضافه مرکب، شبه حروف اضافه، حروف نشانه، حروف ربط و حروف ربط مرکب را شناسایی نمود و برای هر واژه اطلاعاتی آماری شامل فراوانی مطلق، فراوانی نسبی، درصد فراوانی نسبی، فراوانی تجمعی و درصد فراوانی تجمعی را در قالب جداول و نمودارهای خطی و ستونی ارایه نمود. برای انجام کار، فهرست واژگان مربوط به گروه های شش گانه فوق از کتاب های دستور زبان فارسی موجود در بازار استخراج گردید. در مواردی که یک واژه در کتاب های مختلف به صورت های گوناگونی تحلیل و طبقه بندی شده بود، از نظر اکثریت استفاده شده برای رفع مشکل تنوع املایی، برای هر واژه، تمامی صورت های املایی احتمالی، پیش بینی و فهرست گردید. از معیار فاصله ی کامل قبل و بعد از واژه برای تفکیک واژه ها (به جز کسره ی اضافه) استفاده به عمل آمد و در شناسایی زنجیره ی حروف از قاعده ی بزرگترین طول استفاده شد. نرم افزار طراحی شده با درونداد متون نمونه مورد آزمایش قرار گرفت و خطاهای مشاهده شده مرتفع گردید. این نرم افزار قادر است ضمن تحلیل اطلاعات، نتایج را در قالب فایل اکسل ارایه نماید.

این کار، امکان استفاده از جداول و نمودارهای حاصله را به آسان ترین شکل در ساختار مقالات، کتاب ها و آثار دیگر محققان فراهم می آورد و در جهت تسریع نگارش مقالات و آثار علمی از سوی محققان حوزه ی زبان، ابزاری بسیار مناسب محسوب می گردد. نتایج تحلیل متون در نرم افزار حاضر از سوی دو متخصص زبان مورد بررسی و بازبینی قرار گرفت و مطابقت نتایج با فهرست های واژگان طراحی شده مورد تأئید قرار گرفت.

بررسی معنا شناختی بازیابی مدارک در پایگاه اطلاع رسانی مرکز منطقه ای اطلاع رسانی علوم و فناوری و ارائه الگویی معنایی برای بهبود بازیابی

مجری : دکتر فاطمه احمدی نسب

تاریخ شروع : ۹۰/۱۰/۵

تاریخ پایان : ۹۱/۱۰/۵

چکیده :

در راستای انجام طرح حاضر با عنوان "بررسی معنا شناختی بازیابی مدارک در پایگاه اطلاع رسانی مرکز منطقه ای اطلاع رسانی علوم و فناوری و ارائه الگویی معنایی برای بهبود بازیابی " ، پژوهشگر با ورود کلید واژه های مختلف به صورتهای مختلف تصریفی وعبارتی و با عملگرهای مختلف ۱۱۷ کلیدواژه را در موتور جستجوی مرکز منطقه ای اطلاع رسانی علوم و فناوری در سه فیلد جستجو در همه فیلدها، جستجو در عنوان و جستجو در کلیدواژه جستجو نمود تا نحوه عملکرد موتور جستجو را از نظر معنایی بررسی نماید . البته با توجه به ماهیت طرح وی قصد داشت تا جستجو را در چکیده نیز انجام دهد که فعلاً این امکان در پایگاه مرکز وجود ندارد . در ضمن جستوها به مقالات فارسی محدود شد . وی با بررسی مدارک بازیابی شده سعی نمود تا الگوی معنایی بازیابی مقالات فارسی را در موتور جستجوی مرکز بدست آورد . از آنجا که عملکرد این موتور جستجو کلیدواژه ای است و نه معنایی ، بنابر این امکان بازیابی مدارک بر اساس روابط مفهومی هم معنایی ، شمول معنایی ، تقابل معنایی و دیگر روابط معنایی در این موتور جستجو وجود ندارد . بنابر این در این پایگاه بازیابی مدارک درست مطابق با حضور کلیدواژه در مدارک و بر اساس میزان ربط آنها بازیابی می شود . اما در عمل با پدیده ریزش کاذب زیادی از مدارک مرتبط مواجهه هستیم که در واقع تعدادی از آنها به عملکرد موتور جستجوی مرکز و همچنین عملکرد نمایه سازان مرکز مرتبط بوده و علل دیگر را باید در ماهیت خط و زبان فارسی جستجو نمود . در این پژوهش پس از بررسی مسائل مختلف ، سعی شده است تا راهکارهایی حتی المقدور متناسب با امکانات موجود مرکز در راستای رفع این مشکلات و افزایش کارایی این موتور جستجو ارائه شود . بطور خلاصه ، مشکلات مرتبط به موتور جستجوی مرکز و نمایه سازی را می توان در عدم عملکرد عملگر واژهبری ، ایراد شکلی علامت عملگر not و عدم تمایز بین آ و ا در موتور جستجوی مرکز ، نگارش

نویسه ی به جای نویسه همزه ، تعدد ایرادات نگارشی به هنگام نمایه سازی دانست . مشکلات موثر دیگر در امر بازیابی را باید در ماهیت خط و زبان فارسی در وندهای جمع ساز ، وحدت ، نکره، التقای مصوتها، صورتهای بی قاعده، پیوسته نویسی و یا ناپیوسته نویسی و همچنین عدم یکدستی در نگارش و در هم معنایی فراوان در سطح واژگان علمی فارسی جستجو نمود . در راستای رفع این ایرادات پیشنهاد می شود که در نظام بازیابی مرکز تمایز بین آ و ا به رسمیت شناخته شود، برای عملگر not علامت دیگری انتخاب شود، در راهنمای تفصیلی جستجو در موتور جستجوی مرکز توضیحات زبانشناختی بیشتری قرار داده شود، در نمایه سازی ملاحظات زبانشناختی بیشتری رعایت شود و تا حد امکان مطابق با دستور خط و فرنگستان عمل شود . در ضمن یکی از موثرترین شیوه ها در افزایش دقت بازیابی تهیه اصطلاحنامه بر اساس داده های موجود در پایگاه مرکز است که تاحدود زیادی مسائل معناشناختی هم معنایی، شمول معنایی را حل نموده و حتی در ابهام زدایی نقش بسزایی دارد. البته این راه حل ها بر اساس امکانات موجود در مرکز منطقه ای ارائه شده است و در سطح کلان باید با ایجاد ریشه یاب، پیکره زبانی و برچسب زن سعی نمود تا موتور جستجو را از کلیدواژه ای به معنایی تغییر داد. با این حال و در حال حاضر با نگاهی واقع گرایانه باید گفت که، راه حل های پیش گفته دست یافتنی تر و امکان پذیرتر هستند .

بررسی تطبیقی اصطلاحنامه و هستان شناسی : شباهت ها، تفاوت ها، کاربرد و تدوین

مجری : دکتر فاطمه احمدی نسب

تاریخ شروع : ۹۰/۱۰/۵

تاریخ پایان:۹۱/۱۱/۵

چکیده :

هدف اصلی از انجام پژوهش حاضر مقایسه دو ابزار معنایی بازیابی اطلاعات اصطلاحنامه و هستان شناسی است . در این طرح به شباهت ها و تفاوت های این دو ابزار توجه شده است . به نظر می رسد که نباید اصطلاحنامه را ابزاری ناکارآمد دانست بلکه این ابزار به عنوان یک ابزار مفید می تواند در بازیابی اطلاعات مورد استفاده قرار گیرد. هستان شناسی ابزاری است که تناسب بیشتری با وب۳ و هوشمند دارد و شاید بتوان ادعا نمود که هستان شناسی اصطلاحنامه ای است که خود را با وب معنایی مطابقت داده است . در طرح حاضر سعی شده است تا با زبانی ساده وجوه افتراق، اشتراک و کارایی آنها در نظام های بازیابی ارائه شود . امید است که این مختصر بتواند برای پژوهشگران این حوزه مفید واقع شود .

چالش های پردازش زبان طبیعی فارسی

مجری : دکتر حمید علیزاده

تاریخ شروع : ۸۸/۱۱/۱۰

تاریخ پایان: ۸۹/۱۲/۱۰

چکیده :

پردازش زبان طبیعی، فنون و شیوه های پردازش و تحلیل متن را شامل می شود به نحوی که رایانه قادر به درک محتوا و معنای متون نشده و همچون انسان، الگو و ساختارهای معنایی را برای ارتقاء تحلیل و بازیابی اطلاعات استخراج نماید. این پژوهش های انجام نشده در پردازش زبان طبیعی فارسی، نقشه راه پژوهش های آتی در این حوزه را مهیا می کند. حوزه هایی چون تحلیل ساخت واژه، تحلیل انواع کلام، ترجمه ماشینی و وب معنایی و هستان شناسی از جمله زیر شاخه های پردازش زبان طبیعی است که پژوهش های انجام شده در زبان فارسی برای هر کدام مطرح و تلاش هایی که باید در آینده برای تکمیل این پژوهش ها انجام شود تشریح شده است.

امکان سنجی اجرای بازیابی اطلاعات بین زبانی با نظام ترجمه ماشینی گوگل

مجری : دکتر حمید علیزاده

تاریخ شروع : ۹۰/۲/۸

تاریخ پایان: ۹۱/۴/۸

چکیده :

بازیابی اطلاعات بین زبانی (CLIR) از حوزه های داغ پژوهشی در زمینه بازیابی اطلاعات است. در این پژوهش، استفاده از نظام ترجمه ماشینی گوگل، برای ترجمه عبارت های جستجو جهت استفاده در فرآیند بازیابی اطلاعات بین زبانی فارسی – انگلیسی بررسی شد. یافته ها نشان داد که استفاده از مترجم ماشینی گوگل برای ترجمه عبارت های جستجو باعث کارآمدی بازیابی اطلاعات بین زبانی می گردد.

طرحهای در دست اجرا در سال ۱۳۹۱

کاربردی کردن مفردسازی ماشینی اسامی جمع در زبان فارسی

مجری : شاپور رضا برنجیان

تاریخ شروع : ۳۰/ ۷/ ۹۱

تاریخ پایان : ۹۲/۱۱/۳۰

چکیده :

اهمیت بررسی و تحقیقات در حیطه زبان فارسی به ویژه در برخورد این زبان و خط آن با علم رایانه در ماشینهای ترجمه، ریشه یاب و ... بر کسی پوشیده نیست اگر چه تاکنون پژوهشهایی نیز در این زمینه صورت پذیرفته، اما هنوز موارد بسیاری یافت می شود که هیچ گونه بررسی درباره آنها انجام نپذیرفته و تحقیقات در این زمینه ضروری به نظر می رسد. مفرد سازی صیغه های جمع اسامی نیز یکی از این موارد می باشد .

مسئله اسم فارسی و ویژگی های آن در برابر موتورهای جستجو در شبکه های کامپیوتری نیز سالهاست که مورد توجه اطلاع رسانها قرار گرفته و این امر کم و بیش در کتابخانه ها و مراکز اطلاع رسانی مورد بررسی قرار گرفته است . طرح حاضر تلاشی است در جهت هموارتر کردن این راه ناهموار.

گسترش جستجوی اسامی نویسندگان در پایگاه مقالات مرکز منطقه ای بر اساس فهرست اشکال مختلف نگارشی اسامی فارسی

مجری : شاپور رضا برنجیان

تاریخ شروع : ۹۱/۷/۳۰

تاریخ پایان : ۹۲/۷/۳۰

چکیده :

در این طرح نمای کلی از نرم افزار گسترش جستجوی اسامی نویسندگان در پایگاه مقالات مرکز منطقه ای بر اساس فهرست اشکال مختلف نگارشی اسامی فارسی در مرکز منطقه ای اطلاع رسانی علوم و فناوری تهیه و ارائه گردیده و در آن عملکرد نرم افزار به گونه ای است که با ارائه یک اسم می توان اسناد نوشته شده به اشکال مختلف آن اسم (خواه با spaice یا با half spaice یا بدون spaice نوشته شده باشد) آن اسم را مشاهده نمود و چنانچه اسم مذکور دارای نوشتار دیگریست، جستجوگر می تواند با تغییر نوشتار و جستجوی مجدد به هدف خود برسد و با این کار از حذف اطلاعات و مدارک مورد نیاز جلوگیری به عمل آورد .

گسترش جستجوی اسامی نویسندگان در پایگاه مقالات مرکز منطقه ای بر اساس فهرست اشکال مختلف نگارشی اسامی فارسی

مجری : شاپور رضا برنجیان

تاریخ شروع : ۹۱ / ۷/ ۳۰

تاریخ پایان : ۹۲/۷/۳۰

چکیده :

در این طرح نمای کلی از نرم افزار گسترش جستجوی اسامی نویسندگان در پایگاه مقالات مرکز منطقه ای بر اساس فهرست اشکال مختلف نگارشی اسامی فارسی در مرکز منطقه ای اطلاع رسانی علوم و فناوری تهیه و ارائه گردیده و در آن عملکرد نرم افزار به گونه ای است که با ارائه یک اسم می توان اسناد نوشته شده به اشکال مختلف آن اسم (خواه با spaice یا با half spaice یا بدون spaice نوشته شده باشد) آن اسم را مشاهده نمود و چنانچه اسم مذکور دارای نوشتار دیگریست، جستجوگر می تواند با تغییر نوشتار و جستجوی مجدد به هدف خود برسد و با این کار از حذف اطلاعات و مدارک مورد نیاز جلوگیری به عمل آورد .

بررسی وضعیت چکیده نویسی انگلیسی در مجلات فارسی علمی– پژوهشی وزارت علوم، تحقیقات و فناوری (حوزه ی علوم پایه) سال ۱۳۹۰ و ارایه راهکارها یی برای بهبود آن

مجری : دکتر محمدرضا فلاحتی فومنی قدیمی

تاریخ شروع : ۹۱/۶/۵

تاریخ پایان : ۹۲/۸/۵

چکیده :

در چند سال اخیر و پیروی برنامه ی رتبه بندی مجلات فارسی در کشور، وجود چکیده ی انگلیسی در مجلات علمی پژوهشی ضروری است. نگاهی گذرا به مجلات رتبه دار فارسی نشان می دهد که هر مجله برای نگارش چکیده دستورالعملی را ارایه نموده که انتظار می رود از سوی نویسندگان رعایت شود. بررسی نمونه مقالات این دسته از مجلات نشان می دهد که چکیده های انگلیسی بکار رفته در آنها دارای اشکالات فراوانی است. هدف پژوهش حاضر آن است تا چکیده را در مجلات علمی-پژوهشی مربوط به حوزه ی علوم پایه (یکی از حوزه های شش گانه ی موضوعی) مورد بررسی و تجزیه تحلیل زبانشناختی و اطلاع رسانی قرار دهد. علت انتخاب حوزه ی علوم پایه تناسب بیشتر این حوزه با ماموریتهای مرکز منطقه ای اطلاع رسانی علوم و فناوری بوده است. اشکالات چکیده ها شناسایی و گروه بندی خواهد شد و اطلاعات توصیفی و تحلیل در هر مورد ارایه خواهد گردید. در پایان دستورالعمل و راهکارهایی برای بهبود وضع چکیده نویسی به انگلیسی ارایه خواهد گردید.

بررسی وضعیت واژگان تخصصی و فرایندهای واژه گزینی در حوزه علم اطلاعات و دانش شناسی در زبان فارسی

مجری : دکتر فاطمه احمدی نسب

تاریخ شروع : ۹۱/۱۰/۵

تاریخ پایان: ۹۲/۸/۵

چکیده:

برخورد زبان فارسی با ورود جدیدترین یافته های علمی به جامعه علمی ایران با چالش عظیمی به نام واژه گزینی همراه‌است. یکی از رسالتهای پژوهشگران در هر حوزه ای معادل یابی درست برای مفاهیم علمی است چرا که انتخابهای نابجا می‌تواند به زبان فارسی آسیب بزند و علاوه بر این با ایجاد هم معنایی و چند معنایی زیاد باعث عدم شفافیت و در نتیجه عدم تبادل و تفاهم سریع علمی بین پژوهشگران آن حوزه از دانش شود. هدف از انجام این طرح ارائه تصویری روشن از واژگان تخصصی حوزه علم اطلاعات و دانش‌شناسی است تا با آگاه ساختن متخصصان این حوزه از نتایج پژوهش، آنها را در واژه‌گزینی بهتر و رفع کاستی های این حوزه یاری رساند. بدیهی است که یکی از راه های بومی ساختن یک حوزه از دانش، گسترش واژه سازی در آن حوزه بر اساس قواعد زبان فارسی و جلوگیری از هرج و مرج و تک‌روی در واژه‌سازی برای مفاهیم علمی است. این طرح در راستای این هدف ارائه شده است و به بررسی و تحلیل ۳۰۰۰ واژه تخصصی حوزه علم اطلاعات و دانش شناسی می پردازد تا شیوه های واژه سازی در این حوزه را به دست دهد.

بررسی میزان مطابقت مجلات برتر ISC با دستور خط فارسی فرهنگستان زبان و ادب فارسی

مجری : دکتر فاطمه احمدی نسب

تاریخ شروع : ۹۱/۱۰/۵

تاریخ پایان: ۹۲/۱۱/۵

چکیده:

زبان و خط فارسی یکی از زبان‌های مهم و رایج دنیا است و منابع و مدارک زیادی به این خط و زبان نوشته می‌شسود و در شبکه جهان‌گستر قرار می‌گیرد. علاوه بر این پایگاه‌های متعدد مقالات فارسی نیز موجود است. اما به علت ماهیت خط فارسی، بازیابی اطلاعات - چه در پایگاه‌های اطلاع‌رسانی و چه وب جهان‌گستر - با ریزش کاذب و عدم بازیافت منابع مرتبط مواجهه می‌شود. یکی از راه‌های بهبود بازیابی منابع فارسی، رعایت یکدستی و یکسان‌سازی متون فارسی از لحاظ خط است. در راستای این مهم، پیروی از یک دستورالعمل واحد راهگشاست. از آنجا که دستورخط فارسی مصوب فرهنگستان در جهت یکسان‌سازی خط فارسی مدون شده است و پیروی از آن برای نهادهای دولتی الزامی است، مسلما تبعیت از آن، در جهت بهبود بازیابی اطلاعات مؤثر است. . در این طرح میزان و وضعیت پیروی از این دستورالعمل در مجلات برتر ISC بررسی می‌شود. جامعه پژوهش را ۷ مجله برتر گروه‌های هفتگانه حوزه‌های عام مجلات فارسی نمایه شده در ISC تشکیل می دهد.

مقالات

دگر گونی های زبان و علل واژه گزینی

نویسنده : شاپور رضا برنجیان

چاپ شده در مجله ی اباختر ، سال ششم ، شماره پیاپی ۲۱ و ۲۳ |، تابستان ۱۳۹۱

چکیده :

در این مقاله ابتدا به واژه نامه دساتیر و مصوبه مجلس شورای الامی در خصوص ممنوعیت به کارگیری اسامی و عناوین و اصطلاحات بیگانه اشاره شده و همچنین تعدادی از واژه های ساخته شده توسط افراد غیر متخصص و غیر مسئول و معایب این گونه واژه گزینی ذکر گردیده و به تعدادی از اصول و ضوابط واژه گزینی اشاره شده است . در نهایت به دگرگونی های واژگانی ، معنایی و بسامدی – که از دگرگونی هایی هستند که بر اثر گذشت زمان در زبان پدید می آیند – به طور اختصار شده و نیاز جامعه به واژه گزینی مورد بحث و بررسی قرار گرفته است .

کلید واژه : دساتیر ، دگرگونی های زبان ، علل واژه گزینی ، اصول واژه گزینی ، دگرگونی معنایی ، دگرگونی واژگانی ، دگرگونی بسامدی

سنجش اثر بخشی سلسله کارگاه های پیشنهادی آموزشی سواد اطلاعاتی به شاغلین تحقیقاتی کتابخانه ها و مراکز اطلاع رسانی

نویسنده : سیامک، مرضیه و احمدی نسب، فاطمه (۱۳۹۱)

چاپ شده در فصلنامه کتابداری و اطلاع رسانی ۶۰ (۴).

چکیده:

متن حاضر بر آن است تا با توجه به نتایج طرح «ارتقای مهارتهای سواد اطلاعاتی شاغلان تحقیقاتی کتابخانه‌ها و مراکز اطلاع‌رسانی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» و استانداردهای سواد اطلاعاتی برای آموزش عالی، عنوانهای سلسله جلسات پیشنهادی مجری را برای آموزش سواد اطلاعاتی به شاغلان تحقیقاتی کتابخانه‌ها و مراکز اطلاع‌رسانی برای تشکیل سلسله کارگاه‌های آموزشی ارائه دهد و سپس به سنجش اثربخشی این عنوانها بپردازد. به طور نمونه، در سلسله کارگاه‌های آموزشی برگزار شده در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوی.

کلیدواژه‌ها: سواد اطلاعاتی، شاغلان تحقیقاتی کتابخانه‌ها و مراکز اطلاع‌رسانی، کارگاه آموزشی، سنجش اثربخشی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

Another look to the Construction of Present Stem in Persian Language

Shapour Reza Berenjian

Regional Information Center for Science and Technology (RICeST), Shiraz, Iran
shapour_berenjian@radiffmail.com

Abstract
In Persian language, the second person singular of imperative verb is used to make the present stem. We present a pattern for making present stem directly from infinitive verbs of Persian language instead of using imperative form. In this study, the Persian verbs are categorized in 12 groups regarding the methods used to make their present stems, and the making pattern of each group is presented here.
[Shapour Reza Berenjian, Ali Reza Berenjian. Another look to the Construction of Present Stem in Persian Language. J Am Sci 2012;8(6):79-83]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 9

Keywords: Present stem, Persian infinitive verbs, Persian language, Grammar

The Study of Thesaural Relationships from a Semantic Point of View

Dr. Ja'far Mehrad

Fatemeh Ahmadi Nasab

Abstract

Thesaurus is one, out of many, precious tools in information technology by which information specialists can optimize storage and retrieval of documents in scientific databases and on web. In recent years, there is a shift from thesaurus to ontology by downgrading thesaurus in favor of ontology. The reason is claimed to be that thesaurus cannot meet the needs of information management because it cannot create a knowledge-rich description of documents. The so-called reason is that the thesaural relationships are restricted and insufficient. In this paper the writers show that thesaural relationships are not inadequate and although it appears that they are restricted, not only they are not, but also they cover all semantic relations and they can increase the possibility of successful storage and retrieval of documents. This study shows that thesauri are semantically optimal and they cover all lexical relations, therefore thesauri can continue as suitable tools for knowledge management.

Keywords: thesaurus, thesaural relationships, lexical relations

آسیب شناسی زبان و خط فارسی در بازیابی اطلاعات : نگاهی به موتورهای کاوش و پایگاه های بر خط

نویسنده : دکتر فاطمه احمدی نسب ، دکتر شعله ارسطوپور

چاپ شده در مجموعه مقالات نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب

چکیده :

همانگونه که در بسیاری از متون نیز مطرح است زبان فارسی به واسطه ویژگی های خاص خود چه از نظر رسم الخط و چه از نظر صرف و معنا با چالش های منحصر به فردی در زمینه ذخیره و بازیابی اطلاعات رو به روست . در مقالات مختلف به این مشکلات در قالب ها و سطوح مختلف بدون توجه به نوع شناسی این چالش ها اشاره شده است . به عبارت بهتر کمتر مقاله ای در زبان فارسی در حوزه اطلاع رسانی از دیدی مبتنی بر علم زبانشناسی مدرن به بحث دسته بندی این مشکلات و پاسخ دهی به آنها پرداخته است . نوشتار حاضر بر آن است تا عمده ترین چالش های مطرح در این زمینه را در سه گروه رسم الخط ، مسائل صرفی و مسائل معنایی مورد اشاره قرار داده و پس از بیان نمونه هایی در پیوند با هر یک از این چالش ها ، و ارائه ساختواره ای درختی از انواع مسائل مطرح در این سه گروه ، با نگاهی به پژوهش های صورت گرفته در زمینه زبان فارسی و عربی به ارائه راهکارهایی جهت حل این مسائل بپردازد.

Inconsistent transliteration of Iranian university names: a hazard to Iran's ranking in ISI Web of Science

Mohammad Reza Falahati Qadimi Fumani,

Marzieh Goltaji,

Pardis Parto

Abstract

Today, university ranking has turned into a critical issue in the world. Each university is identified with a surface form under which the whole performance of that university is assessed. This article intends to provide a clear picture of the inconsistencies observed in recording Iranian university titles by their affiliated authors and to clarify the negative impact of such inconsistencies in positioning Iranian universities in global university ranking systems. To collect various surface forms of Iranian university names, use was made of ISI Web of Science through keywords Cu = Iran and py = 2000–2009. Only MSRT universities were considered. Two M.A. experts listed all variant forms of a single university under that name. The form publicized in a university's website was considered as its entry name. The major sources of variation identified were as follows: *Acronyms, misspellings, abbreviations, space variations, syntactic permutation, application of vowels/consonants and vowel/consonant combinations, /a/vs./aa/, Tashdid, Kasra ezafe, redundancy, downcasing, voiceless glottal stop sound /?/, shortening* and *deletion of titles.* It was found that at its present shape Iranian universities are not receiving the rank they really deserve simply because authors affiliated to a university use university title forms inconsistently. It was recommended that authors follow the surface form publicized by universities in their websites, use the help of an editor in their works, and not be credited for their articles in case the forms deviate from those publicized through the websites. A spell checker, as an add-ins software is highly needed to homogenize Iranian university surface forms by replacing the variants by the dominant form proposed

سخنرانی

پیکره زبان و انواع آن

سخنران : شاپور رضا برنجیان

زمان : ۱۳۹۱/۸/۳۰

مکان : سالن کنفرانس مرکز منطقه ای اطلاع رسانی علوم و فناوری

چکیده :

یک تعریف کلی برای پیکره زبان وجود دارد و آن اینکه : در دانش زبان ، پیکره ، مجموعه ای از متون نوشتاری یا گفتاری آوانویسی شده است که می توان آن را به عنوان مبنایی برای تحلیل و توصیف زبانی به کار برد و یا ، پیکره مجموعه ای از نمونه های زبان طبیعی است که به طریقه الکترونیکی ذخیره شده اند . علت ایجاد این شاخه در زبان شناسی رایانه ای ، نیاز اساسی پژوهشی های زبانشناسی به داده های واقعی زبانی به صورت گسترده بوده است . زبان شناسی پیکره ای نماینده یک حوزه مشخص از مطالعات زبانی نیست بلکه بنیادی روش شناختی ، برای پژوهش های زبانی است .

بررسی تنوع املایی نام و نام خانوادگی نویسندگان خارجی در فارسی

سخنران : دکتر محمد رضا فلاحتی قدیمی فومنی

زمان : دی ماه ۱۳۹۰

مکان : سالن کنفرانس مرکز منطقه ای اطلاع رسانی علوم و فناوری

چکیده:

با توجه به موفقیت نسبی روش رخداد-محور در تعیین صورت املایی غالب نام و نام خانوادگی نویسندگان خارجی (نوشته شده با حروف انگلیسی) به نظر می رسد می توان از این روش نیز به صورت انفرادی و یا ترکیب با روش های دیگر برای فارسی نویسی اسامی خارجی استفاده نمود.

استناد یا استفاده واقعی: نگاهی به دو رویکرد در خصوص ارزیابی و رتبه گذاری مجلات علمی

سخنران : دکتر محمد رضا فلاحتی قدیمی فومنی

زمان : ۱۳۹۰/۱۲/۲۵

مکان : سالن کنفرانس مرکز منطقه ای اطلاع رسانی علوم و فناوری

چکیده:

در این مقاله به مزایا و نقاط ضعف دو روش استناد و استفاده ی واقعی پرداخته شده است. نشان داده شد که هر روش به تنهایی دارای معایبی است و در صورت تلفیق این دو می توان به معیارهای دقیق تری برای رتبه گزاری منابع علمی پرداخت.

وضعیت واژه گزینی در حوزه زبانشناسی و چالش های تدوین اصطلاحنامه زبانشناسی

سخنران: دکتر فاطمه احمدی نسب

زمان : ۹۱/۸/۲۹

مکان : سالن کنفرانس مرکز منطقه ای اطلاع رسانی علوم و فناوری

چکیده :

در این سخنرانی ، سخنران پس از ارائه سخنانی در باب ضرورت واژه گزینی و اهمیت آن و نقش فرهنگستان زبان و ادب فارسی دراین حوزه، به نقاط قوت و ضعف فرهنگستان در این حوزه پرداخته و سپس با ارائه مثالهایی از حوزه زبانشناسی و نقد برابر نهادهای فرهنگستان ضرورت تهیه اصطلاحنامه در حوزه زبانشناسی را نشان داده است و به چالشهای روبروی اصطلاحنامه نویسان این حوزه از دانش پرداخته است .

خط فارسی در بازیابی اطلاعات : چالش ها، راهکارها و سیاست های فرهنگستان

سخنران : دکتر فاطمه احمدی نسب

زمان : ۹۱/۱۱/۴

مکان : کنفرانس محتوای ملی در فضای مجازی

چکیده :

خط فارسی، به علت ویژگی های خاص خود ذخیره و بازیابی اطلاعات را با چالش های عمده ای روبه رو کرده است که یکی از مهمترین آنها بحث پیوسته نویسی خط فارسی است ، زیرا که جست و جوها معمولاً به صورت گروه اسمی انجام می شود . در این نوشتار، مسائل مرتبط با بازیابی گروه های اسمی فارسی در زمینه پیوست نویسی شامل علامت نکره، تکواژ میانجی، وند جمع ساز، جمع بی قاعده و تکواژ مفید واژگانی مطرح و نمونه هایی برای آن ذکر شده و سپس راهکارهایی برای رفع این مسائل در حوزه نمایه سازی و همچنین اصلاحاتی برای خط فارسی مطرح شده است.