

به نام خداوند مهربان

عنوان:

بررسی ساختواژی هم نگاره های اسمی و صفتی به  
منظور کمک به برچسب دهی «اسم» به کلیدواژه ها در  
پیکره های علمی

نگارنده:

الهام علایی ابوذر

استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

# ضرورت و اهمیت پژوهش:

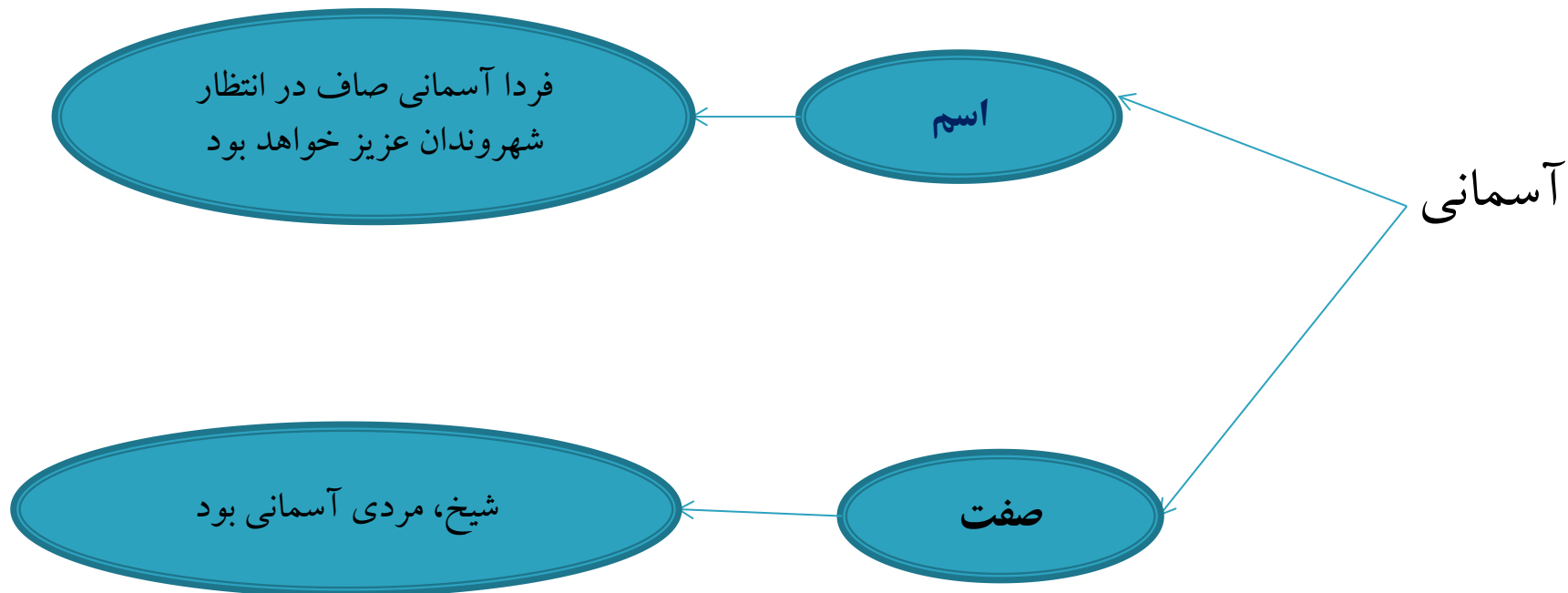
در تهیه سامانه های برچسب گذاری برای اسم ها در زبان فارسی، بررسی ساختواری اسم ها و صفت ها از دو نظر حائز اهمیت است:

# ۱- کلمات خارج از واژگان (Out Of Vocabulary words)

تلفنیدن: تلفن + ید + ن

اینترنت خوانی: اینترنت + خوان + ی

## ۲- هم نگاره ها (homographs)

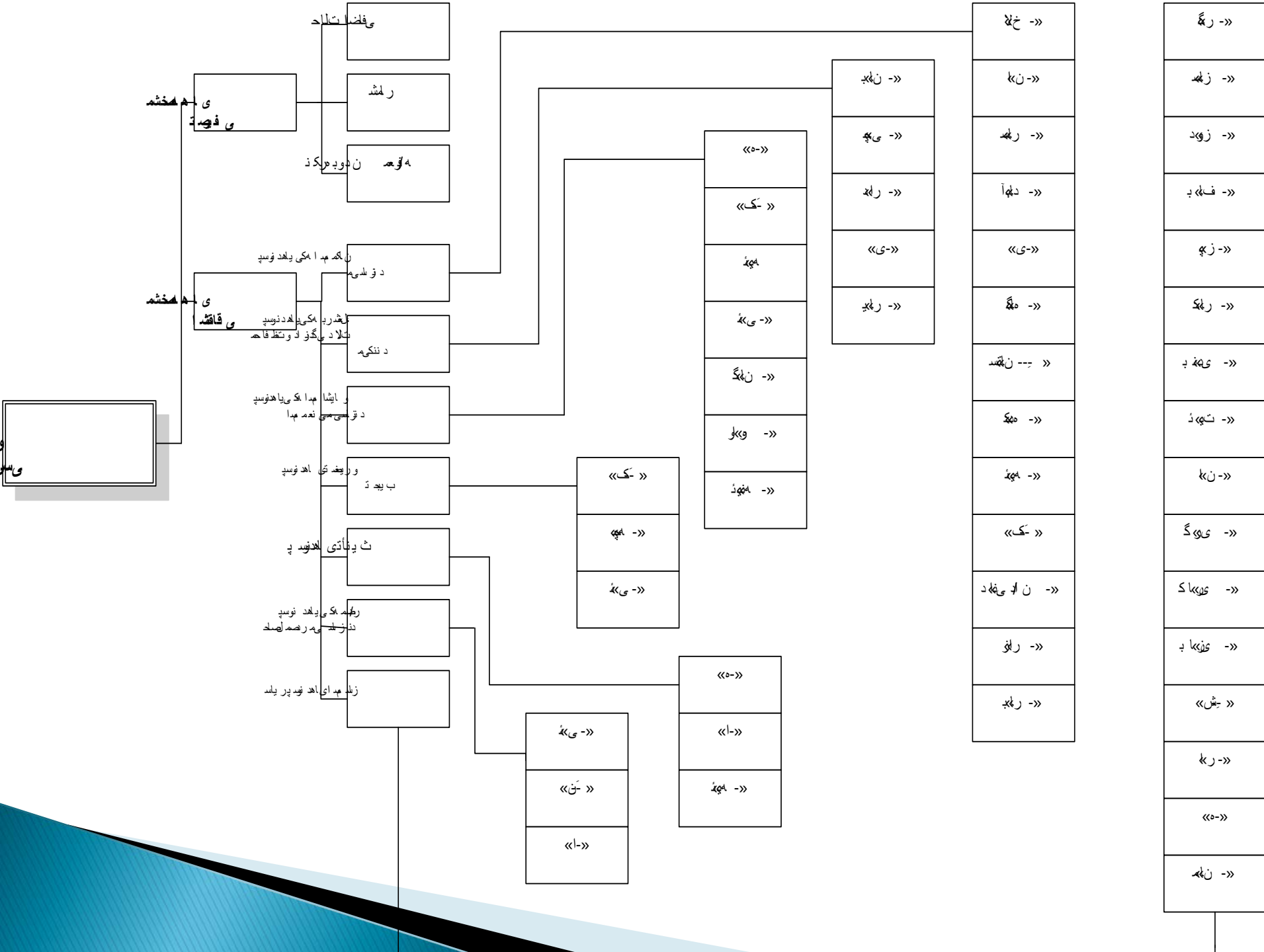


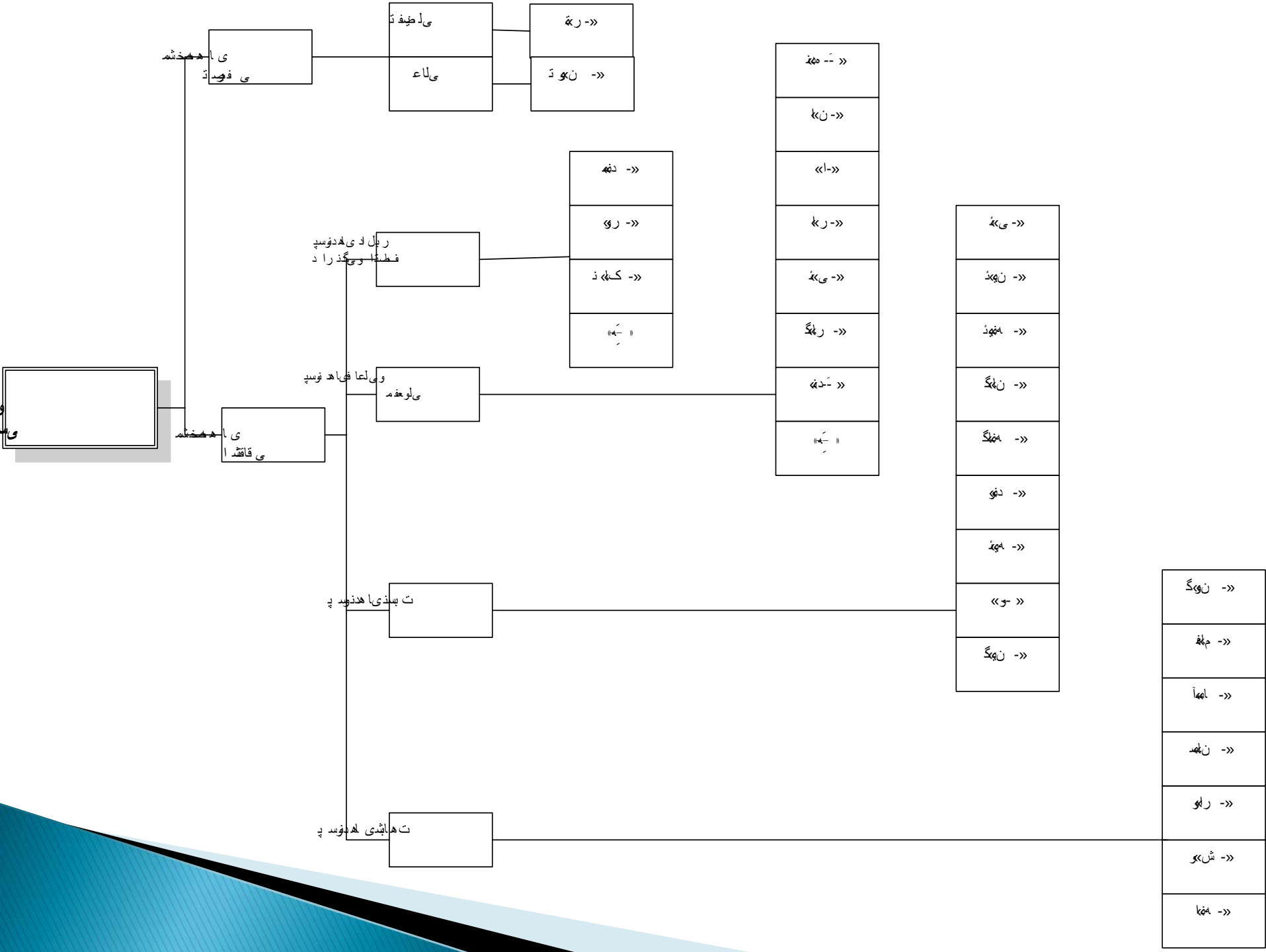
علی اشرف صادقی - شیوه ها و امکانات واژه سازی در زبان فارسی  
معاصر ۱ تا ۱۲ : ۱۳۷۲ - ۱۳۷۰ «

فریبا قطره (۱۳۸۶) - مشخصه های تصریفی در زبان فارسی امروز

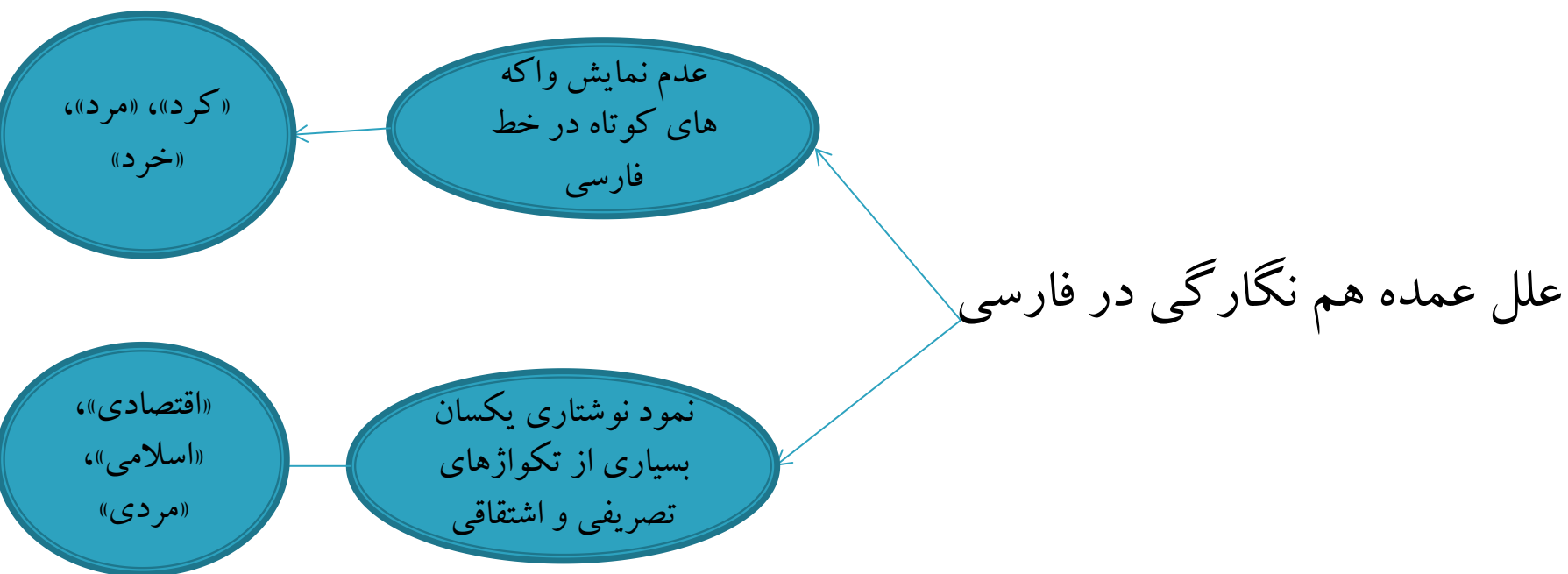
خسرو کشانی (۱۳۷۱) - اشتقاق پسوندی در زبان فارسی امروز

ژیلبر لازار (۱۳۸۹) - دستور زبان فارسی معاصر





# ساختار نوشتاری فارسی و هم نگاره ها:



# پیکره های مورد استفاده در پژوهش:

۱- پیکره متنی زبان فارسی (پیکره بی جن خان)

۲- پیکره وابستگی نحوی زبان فارسی

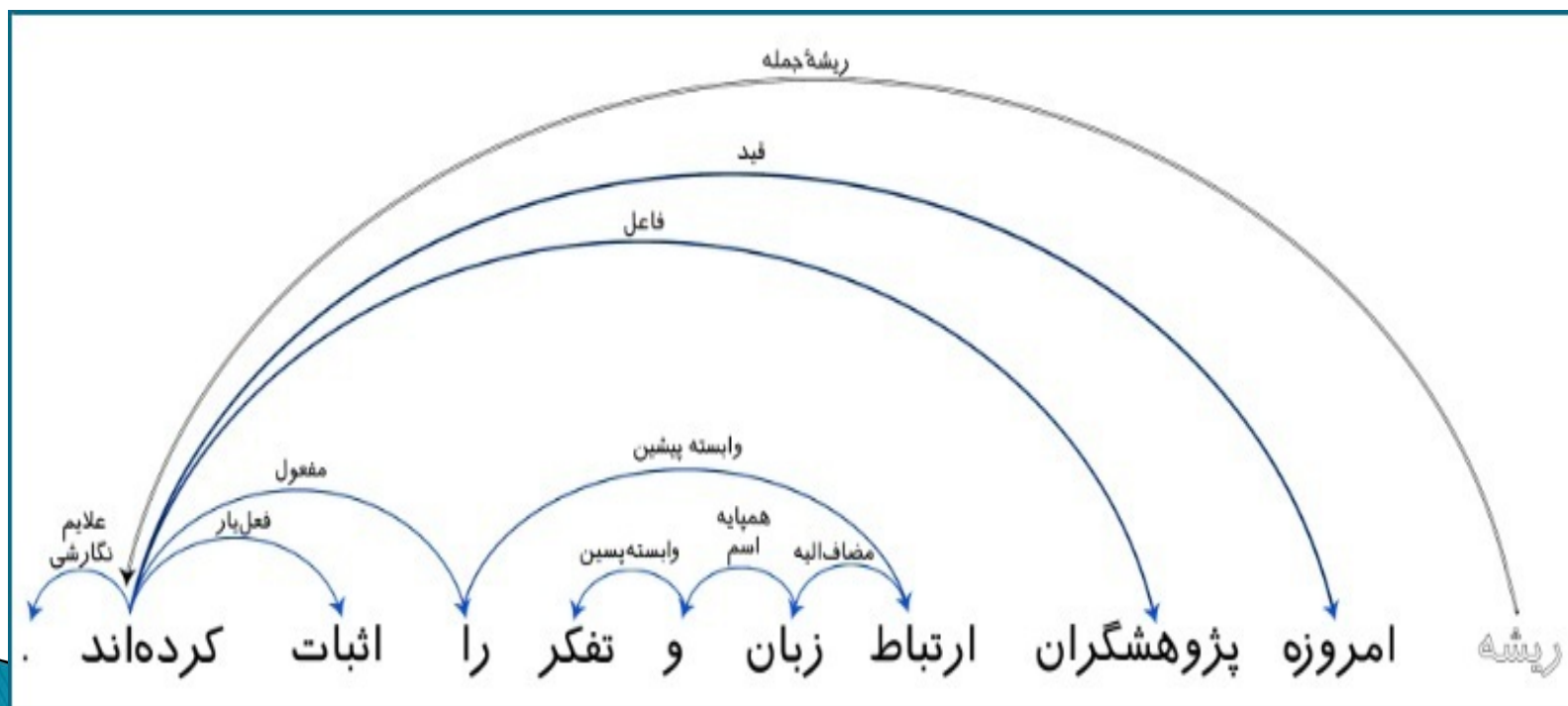
۳- پایگاه دادگان زبان فارسی

# ویژگی های پیکره متنی زبان فارسی / پیکره بی جن خان:

پیکره متنی استاندارد زبان فارسی که در پژوهشکده پردازش هوشمند علائم تهیه شده است، مجموعه ای از متون نوشتاری و گفتاری زبان فارسی به صورت رسمی است که از منابع واقعی همچون روزنامه ها، سایت ها و مستندات از قبل تایپ شده، جمع آوری شده، تصحیح کرده و برچسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده است و دارای تنوع بسیار زیادی است.

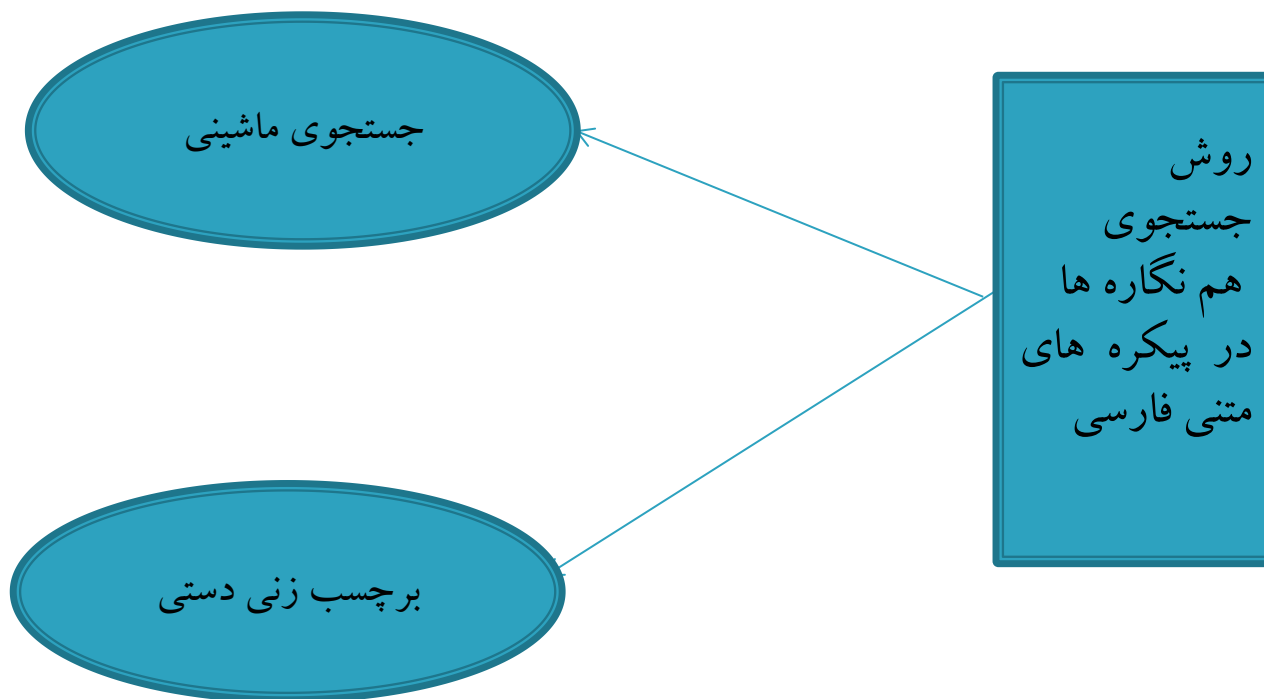
# ویژگی های پیکره وابستگی نحوی زبان فارسی:

مجموعه ای است شامل ۳۰۰۰۰ جملهٔ برچسب خورده با اطلاعات نحوی و ساخت واژگی.



# ویژگی های پایگاه دادگان زبان فارسی:

پایگاه دادگان زبان فارسی، مجموعه‌ای نرم‌افزاری برای ذخیره، پردازش و ارائه داده‌های زبانی فارسی است. این پایگاه دربرگیرنده پیکره‌های گوناگونی از زبان فارسی است که با وجود حجمی عظیم و با گستردگی و گوناگونی‌های بسیار، دارای ساختاری بسامان و منطقی است و امکان هرگونه جستجو و دستیابی سریع به آگاهی‌های مورد نیاز را در هر زمان فراهم آورده است. پایگاه دادگان زبان فارسی تنها مجموعه‌ای از مواد خام زبانی نیست بلکه دارای متن‌های نشانه‌گذاری شده از جمله شناسنامه متن، برچسب‌های دستوری، آوایی، ریشه‌ای و معنایی است که همواره افزایش می‌یابد. این دادگان مجهز به نرم‌افزارهای اختصاصی جستجو، تقطیع و تحلیل متن است.



نمونه ای از هم نگاره هایی که به روش جستجوی ماشینی از پیکره های  
متنی زبان فارسی و پیکره وابستگی نحوی زبان فارسی تهیه شده اند

N:1 ADJ:133 (فراوانی کلمه «اقتصادی» به عنوان «صفت» ۱۳۳ و به عنوان «اسم»، ۱ است) اقتصادی

N:4 ADJ:28 (فراوانی کلمه «انقلابی» به عنوان «صفت» ۲۸ و به عنوان «اسم»، ۴ است) انقلابی

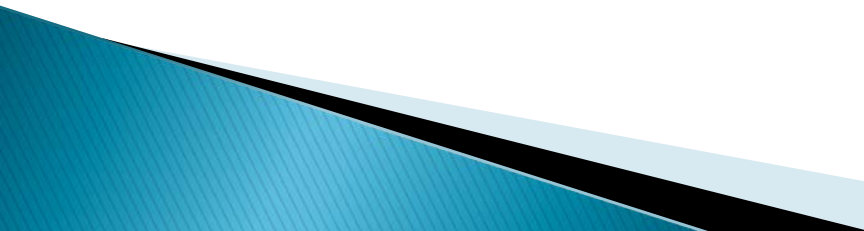
N:51 ADJ:94 (فراوانی کلمه «جوان» به عنوان «صفت» ۹۴ و به عنوان «اسم»، ۵۱ است) جوان

# نمونه ای از برچسب دهی دستی به هم نگاره های موجود در پایگاه دادگان زبان فارسی

(فروانی این کلمه در پایگاه دادگان ۲۶۱۷۵ است) N-IN , ADJ, N+V-PRE فرهنگی 26175

(فروانی این کلمه در پایگاه دادگان ۸۴ است) N-IN , N(NO-V), N+V-PRE کاسبی 84

(فروانی این کلمه در پایگاه دادگان ۱۷۶۷۰ است) N-IN , N , N+V-PRE, ADJ شهرداری 17670



# یافته های پژوهش:

بررسی کلی هم نگاره ها در پیکره های مورد مطالعه نشان می دهد که تعداد هم نگاره ها در پیکره ها قابل توجه است و حتی می توان گفت، بیشتر هم نگاره ها کلماتی هستند که فراوانی بالایی در پیکره ها دارند. اکثر این هم نگاره ها، به نظر می رسد در اثر یکسان بودن نمود نوشتاری تکواژ یاء نکره ، یاء اسم ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنی یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسه دوم شخص مفرد و یاء صفت ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) و یاء متصل به گروه اسمی باشند.

# یکسان بودن نمود نوشتاری تکواژهایی که در نهایت منجر به شکل گیری هم نگاره ها می شوند:

الف. یاء نکره: کشاورزی را دیدم.

ب. یاء اسم ساز: وی چند سال کشاورزی را دنبال می کرد.

پ. یاء شناسه دوم شخص مفرد: تو، خودت کشاورزی.

# کاربرد جمع آوری داده ها:

۱- نشان دهنده تعداد بسیار زیاد هم نگاره ها در پیکره های متنی فارسی است که خود نشان می دهد در صورت رفع ابهام نکردن از هم نگاره ها در متون گوناگون، پردازش متن با مسائل و مشکلاتی روبرو خواهد شد.

۲- فهرست تهیه شده از هم نگاره ها از پیکره های گوناگون در طرح های بعدی مورد استفاده قرار خواهد گرفت؛ به این صورت که هم نگاره هایی که فراوانی بالاتری دارند در بافت نحوی مورد مطالعه قرار خواهند گرفت و با توجه به ساخت نحوی که در آن به کار رفته اند، برچسب های درست به آن ها داده خواهد شد.

# سپاسگزاری:

از جناب آقای دکتر مصطفی عاصی که بسیار از نظرات عالمانه ایشان در پیشبرد این پژوهش بهرمند شدم، کمال تشکر را می نمایم. همچنین از آقای مرتضی رضایی که در بخش جستجوی داده ها بنده را یاری رساندند، بسیار سپاس گزارم.

از صبر و توجه شما سپاسگزارم