دانشکده علوم و فنون نوین
گروه فناوری‌های بین رشته‌ای

آزمایشگاه شبکه‌های کامپیوتری و اطلاعاتی

تحلیل‌گر تصریفی کلمات فارسی غیر رسمی

دانشجو: **داود حیدرپور**

رشته تحصیلی: **زبان‌شناسی رایانشی**

استاد راهنما: **دکتر صالحی، دکتر بی‌جن‌خان**

استاد مشاور: **دکتر ویسی**

تاریخ دفاع: **۳۰ بهمن ۱۳۹۶**

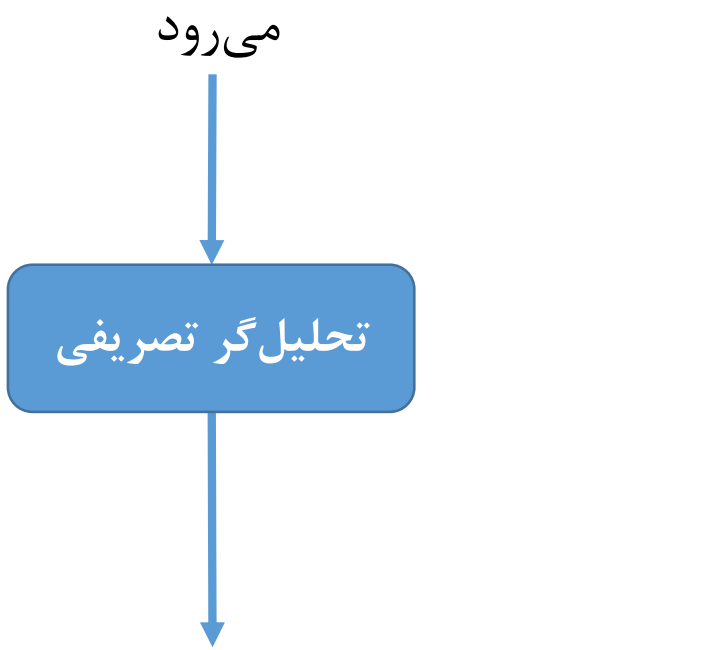
فهرست مطالب


مقدمه
کارهای پیشین
راهکار پیشنهادی
ارزیابی راه حل و مقایسه
نتیجه‌گیری و کارهای آتی

مقدمه

تحلیل‌گر تصریفی چیست؟

- جمله: او به مدرسه می‌رود.
- کاربردها
  - سیستم‌های پرسش و پاسخ
  - خلاصه‌سازی متن
  - پردازش‌های زبانی
  - و ...



گونه‌های فارسی

- گونه رسمی فارسی
  - روزنامه‌ها و نشریات
  - قوانین و احکام
  - مکاتبات
  - متون آموزشی و کتاب‌ها
- گونه غیر رسمی
  - شبکه‌های اجتماعی
  - پیام رسان‌های تلفن همراه
  - وبلاگ‌ها
  - و ...

اون به مدرسه رفت و بعد از کلاساش برگشت خونه.

کلمات رسمی

تفاوت ساخت رسمی و غیر رسمی

- رسمی : او به مدرسه می‌رود ← غیر رسمی : (اون) به مدرسه میره.
- کلمه غیر رسمی
  - واحدهای واژگانی مستقل
    - آسمون: آسمان
    - نون: نان
    - هیچکدوم: هیچکدام
    - اصاب: اعصاب
    - اگه: اگر
  - نگارش متفاوت از رسمی
    - عرزش: ارزش
    - عاریایی: آریایی
    - دخدر: دختر
    - انرجی: انرژی
    - امبار: انبار

تحلیل‌گر تصریفی

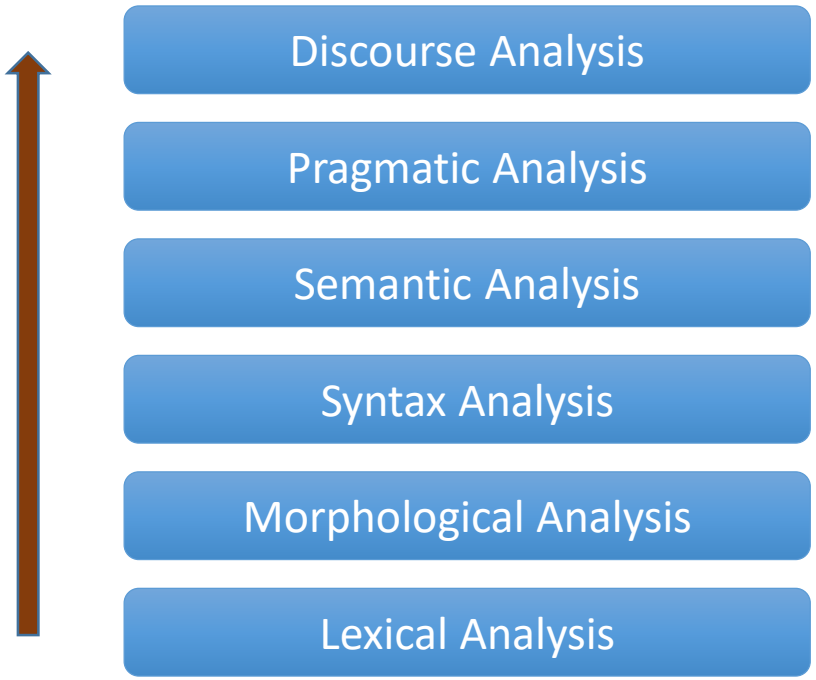
فعل مضارع اخباری غیر رسمی

=

ر + شناسه سوم شخص مفرد غیر رسمی

اهمیت موضوع و کاربردها

- گسترش تصاعدی فارسی غیر رسمی
- عدم وجود منابع برای فارسی غیر رسمی
- فقدان وجود ابزار تحلیل تصریفی فارسی غیر رسمی
- کاربردها:
  - نظر کاوی
  - بازشناسی گفتار
  - دسته‌بندی / طبقه‌بندی متون محاوره
  - آموزش فارسی
  - و ....



(جورافسکی و مارتین، ۲۰۰۸)

چالش‌ها و نیازمندی‌ها

- منابع مورد نیاز برای تحلیل تصریفی کلمه بنابر جورافسکی و مارتین (۲۰۰۸)

	فارسی رسمی	غیر رسمی	
• واژگان	√	×	خسته(صفت)
• قواعد تصریفی	√	×	خسته + م (واژه‌بیست ربطی اول شخص مفرد)
• قواعد نگارشی	√	×	خسته‌ام

کارهای پیشین

روش مناسب

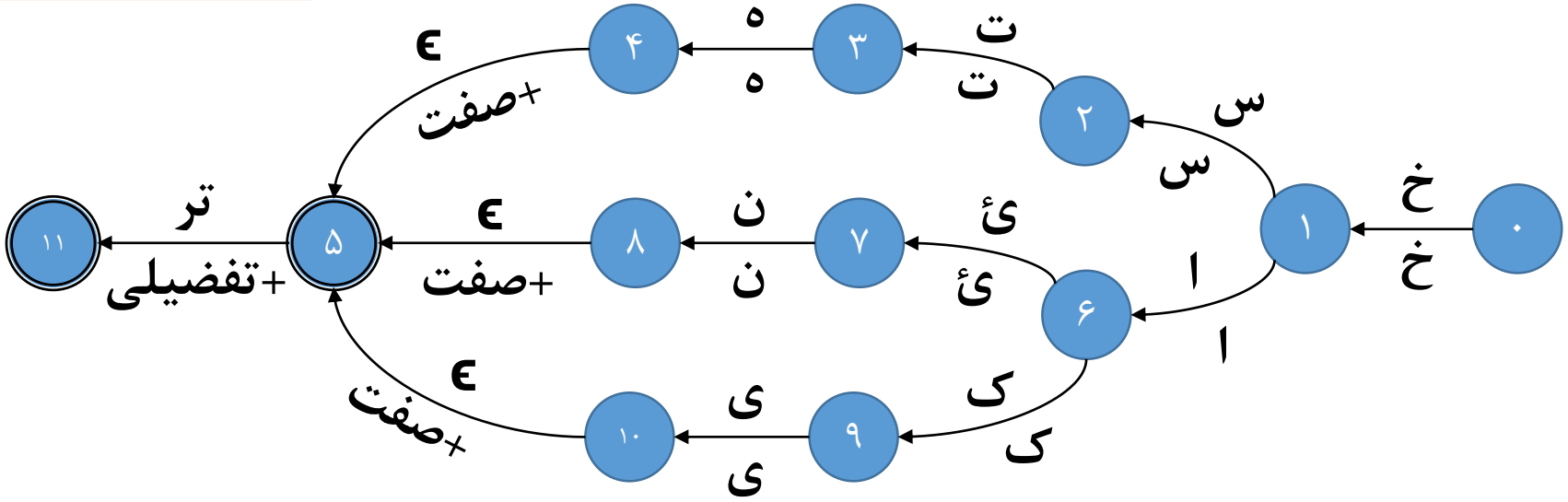
- آماری
  - تجزیه، تقطیع کلمه
    - درختان ← درخت + ان
    - رفتند ← رفت + ند
  - برچسب زنی قسم کلمه
    - درختان ← اسم، رفتند ← فعل
- مکاشفه‌ای
  - ریشه‌یاب
    - درختان، درخت‌ها، درختی ← درخت
    - رفتند ← رفت / رو
- قاعده‌بنیاد
  - تحلیل تصریفی
    - درختان: اسم = درخت + جمع جانداران
    - رفتند: فعل ماضی ساده = رفت + شناسه سوم شخص جمع

تحلیل تصریفی قاعده بنیاد

- مبدل یا اتوماتای حالت متناهی
  - منابع زبانی (جورافسکی و مارتین ۲۰۰۸)
    - واژگان
    - قواعد تصریفی
    - قواعد نگارشی
- مثال ساخت تحلیل‌گر با توجه به منابع زبانی زیر
  - واژگان: صفت‌های خسته، خائن، خاکی
  - قواعد تصریفی: صفت، صفت+تفضیلی. مثال: خائن، خاکی‌تر
  - بدون قواعد نگارشی (برای سادگی)

اتوماتا یا مبدل حالت متناهی

FST – Finite State Transducer:



خسته = خسته + صفت

خائن = خائن + صفت

خاکی= خاکی + صفت

خسته‌تر =	خسته + صفت + تفضیلی
خائن‌تر =	خائن + صفت + تفضیلی
خاکی‌تر =	خاکی + صفت + تفضیلی

خسته‌تر ⟺ مبدل (تحلیل‌گر تصریفی) ⟺ صفت = خسته + تفضیلی

بررسی کارهای انجام شده فارسی رسمی و غیر رسمی

مقاله	رویکرد	وابستگی به بافت	نوع ابزار	منبع واژگان	ارزیابی	مزایا	معایب
مگردومیان، ۲۰۰۰	تصریف	مستقل از متن	مبدل	پروژه شیراز	دقت ۹۵٪ بر روی پیکره ۷مگاباتی متون خبری	• یکپارچه بودن • پوشش کامل زبان رسمی	فقدان پشتیبانی از فارسی غیر رسمی
شمس‌فرد، ۲۰۱۰	تصریف و صرف جزئی	مبتنی بر بافت	اوتوماتا و نگاه به واژگان	واژگان زایا	برای ۶۰۰ کلمه پوشش ۹۸٪ و درستی ۹۳٪	مبتنی بر بافت بودن	• خلط اشتقاق و تصریف • فقدان پشتیبانی از فارسی غیر رسمی • یکپارچه نبودن
روحانیان ، ۱۳۹۳	تصریف	مستقل از متن	مبدل	واژگان زایا	۹۵ درصد	یکپارچه بودن	• فقدان پشتیبانی از فارسی غیر رسمی • عدم پوشش کامل زبان
مگردومیان، ۲۰۰۶ و ۲۰۰۸	تصریف	-	-	-	–	-	-

راهکار پیشنهادی

منابع زبانی

- واژگان
  - رسمی: زایا (اسلامی و همکاران، ۱۳۸۳)
  - غیر رسمی: استخراج از پیکره، معادل سازی از روی رسمی
- قواعد تصریفی
  - رسمی: زایا، دستور زبان فارسی (گیوی و انوری، ۱۳۹۱)
  - غیر رسمی: استنتاج از پیکره، شم زبانی، تحلیل وبلاگ‌ها (مگردومیان، ۲۰۰۸)
- قواعد نگارشی
  - رسمی: مبتنی بر دستور خط فارسی فرهنگستان (صادقی و مقدم، ۱۳۸۵)
  - غیر رسمی: استنتاج از قواعد تصریفی

قسم کلمه	رسمی	غیر رسمی
اسم	۲۷،۹۴۰ (۱،۹۵۱ جاندار، ۷۹۱ جمع مکسر، ۶۵۰ سایر جمع‌های عربی)	۸۰
صفت	۱۶،۶۵۰	۴۰
قید	۱،۳۷۴	۲۲۷
حرف ربط	۸۹	۲۴
حرف ربط گروهی	۲۱۳	۲
صفت شمارشی	۴۸	۱۷
حرف اضافه	۲۷	۴
حرف اضافه گروهی	۱۳۷	-
ضمیر شخصی	۱۰	۳
ضمیر یا صفت مبهم	۲۷	۸
ضمیر یا صفت اشاره	۲۲	۱۰
ضمیر یا صفت پرسشی	۲۸	۴
اسم خاص مکان	۱،۵۷۷	۵
اسم خاص اشخاص	۱،۹۵۰	۱۳
اسم خاص فامیل	۲،۷۱۵	-
شبه جمله	۱۹۳	۲۳
شاخص	۲۹	۳
بن مضارع ساده	۳۷۹ (۸۳ لازم، ۲۹۶ متعدی)	۸۷ (۹ لازم، ۷۸ متعدی)
بن ماضی ساده	۵۱۶ (۱۶۰ لازم، ۳۵۶ متعدی)	۸۹ (۱۳ لازم، ۷۶ متعدی)
ماضی پیشوندی	۴۹ (۱۲ لازم، ۳۷ متعدی)	۲۳ (۱۱ لازم، ۱۲ متعدی)
مضارع پیشوندی	۴۹ (۹ لازم، ۴۰ متعدی)	۲۲ (۶ لازم، ۱۶ متعدی)

ساختار کلمات

- فعل‌ها
  - ساده
  - پیش‌وندی
- غیر فعلی‌ها (به شرط موجود بودن در واژگان)
  - واحدهای چند قطعه‌ای
  - قطعه‌های چند واحدی
- کلمه قابل پذیرش (جداساز در چه سطحی باید جداسازی کند؟)
  - (تک‌واژ وابسته+) تک‌واژ مستقل (+تک‌واژ وابسته)
- کلمه رسمی و غیر رسمی
  - کلمه رسمی از اجزاء و ساخت رسمی بهره می‌برد.
  - کلمه غیر رسمی حداقل یک جزء غیر رسمی یا ساخت غیر رسمی دارد و یا هر دو غیر رسمی است.

فعل‌ها

- ساده
- پیش‌وندی

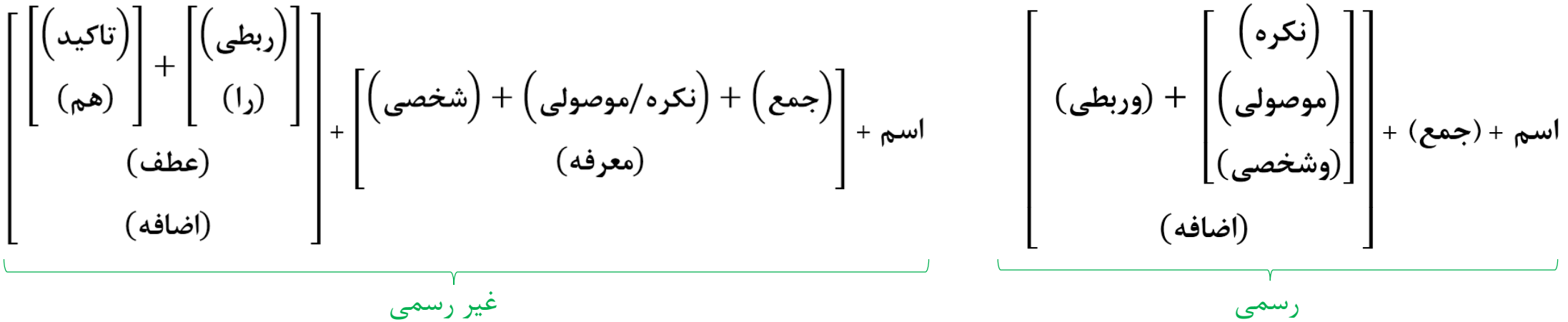
رسمی

$$\underbrace{\underbrace{(\text{پیش‌وند}) + \left[\begin{matrix}(\text{تک‌واژ نفی}) + (\text{تک‌واژ استمراری}) \\ (\text{تک‌واژ التزامی})\end{matrix}\right] + \left[\begin{matrix}\text{بن ماضی} \\ \text{بن مضارع}\end{matrix}\right] + \text{شناسه}}_{\text{رسمی}} + \left[\begin{matrix}(\text{واژه‌بست‌های مفعولی}) \\ (\text{واژه‌بست فاعلی})\end{matrix}\right] + (\text{تک‌واژ محاوره})}_{\text{غیر رسمی}}$$

- موارد ویژه
  - فعل‌های ناقص
  - تصریف ویژه برخی ریشه‌های غیر رسمی در ساخت مضارع و امری
- مثال
  - بردار، پاشین، نرو، نذاشتم، گفت، خوردشون، چرخوندتم، برن، می‌زنمتا

غیر فعلی‌ها

- اسم



- صفت، قید، اشاره، مبهم، پرسشی، تعجبی، ضمیر شخصی، ضمیر مشترک، حرف اضافه، شماره، حرف ربط، مصدر، صفت مفعولی، شبه جمله، شاخص.
- مثال
  - آلودگیست، ادماینا، بهترینشونه، بعدنم، اینارو، اوناهاش، هیشکیو، کدومیشونه، کوشن، شماهایین، خودشونن، باهاشون، دومی، سه‌تا، اگه، برنگشتن، پاشدن، دورمده‌هه، جون، عمه‌ام.

قواعد نگارشی یا بازنویسی

- مشخص کردن مرز تک‌واژها
  - در مرحله تولید با استاندارد تعریف شده هر قاعده یک کلمه تولید می‌کند.
  - در مرحله شناسایی مرز تک‌واژها هر سه نویسه فاصله، نیم‌فاصله و اتصال (تهی) است.
- همه تک‌واژها (مستقل و وابسته) در صورتی که به نویسه واکه منتهی شوند، توسط نشانه‌های فرامتنی علامت گذاری می‌شوند.
  - برخی از این تک‌واژها برای اتصال به برخی دیگر تک‌واژها واج میانجی لازم دارند.
  - واکه پایانی برخی از این تک‌واژها در هنگام اتصال به تک‌واژی دیگر حذف یا تبدیل می‌شود.

پیاده سازی

- فن‌آوری زیراکس
  - صورت‌بندی لکس
    - قواعد تصریفی
    - قواعد نگارشی
- دستورالعمل زبان قاعده‌مند و مقید
  - قواعد نگارشی
- ابزار فوما

نمونه‌ای از پیاده سازی با صورت‌بندی لکس

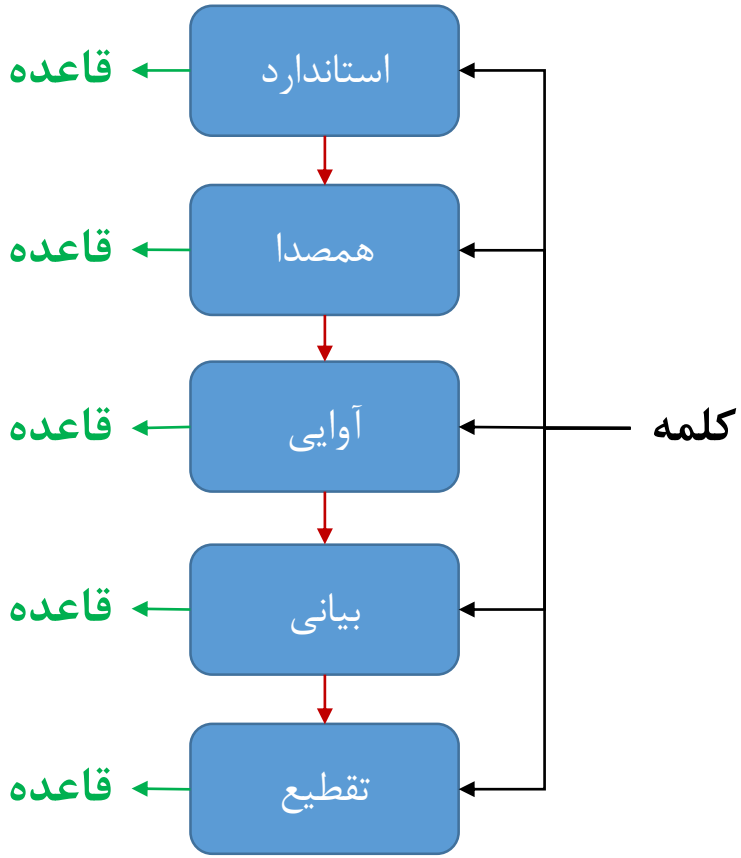
```
@D.VOWEL@@C.FORMAL@@P.VOWEL.HEH@+وربطی۳:@D.VOWEL@@C.FORMAL@@P.VOWEL.HEH@*ه takid2;
@D.VOWEL.HEH@@C.VOWEL@+وربطی۳:@D.VOWEL.HEH@@C.VOWEL@*ست takid2;
@R.VOWEL.HEH@@C.VOWEL@+وربطی۳:@R.VOWEL.HEH@@C.VOWEL@×ست takid2;
@R.VOWEL.VAV@@C.VOWEL@@C.FORMAL@@P.VOWEL.HEH@+وربطی۳:@R.VOWEL.VAV@@C.VOWEL@@C.FORM
@R.VOWEL.VAV@@C.VOWEL@@P.VOWEL.HEH@+وربطی۳:@R.VOWEL.VAV@@C.VOWEL@@P.VOWEL.HEH@*ست
@R.VOWEL.HEH@@C.VOWEL@@C.FORMAL@+وربطی۳:@R.VOWEL.HEH@@C.VOWEL@@C.FORMAL@¤س takid2;
@R.VOWEL.YEH@@C.VOWEL@@C.FORMAL@@P.VOWEL.HEH@+وربطی۳:@R.VOWEL.YEH@@C.VOWEL@@C.FORM
@R.VOWEL.YEH@@C.VOWEL@@P.VOWEL.HEH@+وربطی۳:@R.VOWEL.YEH@@C.VOWEL@@P.VOWEL.HEH@*ست
@R.VOWEL.ALEF@@C.VOWEL@@C.FORMAL@+وربطی۳:@R.VOWEL.ALEF@@C.VOWEL@@C.FORMAL@*س takid2;
@R.VOWEL.ALEF@@C.VOWEL@+وربطی۳:@R.VOWEL.ALEF@@C.VOWEL@*ست takid2;
```

نمونه‌ای از دستورالعمل مقید

```
define iPutSpace "^" -> SPACE;
define putZWNJ "×" -> ZWNJ;
define copulaElim ["HEH" "¤"] (->) "×" || _ SIN .#.;
define words NOUNS .o. copulaElim;
regex words;
```

مبدل‌ها

- مبدل استاندارد
  - واژگان، قواعد تصریفی، قواعد نگارشی
  - غیر متعین برای فاصله، نیم‌فاصله و اتصال
  - غیر متعین برای ا/آ
- مبدل‌های ثانوی (شامل مبدل استاندارد نیز می‌شوند)
  - همصدا
    - غیر متعین برای حروف همصدا
    - شناسایی نویسه ه به عنوان بازنمایی کسره اضافه
  - آوایی
    - استفاده از الگوریتم آوایی برای شناسایی کلمات
  - بیانی
    - حذف تکرار حروف.
  - تقطیع
    - تقطیع کلمات به هم چسبیده و تحلیل هر یک



- مجموعه داده
- متریک‌های ارزیابی

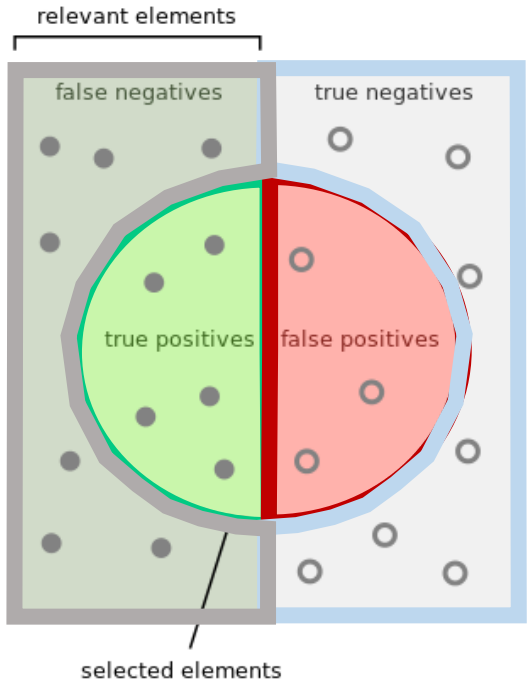
پیکره

- شناسایی ۲۶ زیر سیاق غیر رسمی
  - دسترسی به منابع ۱۹ زیر سیاق آن
- جمع‌آوری نزدیک به ۵۰ هزار توکن
- رعایت استانداردهای جمع‌آوری پیکره اسنپ‌شات در نمونه گیری و آماده سازی
- مشخص کردن کلمه‌های غیر رسمی
- انتخاب جمله‌های تصادفی از هر سیاق
  - بیش از هزار کلمه

شماره	سیاق		موجود
۱	گفت وگوی غیر خیالی	گفت وگوی عادی (خودانگیز)	
۲		گفت وگوی تلفنی	
۳		مصاحبه	√
۴		مناظره	
۵		صحنه دادگاه	
۶	گفت وگوی خیالی	رمان (اصل فارسی)	√
۷		نمایش‌نامه (اصل فارسی)	√
۸		فیلم‌نامه (اصل فارسی)	√
۹		زیرنویس (ترجمه به فارسی)	√
۱۰		زیرنویس (اصل فارسی)	
۱۱	نامه‌ی شخصی		√
۱۲	خاطرات روزانه		√
۱۳	شعر محاوره		√

شماره	سیاق		موجود
۱۴	سخنرانی	از پیش آماده شده	√
۱۵		خودانگیز	
۱۶	شوی رادیو و تویزیونی	اجرا از روی متن	√
۱۷		اجرای خودانگیز یا نیمه خودانگیز	
۱۸	وبلاگ	پست‌ها	√
۱۹		پاسخ‌ها	√
۲۰	شبکه‌های اجتماعی	پست‌ها	√
۲۱		پاسخ‌ها	√
۲۲	پیام‌رسان تلفنی	مکالمه‌ی شخصی	√
۲۳		مکالمه‌ی گروهی	√
۲۴		گروه اطلاع رسانی	√
۲۵	نظر کاربر فضای مجازی	مصرف کننده‌ی کالا یا خدمات	√
۲۶		مقالات خبری و متفرقه	√

شاخص ارزیابی مستقل از متن



- مثبتِ صحیح (TP)
  - قاعده‌ای تولید شده که به ازای یک تایپِ کلمه صحیح است.
- مثبتِ غلط (FP)
  - قاعده‌ای تولید شده که به ازای تمام تایپ‌های کلمه غلط است.
- منفیِ غلط (FN)
  - قاعده‌ای که تولید نشده اما می‌بایست برای تایپی از کلمه تولید می‌شده است.
- منفیِ صحیح (TN)
  - قاعده‌ای که نباید تولید می‌شده و تولید نشده است.

معیارهای ارزیابی

- فراخوانی(Recall)

$$\text{Recall} = \frac{\text{مثبتِ صحیح}\,(TP)}{\text{مثبتِ صحیح}\,(TP) + \text{منفیِ غلط}\,(FN)} = \text{فراخوانی}$$

  - چه مقدار از کل قاعده‌هایی که می‌بایست تولید می‌شده، تولید شده است.

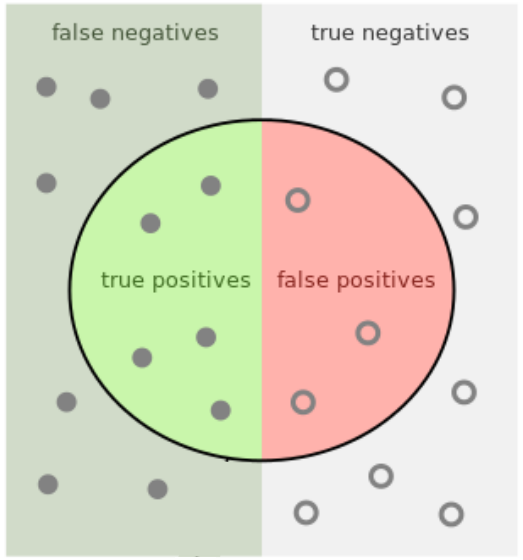
- درستی (Precision)

$$\text{Precision} = \frac{\text{مثبتِ صحیح}\,(TP)}{\text{مثبتِ صحیح}\,(TP) + \text{مثبت غلط}\,(FP)} = \text{درستی}$$

  - چه مقدار از قاعده‌های تولید شده صحیح‌اند.

- معیار اف (۱) (F1 Measure)

$$\text{معیار اف}(1) = 2 \times \frac{\text{فراخوانی} \times \text{درستی}}{\text{فراخوانی} + \text{درستی}}$$



ارزیابی کل کلمات

- ارزیابی بر روی بیش از هزار واژه انجام شده است.
- عمده خطاها (بیشتر از ۶۵٪) به علت کمبود واژگان است.
- به دلیل تولید قاعده‌های بیشتر در مبدل‌های ثانویه، مثبتِ غلط افزایش پیدا می‌کند.
  - معیار درستی را کاهش می‌دهد. در نتیجه معیار اف نیز کاهش می‌یابد.

معیار	استاندارد	همصدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۵.۲۳	٪۹۵.۵	٪۹۵.۵۱	٪۹۵.۵۱	٪۹۵.۶۵
درستی	٪۹۹.۵۶	٪۹۹.۷	٪۹۹.۳۴	٪۹۹.۳۴	٪۹۸.۴۱
معیار اف (۱)	٪۹۷.۳۴	٪۹۷.۵۵	٪۹۷.۳۸	٪۹۷.۳۸	٪۹۷.۰۱

ارزیابی کلمات رسمی موجود

- بیش‌تر از ۶۰٪ کلمات رسمی هستند.
- بیش از ۹۰٪ منفی‌های غلط به دلیل کمبود واژگان است (۴۳ مورد)
- ۴.۲٪ خطاها، خطای نگارشی (املایی) است (۲ مورد).
- ۲٪ خطاها، به دلیل نقص در ساختار قواعد تصریفی است (۱ مورد).
  - حرف اضافه + اشاره که کلمه ازین را تولید می‌کند درقواعد و واژگان نیست.
- مبدل‌های ثانویه غیر از یک مورد تاثیری در خروجی ندارند.
- مبدل تقطیع دو قطعه به هم چسبیده را اصلاح کرده است (۱ مورد خطا، ۲٪ از کل خطاها).
  - مثبت‌های غلط زیادی نیز تولید کرده که درستی را کاهش داده است.

معیار	استاندارد	همصدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۴.۲۲	٪۹۴.۲۲	٪۹۴.۲۲	٪۹۴.۲۲	٪۹۴.۳۴
درستی	٪۹۹.۵۱	٪۹۹.۵۱	٪۹۹.۵۱	٪۹۹.۵۱	٪۹۷.۹۶
معیار اف (۱)	٪۹۶.۷۹	٪۹۶.۷۹	٪۹۶.۷۹	٪۹۶.۷۹	٪۹۶.۱۱

ارزیابی کلمات غیر رسمی موجود

- نزدیک به ۴۰٪ کلمات ارزیابی، غیر رسمی است.
- بیش از ۷۲٪ خطاها به دلیل کمبود واژگان است.
  - بیش از ۹۲٪ درصد خطای کمبود واژگان را مبدل آوایی اصلاح کرد.
- مبدل همصدا کسره اضافه‌ای که با حرف ه نمایش داده شده بود شناسایی کرد (۲ مورد، ۱۱٪ کل خطاها).

معیار	استاندارد	همصدا	آوایی	بیانی	تقطیع
فراخوانی	٪۹۶.۲۶	٪۹۶.۶۲	٪۹۷.۱۵	٪۹۷.۱۵	٪۹۷.۱۵
درستی	٪۹۹.۶۳	٪۹۹.۶۳	٪۹۹.۰۹	٪۹۹.۰۹	۹۹.۰۹
معیار اف (۱)	٪۹۷.۹۱	٪۹۸.۳۶	٪۹۸.۱۱	٪۹۸.۱۱	٪۹۸.۱۱

نتیجه گیری

خلاصه

- قاعده بنیاد
  - بر اساس مبدل حالت متناهی و با صورت‌بندی استاندارد زیراکس پیاده شده است.
  - مستقل از متن است.
- منابع زبانی رسمی از واژگان زایا، دستور زبان فارسی استفاده شده است.
- منابع زبانی غیر رسمی با تحلیل پیکره، معادل سازی کلمات رسمی و تحلیل وبلاگ‌ها (مگردومیان، ۲۰۰۸) انجام شده است.

دستاورد

- تحلیل‌گر تصریفی فارسی معاصر
  - اولین ابزار برای تحلیل تصریفی فارسی غیر رسمی
  - اولین منابع زبانی برای فارسی غیر رسمی
    - واژگان
    - قواعد تصریفی (ساختار جامع و مانع فعل و غیر فعل)
    - قواعد نگارشی
  - فراهم کردن اولین دادگان ارزیابی برای فارسی غیر رسمی (dataset)

کارهای آتی

- بهبود
  - عمده خطاها به دلیل کمبود واژگان است.
  - چند مورد قواعد تصریفی رسمی و غیر رسمی می‌بایست اصلاح و یا اضافه شود.
- توسعه تحلیل‌گر
  - مبتنی بر بافت کردن تحلیل‌گر
  - شناسایی و رفع خطای املایی مبتنی بر بافت
  - قرار دادن تحلیل تصریفی در پایین‌ترین سطح پردازش با افزودن تصریف تک‌واژهای وابسته
    - بی‌نیاز شدن از جداساز
  - مشخص کردن برچسب دقیق کلمه در پس‌پردازش
- ابزارهای دیگر
  - ساخت تقطیع‌گر واژه
  - ساخت مبدل رسمی به غیر رسمی و برعکس
  - ساخت ریشه‌یاب
  - استفاده برای استخراج ویژگی برای سامانه‌های یادگیری ماشین

		رسمی	غیر رسمی
همه کلمات		۸۵۷	۲۰۶
کلمات دارای تصریف		۱۹۸	۱۴۴
کلمات بدون تصریف		۶۵۱	۶۲
نسبت تعداد تک‌واژهای ساختمان تصریفی	به کل کلمات	۰.۲۷۷۷	۰.۹۵۶۳
	به کلماتی که تصریفی‌اند	۱.۲۰۲	۱.۳۶۸

مراجع

1. آرمین، نادیه و شمس‌فرد، مهرنوش. ۱۳۸۹. «تبدیل متن محاوره‌ای فارسی به رسمی به کمک ان-گرام‌ها» در دانشگاه صنعتی شریف.
2. اسلامی، محرم و علی‌زاده لجمیری، صدیقه. ۱۳۸۸. «ساختار تصریفی کلمه در زبان فارسی» در دانشگاه تبریز. نشریه زبان و ادب فارسی. شماره ۲۱۱.
3. اسلامی، محرم، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری و طاهره زندی. ۱۳۸۳. «واژگان زایای زبان فارسی» در اولین کارگاه پژوهشی زبان فارسی و رایانه.
4. احمدی گیوی، حسن و انوری، حسن. ۱۳۹۱. «دستور زبان فارسی». ویرایش چهارم. انتشارات فاطمی.
5. روحانیان، مجتبی، غفاری، کامران و وزیرنژاد، بهرام. ۱۳۹۳. «طراحی تحلیلگر ساختواژی برای زبان پارسی، با استفاده از ساختار تراگذر و روش دوسطحی» در سومین همایش ملی زبان شناسی رایانشی، دانشگاه صنعتی شرف.
6. صادقی، علی‌اشرف و زندی مقدم، زهرا. ۱۳۸۵. «فرهنگ املایی زبان فارسی» فرهنگستان زبان و ادب فارسی.
7. Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing, 2nd edition*. Prentice Hall.
8. Karine Megerdoomian. (2000). Persian Computational Morphology: A Unification-Based Approach. Computing Research Laboratory, New Mexico State University.
9. Karine Megerdoomian. (2006). Extending a Persian Morphological Analyzer to Blogs.
10. Karine Megerdoomian. (2008). *Analysis of Farsi Weblogs*.
11. Kenneth R Beesley, & Lauri Karttunen. (2003). *Finite state morphology (Xerox)*. Stanford, Calif.
12. Sarabi, Z., Mahyar, H., & Farhoodi, M. (2013). ParsiPardaz: Persian Language Processing Toolkit. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on* (pp. 73–79). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6682862
13. Shamsfard, M., Jafari, H. S., & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. In *LREC*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/809_Paper.pdf

پایان