امکان‌سنجی پروژه‌های زیرساختی کاربری خط و زبان فارسی در محیط رایانه‌ای

فهرست مطالب

عنوان — شماره صفحه

عنوان **شماره صفحه**

1. مقدمه

امروزه با گسترش کاربرد کامپیوتر، نیاز به استفاده از توانایی‌های غیرقابل چشم‌پوشی آن در حوزه‌ی زبان شناسی نیز به شدت احساس می‌شود. حوزه‌های پردازش زبان طبیعی[1] و زبان شناسی رایانه‌ای[2] به تلاش برای ماشینی کردن فرایند زبان شناسی سنتی می‌پردازند. زبان شناسی رایانه‌ای شاخه‌ای از هوش مصنوعی است که منشا پیدایش آن را می‌توان با هم‌زمان با شکل‌گیری تلاش‌هایی برای تولید ماشین خودکار ترجمه برای ترجمه مجلات علمی روسی به انگلیسی در دهه 50 میلادی در ایالات متحده آمریکا دانست [1]. پس از شکست پروژه ترجمه ماشینی، مشخص گردید که پردازش خودکار زبان طبیعی بسیار پیچیده‌تر از آن چه که تصور می‌شد است.

پردازش زبان طبیعی عبارتست از استفاده از رایانه برای پردازش زبان گفتاری و نوشتاری. پردازش زبان طبیعی می‌تواند در سطوح مختلف از زبان صورت گیرد، که پردارش در تمامی این سطوح برای ترجمه ماشینی لازم است. این سطوح را می‌توان به شکل زیر تقسیم‌بندی نمود [2]:

1) آواشناسی و صدا شناسی[3] که به تشخیص آواها و صداها و بازشناسی گفتار می‌پردازد.
2) ریخت‌شناسی[4] که به ساختارهای کلمات و ریشه‌یابی واژگان می‌پردازد.
3) نحو[5] که به ارتباط کلمات به همدیگر و مباحث دستوری آن‌ها در گروه‌ها و جملات می‌پردازد.
4) معناشناسی[1] که به ارتباطات معنایی کلمات می‌پردازد.

---

[1] Natural Language Processing - NLP

[2] Computational Linguistics

[3] Phonetics and Phonology

[4] Morphology

[5] Syntax

5) عمل‌گرایی[۲] که کاربردهای زبان برای رساندن یک مطلب به مخاطب یا مخاطبـان، در حالـت عملی و یا در نوشتار و گفتار طبیعی می‌پردازد.

6) مباحثه[۳] که به ارتباطات کلی یک زبان فرای یک یا چند جمله خاص می‌پردازد.

از کاربردهای اصلی پردازش زبان طبیعی می‌توان موارد زیر را ذکر کرد:

1) خلاصه‌سازی خودکار[٤]

2) کمک به خواندن زبان‌های طبیعی دیگر[٥]

3) کمک به نوشتن به زبان‌های طبیعی دیگر[٦]

4) استخراج اطلاعات[۷]

5) بازیابی اطلاعات[۸]

6) ترجمه ماشینی[۹]

7) تشخیص واحدهای اسمی[۱۰]

[۱] Semantics

[۲] Pragmatics

[۳] Discourse

[٤] Automatic Summarization

[٥] Foreign language reading aid

[٦] Foreign language writing aid

[۷] Information Extraction

[۸] Information Retrieval

[۹] Machine Translation

[۱۰] Name Entity Recognition

8) تولید زبان طبیعی[۱]

9) فهم زبان طبیعی[۲]

10) نویسه‌خوان نوری[۳]

11) تحلیل مرجع‌دارها[٤]

12) سیستم سوال، پاسخ[٥]

13) تشخیص گفتار[٦]

14) مبدل متن به گفتار[۷]

15) نظام‌های مکالمه گفتاری[۸]

16) ساده‌سازی متن[۹]

17) تایید متن[۱۰]

---

[۱] Natural language generation

[۲] Natural language Understanding

[۳] Optical Character Recognition - OCR

[٤] Anaphora Reservation

[٥] Question Answering System

[٦] Speech Recognition

[۷] Text-to-Speech

[۸] Spoken Dialogue System

[۹] Text Simplification

[۱۰] Text Proofing

برخی مسائل اساسی در پردازش زبان طبیعی را می‌توان به صورت زیر برشمرد:

1) قطعه‌بندی گفتار[۱]
2) قطعه‌بندی متن[۲]
3) برچسب‌گذاری نقش کلمه[۳]
4) ابهام‌زدایی از نقش کلمه[٤]
5) ابهام‌زدایی نحوی[٥]
6) هنجارسازی[٦]
7) تشخیص اعمال گفتاری[۷]

پردازش زبان طبیعی در حوزه متن یا خط و زبان به پردازش پیکره‌های متنی[۸] به عنوان نمایانگر زبان می‌پردازد. بنابر این مسائل اصلی در حوزه خط و زبان را می‌توان به موارد زیر کاهش داد:

---

[۱] Speech Segmentation
[۲] Text Segmentation
[۳] Part-of-Speech Tagging
[٤] Word Sense Disambiguation
[٥] Syntactic Disambiguation
[٦] Normalization
[۷] Speech acts
[۸] Text Corpus

1) قطعه‌بندی متن
2) برچسب‌گذاری نقش کلمه
3) ابهام‌زدایی از نقش کلمه
4) ابهام‌زدایی نحوی
5) هنجارسازی

مسائل فوق تقریبا در تمامی کاربردهای پردازش زبان طبیعی در حوزه خط و زبان مطرح هستند و به عنوان مسائل زیرساختی خط و زبان می‌باید مورد بررسی قرار گیرند. زبان فارسی به عنوان یکی از زبان‌های طبیعی نیز از این قاعده مستثنی نیست. از طرفی همان طور که مطرح شد، در حوزه خط و زبان، پیکره‌های متنی به عنوان نمادی از زبان هستند که می‌باید با تحلیل آن‌ها به استخراج اجزا، قواعد و ساز و کار زبان پی برد.

دادگان مهم و زیرساختی حاصل از تحلیل پیکره‌های متنی در پردازش زبان طبیعی و حوزه خط و زبان را می‌توان به موارد زیر تقسیم نمود:

1) پیکره متنی
2) واژگان زبان (واژگان عمومی و تخصصی)
3) گنجواژه یا اصطلاح‌نامه[1]
4) الگوهای زبان
5) پیکره‌های برچسب داده‌ای

[1] Thesaurus

6) پیکره‌های تخصصی

پروژه‌های زیرساختی در حوزه خط و زبان فارسی را می‌توان پروژه‌هایی برای ایجاد دادگان زیرساختی در حوزه خط و زبان فارسی در نظر گرفت. از طرفی ابزارهایی نیز جهت حل مسائل زیرساختی مطرح شده در حوزه خط و زبان نیز نیاز هستند تا بتوان اقدام به تحلیل پیکره‌های متنی زبان جهت تولید دادگان زیرساختی نمود. بنابراین پروژه‌های زیرساختی خط و زبان فارسی تلفیقی از ابزارها و دادگان زیرساختی زبان خواهد بود.

2. دادگان زیرساختی در حوزه خط و زبان فارسی

همان گونه که در بخش پیشین نیز مطرح شد، برای انجام هرگونه فعالیت در حوزه کاربر خط و زبان فارسی، نیاز به یک سری دادگان زیرساختی است که خصوصیات زبان را بیان کنند. این دادگان شامل: 1) پیکره‌های متنی، 2) واژگان زبان، 3) گنج‌واژه‌ها، 4) الگوهای زبان، 5) هم‌نشینی‌های رایج در زبان، 6) پیکره‌های برچسب داده‌ای، 7) پیکره‌های تخصصی هستند که به تفصیل به هر یک خواهیم پرداخت.

2-1. پیکره متنی

در علوم زبان شناسی و زبان شناسی رایانه‌ای، پیکره متنی، حجم بسیاری از متون ساخت‌یافته آن زبان است. پیکره‌های متنی به منظور تحلیل‌های آماری، صحت‌سنجی فرضیه‌ها و بررسی رخداد یا صحت قواعد زبانی در حوزه‌ای مشخص به کار می‌روند. پیکره‌های متنی به عنوان پایگاه دانش[۱] اصلی در زبان‌شناسی رایانه‌ای خصوصا در حوزه خط و زبان شناخته می‌شوند. پیکره‌های می‌توانند تک زبانی[۲] - شامل متون با یک زبان و یا چند زبانی[۳] - شامل متون با چندین زبان باشند، که مورد اخیر معمولا برای مقایسه نظیر-به-نظیر ساختار دهی می‌شود و پیکره موازی منطبق[۴] نامیده می‌شود.

به منظور مناسب‌سازی پیکره‌های متنی برای کاربرد مفیدتر در مطالعات زبان‌شناسی، پیکره‌ها می‌باید حاشیه‌نویسی[۵] شوند. حاشیه‌نویسی عبارت است از تحلیل و افزودن برخی اطلاعات مانند اطلاعاتی در مورد نقش[۶] و یا ریشه[۷] کلمات موجود در متن به پیکره. پیکره‌های متنی اکثرا دارای در سطح ساختاری

---

[۱] Knowledgebase

[۲] Monolingual

[۳] Multilingual

[۴] Aligned Parallel Corpus

[۵] Annotated

[۶] Part-of-Speech - POS

[۷] Lemma

حاشیه‌نویسی می‌شوند. پیکره‌هایی که کاملا تجزیه[1] و حاشیه‌گذاری شده باشند، پیکره‌های تجزیه شده یا بانک درختی[2] گفته می‌شود. عملا پیکره‌های کوچک (بین 1 تا 3 میلیون کلمه) امکان تجزیه کامل را دارند زیرا حصول اطمینان از این که حاشیه‌نویسی کاملا صحیح و سازگار است، عملی بسیار پیچیده، زمان‌گیر و پر هزینه خواهد بود. امکان حاشیه‌نویسی در سطوح ریخت‌شناسی، نحوی، معنایی و کاربردی معمولا برای پیکره‌های کوچک امکان پذیر است.

از چالش‌های مطرح در خصوص پیکره‌های متنی می‌توان به این نکته اشاره کرد که چه پیکره‌ای با چه خصوصیاتی می‌تواند به بهترین و بیشترین شکل خصوصیات زبان را بیان سازد. با توجه به تغییرپذیری زبان در زمان، این نکته که چه متونی با چه ویژگی‌هایی می‌توانند پیکره‌ی مناسبی را شکل دهند اهمیت می‌یابد. این که چه سهمی از پیکره از زبان معیار انتخاب شود، چه میزان آن از زبان فوق معیار باشد و چه بخشی از آن را زبان زیر معیار دربر گیرد از چالش‌های اساسی در انتخاب و ایجاد پیکره‌ها است. همچنین تعریف مناسب از زبان معیار، فوق و زیر معیار نیز در این امر تاثیر گذار است و می‌باید مورد توجه قرار گیرد.

2-2. واژگان زبان

یکی از مهم‌ترین اجزاء از یک زبان کلمات تشکیل دهنده آن زبان هستند که مفاهیم در هر زبان از طریق آن‌ها انتقال می‌یابد. کلمه کوچکترین جزء آزاد از زبان است که معنا و یا کاربرد مشخصی دارد؛ در مقایسه واژک[3] کوچکترین واحد معنایی زبان است. کلمات می‌توانند از یک و یا چند واژک تشکیل شوند. مانند کلمه «آب‌سردکن» که از سه واژک تشکیل شده است. عموما یک کلمه از ترکیب یک ریشه یا بن و وندها[4] حاصل می‌شود. کلماتی که از اتصال بیش از یک ریشه حاصل شده باشند کلمات مرکب[5] نامیده

---

[1] Parse

[2] Treebank

[3] Morpheme

[4] Affix

[5] Compound Word

می‌شوند. کلماتی که از ترکیب کلمات موجود و یا بخشی از آن‌ها حاصل شوند نیز کلمات آمیخته[1] نامیده می‌شوند؛ مانند کلمه «رزمایش» که از آمیخته شدن دو کلمه «رزم» و «آرایش» و با تلفیق معانی به دست آمده.

استخراج و تولید لیستی از کلمات زبان، اصلی‌ترین و یا یکی از اصلی‌ترین نیازمندی‌های تحقیقات در زمینه زبان‌شناسی رایانه‌ای است. اما تشخیص کلمات آن چنان که تصور می‌شود آسان نیست. تشخیص کران کلمات[2] چالش اصلی در تشخیص و استخراج کلمات از پیکره‌های متنی است.

2-2-1. تشخیص کران کلمه

همان طور که پیش از این گذشت، تشخیص کران کلمه چالشی در تشخیص و استخراج کلمات زبان از پیکره‌های متنی است. در شیوه‌های نوین نگارش درست[3]، جداکننده کلمات معمولا نویسه «فاصله» است. اما مشکل جایی است که این نویسه جزئی از کلمه است. مثلا «از ما بهتران» بیانگر یک کلمه است که در میان خود شامل نویسه جدا کننده است. مواردی از قبیل:

1) افعال مرکب
2) افعال مجهول
3) مرکب‌های اتباعی
4) ترکیبات اضافه چند جزئی

مواردی هستند که با فرض رعایت شیوه نگارش صحیح و درست نویسی مشکلاتی را در زمینه تشخیص کران کلمه در زبان‌شناسی رایانه‌ای به وجود می‌آورند. این مشکلات هنگامی نمود بیشتری پیدا می‌کند

[1] Portmanteau
[2] Word Boundary
[3] Orthography

که خط فارسی و دستور زبان و شیوه نگارش صحیح فارسی نیز دارای ابهامات و پیچیدگی‌های خاص دیگری نیز هست. مواردی از قبیل:

1) وجود شبه[1]-فاصله به عنوان یکی از نویسه‌های زبان
2) وجود ابهام در دستور خط زبان فارسی برای قوانین پیوسته و جدانویسی
3) بازگذاشتن دست نویسندگان در فاصله‌گذاری میان کلمات
4) عدم وجود دستورالعمل قطعی برای استفاده از شبه-فاصله
5) عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات؛ استفاده از دستورالعمل مبتنی بر لغت (مانند تک‌هجایی بودن، بسیط‌گونه بودن)
6) عدم وجود رویکرد زبان‌شناسی رایانه‌ای در قوانین دستور خط زبان فارسی
7) تفاوت زیاد میان زبان محاوره و نوشتار
8) قابلیت زایائی زبان در خصوص ایجاد کلمات جدید

مواردی هستند که موجب چندگانه نویسی کلمات در زبان فارسی می‌شوند و تشخیص و استخراج کلمات از پیکره‌های متنی در زبان فارسی را دشوار می‌سازند.

2-2-2. لیست‌های بسامدی

لیست‌های بسامدی شامل لیستی از واژگان زبان به همراه میزان تکرار آن‌ها در پیکره‌ی معرف زبان است. بسامد تکرار واژگان می‌تواند نمودی از میزان رایج بودن کلمه در زبان باشد. لیست‌های بسامدی از کلمات از دادگان اساسی در تولید واژه‌نامه‌های کامپیوتری هستند.

[1] Pseudo-space

امروزه به نظر می‌رسد بسامد واژگان زبان نیز از قوانین توزیع زیپفین[1] که معمولا برای تخمین رفتارهای طبیعی و اجتماعی کاربرد دارد، پیروی می‌کند [3].

3-2. اصطلاح‌نامه‌ها

در نقطه‌ی مقابل واژه‌نامه‌ها که شامل تلفظ و شرح کلمات هستند، اصطلاح‌نامه‌ها شامل شبکه‌ای از لغات مترادف و در برخی موارد متضاد یک کلمه هستند. اصطلاح‌نامه‌ها لزوما شامل تمامی لغات مترادف نیستند اما برای تشخیص تفاوت‌های میان کلمات مشابه و انتخاب بهترین کلمه برای کاربردی خاص بسیار کارآمد هستند.

اصطلاح‌نامه‌ها می‌توانند کاربرد گسترده‌ای در بازیابی اطلاعات که یکی از کاربردهای اصلی پردازش زبان طبیعی است داشته باشند. از اصطلاح‌نامه‌ها می‌توان برای نمایه‌گذاری یا برچسب‌گذاری استفاده کرد. اصطلاح‌نامه‌های مورد کاربرد در بازیابی اطلاعات به گونه‌ای سامان یافته‌اند که رابطه‌ی صریحی میان مفاهیم در آن‌ها وجود داشته باشد. بنابراین این اصطلاح‌نامه‌ها پیچیده‌تر از تنها لیستی از کلمات مشابه و یا متضاد خواهند بود. هر کلمه در زمینه‌ای خاص رفتاری متفاوت دارد بنابراین در این نوع اصطلاح‌نامه‌ها کلمات در زمینه‌های گوناگون بیان می‌شوند و این امکان برای کاربران فراهم می‌شود که مثلا میان «شیر» به عنوان یک حیوان و «شیر» به عنوان مایعی نوشیدنی تمایز قائل شوند. با توجه به تفکیک زمینه‌ها در اصطلاح‌نامه‌های مورد کاربر در بازیابی اطلاعات، در کاربردهای هوش مصنوعی و پردازش زبان طبیعی به این گونه اصطلاح‌نامه‌ها، هستان‌شناسی[2] نیز گفته می‌شود.

کلمات واحدهای معنایی پایه جهت انتقال مفاهیم هستند. کلمات موجود در اصطلاح‌نامه‌ها معمولا کلمات اسمی تکی هستند. افعال قابلیت تبدیل به اسامی را دارند، صفات و قیود نیز معمولا به ندرت به انتقال مفاهیم موثر برای نمایه‌گذاری می‌پردازند بنابر این بخش اعظم اصطلاح‌نامه‌ها را اسامی تشکیل می‌دهند.

روابط میان کلمات می‌توانند به سه گونه‌ی زیر تقسیم شوند:

---

[1] Zipfian

[2] Ontology

1) *رابطه سلسله مراتبی*[1]: برای مشخص کردن کلماتی که در طول هم در یک زمینه قرار دارند به کار می‌رود. «کلمات کلی‌تر[2]» کلماتی هستند که عمومی‌تر از کلمه جاری هستند. مثلا: «وسیله» یک مفهوم کلی‌تر برای «کامپیوتر» است و نیز «کلمات جزئی‌تر[3]» کلماتی هستند که در مفهوم مورد نظر جزئی‌تر از کلمه جاری هستند مثلا «پدر» یک مفهوم جزئی‌تر برای «والدین» است.

2) *رابطه معادل* بودن که برای ارتباط میان کلمات مشابه به کار گرفته می‌شود.

3) *رابطه انجمنی* که برای مشخص کردن رابطه کلمات مرتبطی به کار می‌رود که رابطه آن‌ها نه از جنس سلسله مرتبی است و نه از جنس معادل بودن. این رابطه را می‌توان این گونه مشخص کرد که اگر فردی به دنبال جستجوی مورد «الف» باشد، اگر این فرد در پی آن تمایل به جستجو برای مورد «ب» نیز داشته باشد، این دو مورد با یکدیگر رابطه انجمنی دارند.

4-2. الگوهای زبان

الگوهای زبانی یا مدل‌های زبانی، برای بیان خصوصیات آماری زبان به کار می‌روند. در این مدل‌های آماری احتمال الگوهای مختلف از ترتیب قرارگیری از کلمات مشخص می‌شوند. با استفاده از این الگوها می‌توان کلمات بعدی را پیش‌بینی کرد و الگوهای غلط و یا نامصطلح را تشخیص داد.

محاسبه و تخمین احتمالات الگوهای مختلف در پیکره‌های متنی امری دشوار، پیچیده و بسیار پرهزینه است، جملات و عبارات در پیکره‌ها می‌توانند بسیار زیاد باشند و یا طول زیادی داشته باشند که تعداد الگوها را افزایش می‌دهد همچنین محاسبه احتمال را بسیار پرهزینه می‌سازد. این امر از طرفی ممکن است موجب شود که بسیاری از الگوهای زبان در پیکره‌ی مورد استفاده مشاهده نشوند، چون طول زیاد عبارات جداسازی و قطعه‌سازی را مشکل ساخته و یا از اندازه قالب مورد نظر فراتر خواهند رفت.

---

[1] Hierarchical

[2] Hypernym, Broader Term

[3] Hyponym, Narrower Term

الگوهای زبانی در حقیقت منجر به استخراج هم‌نشینی‌های رایج از کلمات در زبان می‌شود که احتمال هر هم‌نشینی و یا الگو میزان رایج بودن آن را می‌رسانند.

از مدل‌های بسیار مطرح در تحلیل الگوهای زبان می‌توان از مدل چند-وزنی[۱] نام برد. مدل چند-تایی به محاسبه احتمال مشاهده الگوی هم‌نشینی n کلمه در زبان می‌پردازد که این احتمال از طریق محاسبه خصوصیت مرتبه nام مارکوف به دست می‌آید.

2-5. پیکره‌های برچسب داده‌ای

برچسب‌گذاری ادات سخن[۲] عمل انتساب برچسب‌های واژگانی به کلمات و نشانه‌های تشکیل‌دهنده یک متن است، به صورتی که این برچسب‌ها نشان‌دهنده نقش کلمات و نشانه‌ها در جمله باشند. درصد بالایی از کلمات از نقطه‌نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوت دارند. بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه[۳] مورد نظر است. برچسب‌گذاری واژگانی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می‌باشد.

برچسب‌گذاری پیکره‌ها به عنوان یکی از مهم‌ترین حاشیه‌نویسی‌ها برای پیکره‌ها مطرح است. برچسب‌گذاری کاربردهای فراوانی در استخراج الگوهای اجزای واژگانی کلام در زبان دارند، به عبارت دیگر دستورات و قوانین زبان را می‌توان از طریق برچسب‌گذاری به مدلی آماری تبدیل نمود که قابل استفاده در زبان‌شناسی رایانه‌ای هستند. با استفاده از چنین مدل‌هایی می‌توان از کلمات و عبارات رفع ابهام کرد، عبارات نادرست دستوری را تشخیص داد، همچنین می‌توان از آن به عنوان مقدمه‌ای جهت استفاده از لایه‌های معنایی زبان یاد کرد.

---

[۱] N-gram

[۲] Part-Of-Speech tagging

[۳] Context

6-2. پیکره‌های تخصصی

پیکره‌های تخصصی در مقایسه با پیکره‌های عمومی، مطالبی در زمینه‌ی خاص را پوشش می‌دهند. این پیکره‌ها جهت تولید هستان‌شناسی‌های تخصصی و یا اصطلاح‌نامه‌ها در زمینه‌های خاص کاربرد دارند. تولید این پیکره‌ها نیز همانند پیکره‌های عمومی با چالش‌های مشابهی مواجه است به علاوه پالایش کلمات پرکاربرد عمومی برای بهبود کیفیت تحلیل‌ها در زمینه‌ی مورد نظر از چالش‌های دیگر تولید پیکره‌های تخصصی است.

3. ابزارهای زیرساختی در حوزه خط و زبان فارسی

همان‌گونه که پیش‌تر گذشت، برای تولید دادگان زیرساختی در حوزه خط و زبان فارسی نیازمند به ابزارهایی هستیم که متاسفانه این ابزارها، آن گونه که باید، برای زبان فارسی موجود نیستند. حوزه خط و زبان فارسی در این سال‌ها دچار سطحی‌نگری شده است، زبان‌شناسی رایانه‌ای آن گونه که شایسته است مورد توجه قرار نگرفته و همچنین عدم تبیین قوانین لازم و استانداردهای مناسب توسط نهادهای مسئول، که مطابق با پیشرفت فن‌آوری و نیاز و کاربرد روز به پیش رفته باشند، موجب برخوردهای سلیقه‌ای و بعضا بدسلیقگی در این حوزه شده است.

این بخش به معرفی ابزارهای زیرساختی مورد نیاز در حوزه خط و زبان فارسی می‌پردازد. این ابزارها شامل: 1) ابزارهای قطعه‌بندی متن، 2) بازارهای برچسب‌گذاری ادات سخن، 3) ابزارهای ابهام‌زدایی از نقش کلمه، 4) ابزارهای ابهام‌زدایی نحوی، و 5) ابزارهای هنجارسازی هستند و در ادامه به بیان مختصری در خصوص کاربردها و نیازمندی‌های هر یک خواهیم پرداخت.

3-1. ابزارهای هنجارسازی و پیش‌پردازش

هنجارسازی به عنوان پیش‌پردازش نسبت به پردازش‌ها و تحلیل‌های اصلی در زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی از جایگاه بسیار مهمی برخوردار است. در حوزه خط و زبان فارسی این پیش‌پردازش و هنجارسازی در حوزه متن مطرح می‌شود. عمده فعالیت‌ها هنجارسازی به پالایش متن، حذف نویسه‌های اضافه، هنجارسازی نویسه‌ها و از این دست خواهد بود. با توجه به سطوح مختلف زبان‌شناسی، هنجارسازی می‌تواند از سطح هنجارسازی ساختاری و حذف و تبدیل نویسه‌ها تا سطح ریخت‌شناسی و اصلاح فاصله‌گذاری و ریشه‌یابی و حتی تا سطح اصلاح املایی نیز مطرح گردد.

از عمده چالش‌ها در پیکره‌های متنی فارسی که نیاز به هنجارسازی دارند می‌توان به موارد زیر اشاره کرد:

1) وجود نویسه‌های هم‌شکل متعدد با کدهای متفاوت که کاربران مختلف ممکن است به دل‌خواه از گونه‌های متفاوت آن‌ها استفاده کنند که علی‌رغم هم شکل بودن، عملا در زبان‌شناسی رایانه‌ای متفاوت هستند.

2) وجود فاصله‌های اضافی در متن که عمل استخراج کلمات و عبارات را مختل می‌سازد

3) وجود نویسه «شبه-فاصله» که به صورت مجرد نمود ظاهری خاصی ندارد و مشکلات بسیاری را به وجود می‌آورد، از آن جمله می‌توان به موارد زیر اشاره کرد:

أ. وجود شبه-فاصله‌های متعدد کنار هم که چون نمود ظاهری خاصی ندارد به صورت چشمی قابل تشخیص نیست اما زبان‌شناسی رایانه‌ای و تحلیل پیکره را با مشکل مواجه می‌سازد.

ب. وجود شبه-فاصله پس از نویسه‌هایی مانند «ر» و «ذ» که فرم چسبان ندارند. در این حالت شبه-فاصله فاقد کارایی است و نیازی به وجود آن نیست، اما وجود آن در تحلیل پیکره‌های متنی مشکل ایجاد می‌کند.

ج. وجود شبه-فاصله پیش و پس از کلمات، در این حالات نیز وجود شبه-فاصله فاقد کارایی است ولی در تحلیل‌های رایانه‌ای پیکره‌ها، این نویسه‌ها جزئی از کلمه محسوب شده و ایجاد اختلال در پردازش می‌کنند.

با گذر از سطح ساختار پیکره‌ها، در سطحی دگر می‌توان عملیات زیر را نیز به عنوان پیش‌پردازش‌های مورد نیاز در پردازش زبان طبیعی مطرح نمود که البته خود نیازمند زیرابزارهایی ویژه هستند که بعضا نیازمند به دادگان زیرساختی زبان بوده و می‌توانند در مرحله پالایش و بهینه‌سازی دادگان اولیه مورد استفاده قرار گیرند:

1) ریشه‌یابی

2) حذف وندها

3) اصلاح فاصله‌گذاری‌ها و قوانین جدا و پیوسته نویسی که ممکن است بـه علـت: 1) اشـتباه در کاربرد فاصله و شبه-فاصله به جای هم، 2) عدم استفاده از شبه-فاصله در کلمات چند جزئـی، 3) استفاده بی مورد از شبه-فاصله، و 4) استفاده بی مورد از فاصله، رخ داده باشند.

4) اصلاح املای لغات، که مشـکلات ناشـی از ایـن امـر خصوصـا در پیکـره‌هـایی کـه بـا کمـک نوسه‌خوان‌های نوری رایانه‌ای شده‌اند، به شدت محسوس است.

3-2. ابزارهای قطعه‌بندی متن و استخراج الگو

قطعه‌بندی متن به تقسیم متن و یا پیکره‌های متنی به واحدهای معنایی مانند کلمات، جملات، عبارات و یا عناوین اطلاق می‌شود. قطعه‌بندی متن آن گونه که در ذهن تصور می‌شود امری آسان نیست، زیرا در برخی زبان‌ها نشانگرهای کران کلمه به صورت صریح همانند فاصله در زبان انگلیسی، موجود نیست و به عنوان مثال در رسم‌الخط عربی و فارسی، فرم‌های پایانی و میانی از حروف می‌توانند علاوه بر فاصله صریح نشانگر کران کلمات باشند و نیز می‌توانند در مواردی نشانگر کران کلمه نباشند که موجب وجود ابهام می‌گردد. از طرفی در برخی موارد فاصله جزئی از خود کلمه است و نیز چالش‌هایی که در مواجه با شبه-فاصله در زبان فارسی وجود دارد، قطعه‌بندی متن را کاری بسیار مشکل‌تر از موارد مشابه برای زبان‌های دیگر می‌سازد.

3-2-1. قطعه‌بندی کلمات

استخراج کلمات شامل تجزیه‌ی یک متن به کلمات تشکیل دهنده آن است. همان طور که پـیش از ایـن گذشت، تشخیص کران کلمه چالشی در تشخیص و استخراج کلمات زبان از پیکره‌هـای متنـی اسـت. در شیوه‌های نوین نگارش درست، جداکننده کلمات معمولا نویسه «فاصله» است. اما مشکل جایی است کـه این نویسه جزئی از کلمه است. مثلا «از ما بهتران» بیانگر یک کلمه است که در میان خود شامل نویسـه جدا کننده است. مواردی از قبیـل: 1) افعـال مرکـب، 2) افعـال مجهـول، 3) مرکـب‌هـای اتبـاعی، و 4) ترکیبات اضافه چند جزئی، مواردی هستند که با فرض رعایت شـیوه نگـارش صـحیح و درسـت نویسـی مشکلاتی را در زمینه تشخیص کران کلمه در زبان‌شناسی رایانه‌ای بـه وجـود مـی‌آورنـد. ایـن مشـکلات

هنگامی نمود بیشتری پیدا می‌کند که خط فارسی و دستور زبان و شیوه نگارش صحیح فارسی نیز دارای ابهامات و پیچیدگی‌های خاص دیگری نیز هست. مواردی از قبیل: 1) وجود شبه[1]-فاصله به عنوان یکی از نویسه‌های زبان، 2) وجود ابهام در دستور خط زبان فارسی برای قوانین پیوسته و جدانویسی، 3) بازگذاشتن دست نویسندگان در فاصله‌گذاری میان کلمات، 4) عدم وجود دستورالعمل قطعی برای استفاده از شبه-فاصله، 5) عدم وجود قواعدی ثابت برای فاصله‌گذاری ترکیبات؛ استفاده از دستورالعمل مبتنی بر لغت (مانند تک‌هجایی بودن، بسیط‌گونه بودن)، 6) عدم وجود رویکرد زبان‌شناسی رایانه‌ای در قوانین دستور خط زبان فارسی، 7) تفاوت زیاد میان زبان محاوره و نوشتار، 8) قابلیت زایائی زبان در خصوص ایجاد کلمات جدید، مواردی هستند که موجب چندگانه نویسی کلمات در زبان فارسی می‌شوند و تشخیص و استخراج کلمات از پیکره‌های متنی در زبان فارسی را دشوار می‌سازند.

عمل «جداسازی کلمات» نیز عملی است که برای جداسازی کلماتی که به یکدیگر متصل شده‌اند به مار می‌رود که گونه‌ای پیچیده‌تر از قطعه‌بندی کلمات را پوشش می‌دهد و برای پیکره‌ها یا متونی که فاقد فاصله و دیگر جداکننده‌های کلمات نیستند کاربرد دارد. از جداسازی کلمات با نام خطِ تیره گذاری[2] نیز نام برده می‌شود.

3-2-2. قطعه‌بندی جملات

قطعه‌بندی متن مساله تجزیه‌ی متن و یا پیکره‌های متنی به جملات تشکیل دهنده آن است. استفاده از علائم نشانه‌گذاری و دیگر نویسه‌های پایان دهنده در متون رایانه‌ای از روش‌ها تشخیص کران جملات است. اما این مساله نیز پیچیده‌تر از آن است که تصور می‌شود. به عنوان مثال استفاده از نویسه پایان دهنده «.» در مخففها و یا به صورت «...» نشانگر پایان جمله نیست. و یا حتی در جمله پیشین همین متن، «.» به عنوان جزئی از جمله مطرح شده و جمله به صحبت در خصوص آن می‌پردازد و وجود نقطه نشانگر پایان جمله نبوده است. از طرف دیگر هیچ تضمینی برای نشانه‌گذاری صحیح جملات و یا حتی وجود نشانه‌گذاری در جملات و پیکره‌های متنی نیست که این امر مساله قطعه‌بندی جملات را بسیار پیچیده می‌سازد.

---

[1] Pseudo-space

[2] Hyphenation

از دیگر روش‌های قطعه‌بندی جملات استفاده از قوانین و دستورات زبان است که به عنوان مثال در زبان فارسی معمولا جملات به یک فعل ختم می‌شوند اما صرفا تشخیص ادات سخن در متون و تشخیص نقش یک کلمه به عنوان فعل دلالت بر تشخیص پایان جمله نخواهد داشت. به عنوان نمونه جمله‌ی «او گفت: آری خواهم آمد.» با این روش دو جمله تشخیص داده خواهد شد در حالی که تنها یک جمله است. از طرفی این روش کلیه‌ی چالش‌های تشخیص فعل خصوصا تشخیص افعال مرکب را نیز دارا است.

3-2-3. قطعه‌بندی دیگر اجزاء

متون و پیکره‌ها می‌توانند به پاراگراف، عناوین و مباحث نیز شکسته شوند که این گونه قطعه‌بندی‌ها نیز در تحلیل پیکره‌های متنی مطرح هستند. یک پیکره متنی ممکن است از چندین عنوان تشکیل شده باشد. تشخیص کران عناوین مختلف موجود در متن نیز می‌تواند توسط تحلیل عنوان بخش‌ها و پاراگراف‌ها انجام پذیرد که البته تشخیص عنوان خود فعالیتی پیچیده و مجزا است و جزء مسائل طبقه‌بندی متن قرار می‌گیرد [4].

3-3. ابزارهای برچسب‌گذاری ادات سخن

برچسب‌گذاری ادات سخن یا برچسب‌گذاری دستوری عملی است که در آن کلمات با یکی از ادات سخن، بر اساس تعریف خود کلمه و نیز زمینه کاربرد آن در جمله جاری، برچسب‌گذاری می‌شوند. از ساده‌ترین ادات سخن می‌توان به اسم، فعل، صفت و قید اشاره نمود.

برچسب‌گذاری خودکار در زبان‌شناسی رایانه‌ای بر اساس تعریف مجزای کلمات و نیز بر اساس ادات پنهان در سخن مشخص می‌شوند که مورد اخیر امر برچسب‌گذاری را بسیار پیچیده می‌سازد. درصد بالایی از کلمات از نقطه‌نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه‌های مختلف برچسب‌های واژگانی متفاوت دارند. بنابراین برچسب‌گذاری واژگانی عمل ابهام‌زدایی از برچسب‌ها با توجه به زمینه مورد نظر است. برچسب‌گذاری واژگانی عملی اساسی برای بسیاری از حوزه‌های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می‌باشد.

نکته‌ای دیگر که در برچسب‌گذاری پیکره‌های متنی می‌باید مورد توجه قرار گیرد تعیین لیستی سلسله مراتبی از ادات سخن و نقش کلمات در زبان فارسی است که بتوان برای کاربردهای متفاوت سطوح مورد

نظر از لیست مورد بحث را انتخاب نمود و همچنین با توجه به سلسله مراتبی بودن این لیست سطوح مختلف از لیست قابلیت سازگاری بالایی با یکدیگر دارند و می‌توانند به یکدیگر تبدیل شوند.

با توجه به کاربردهای مطرح شده برای برچسب‌گذاری و پیکره‌های برچسب داده‌ای، مجموعه ابزارهای برچسب‌گذاری ادات سخن یکی از ابزارهای زیرساختی مهم در بسیاری از پروژه‌های پردازش زبان طبیعی در حوزه خط و زبان فارسی خصوصا در لایه‌های نحوی و معنایی خواهند بود.

3-4. ابزارهای ابهام‌زدایی

3-4-1. ابهام‌زدایی از مفهوم کلمه

ابهام‌زدایی از مفهوم کلمه به فرایند تشخیص مفهوم کلمه در جمله جاری اطلاق می‌شود که در آن از میان مفاهیم مختلفی که یک کلمه به صورت مجرد می‌تواند بیان کند، یکی از آن‌ها انتخاب می‌شوند.

به عنوان مثال می‌توان به مفاهیم مختلف کلمه «شیر» در جملات زیر اشاره کرد:

1) نوشیدن شیر برای سلامتی مفید است.
2) شیر نشان پرچم ایران بوده است.
3) قیمت شیر آلات ساختمانی در سال جاری افزایش چشم‌گیری داشته است.

تشخیص مفهوم مورد نظر از هر کلمه «شیر» در جملات فوق برای انسان کاملا واضح است اما تشخیص این امر توسط کامپیوتر امری بسیار دشوار و پیچیده است.

تولید ابزارهایی جهت تشخیص مفهوم مورد نظر از کلمه در جملات، از ابزارهای پایه در کاربردهای زبان‌شناسی رایانه‌ای و پردازش طبیعی زبان خصوصا در کاربردهایی در لایه نحو و معناست.

3-4-2. ابهام‌زدایی نحوی

ابهام نحوی در یک جمله شرایطی است که در آن یک جمله به طرق گوناگون تفسیر شود و یا به گونه‌ای تفسیر شود که بیش از یک معنی بدهد. ابهام نحوی در قیاس با ابهام معنایی مستقیما متاثر از وجود

چند ادات سخن برای یک کلمه و یا وجود چند معنی برای آن نیست بلکه ناشی از ارتباط میان کلمات و عبارات درون جمله است [5].

به عنوان مثال جملات زیر دارای ابهام نحوی هستند:

1) «عیسی، پسر مریم، برگزیده‌ی خداوند.» در این جمله مشخص نیست که عیسی برگزیده‌ی خداوند است یا مریم برگزیده‌ی خداوند است.

2) «برخی از مردان و زنان» در این جمله مشخص نیست که برخی از مردان و تمامی زنان مورد نظر است و یا برخی از مردان و برخی از زنان منظور جمله بوده است.

3) «او به مسئول انتخابات گفت که نمی‌تواند تقلب کند» در این جمله مشخص نیست که او نمی‌تواند تقلب کند و یا مسئول انتخابات نباید تقلب کند.

4) «او تو را بیشتر از من دوست دارد» در این جمله مشخص نیست که میزان علاقه‌ی او به تو بیشتر از میزان علاقه‌اش به من است و یا او تو را بیش‌تر از آنچه که من به تو علاقه دارم، دوست داد.

با توجه به این که این نوع از ابهام در زبان فارسی نسبت به زبان انگلیسی کمتر مشاهده می‌شود اما تولید ابزارهایی جهت رفع این ابهام‌ها با استفاده از کاربرد و هم‌آیی جملات و زمینه امری است نه به اهمیت دیگر موارد، اما همچنان مهم.

4. سازوکار تولید دادگان زیرساختی

4-1. مرحله اول

شکل 1، نمای کلی و چرخه تولید دادگان زیرساختی در حوزه خط و زبان فارسی را نشان می‌دهد. این چرخه با تولید «پیکره متنی عمومی» زبان آغاز می‌شود. نکاتی که در تولید پیکره‌های متنی می‌باید مورد توجه قرار گیرند شامل موارد زیر هستند:

1) تعیین خصوصیات و معیارهای یک پیکره
2) تعیین حوزه پیکره
3) تعیین سهم حضور هر یک از گونه‌های معیار، فوق معیار و زیر معیار از زبان در متن پیکره
4) صحت متون پیکره ار نظر قوانین زبانی
5) صحت متون پیکره ار نظر قوانین نوشتاری

پس از تولید «پیکره متنی عمومی» در مرحله اول، یک مرحله هنجارسازی در مرحله ساختار و ریخت‌شناسی بر روی آن صورت خواهد گرفت و با استفاده از ابزار قطعه‌ساز متن، اقدام با استخراج «واژگان زبان» می‌شود. در این میان می‌توان هم‌زمان اقدام به استخراج «لیست بسامدی» از واژگان نیز نمود. در مرحله بعد، با استفاده از ابزارهای قطعه‌سازی و برچسب‌گذاری اقدام به استخراج «الگوهای زبان» خواهد شد. در این مرحله نکته‌ای که باید مورد توجه قرار گیرد تعداد کلمات موجود در الگو است. مطالعات خاصی در این زمینه صورت نگرفته اما تجربیات موجود در زمینه‌ی ترجمه ماشینی در زبان انگلیسی این‌گونه بوده که محاسبه احتمال رخداد الگوها برای دنباله‌های 2 تا 5 تایی از کلمات مناسب است و بیش از آن تقریبا کارایی نخواهد داشت.

هم‌زمان و یا پس از تولید و محاسبه الگوهای زبانی توصیه به تولید «پیکره‌های تخصص» است که در مرحله بعد با استفاده از آن‌ها اقدام به تولید «اصطلاح‌نامه» شود. با استفاده از اطلاعات هستان‌شناسی، و واژگان زبان و پیکره‌های متنی و تخصصی اقدام به تولید اصطلاح‌نامه‌ها در سطح مقدماتی خواهد شد.

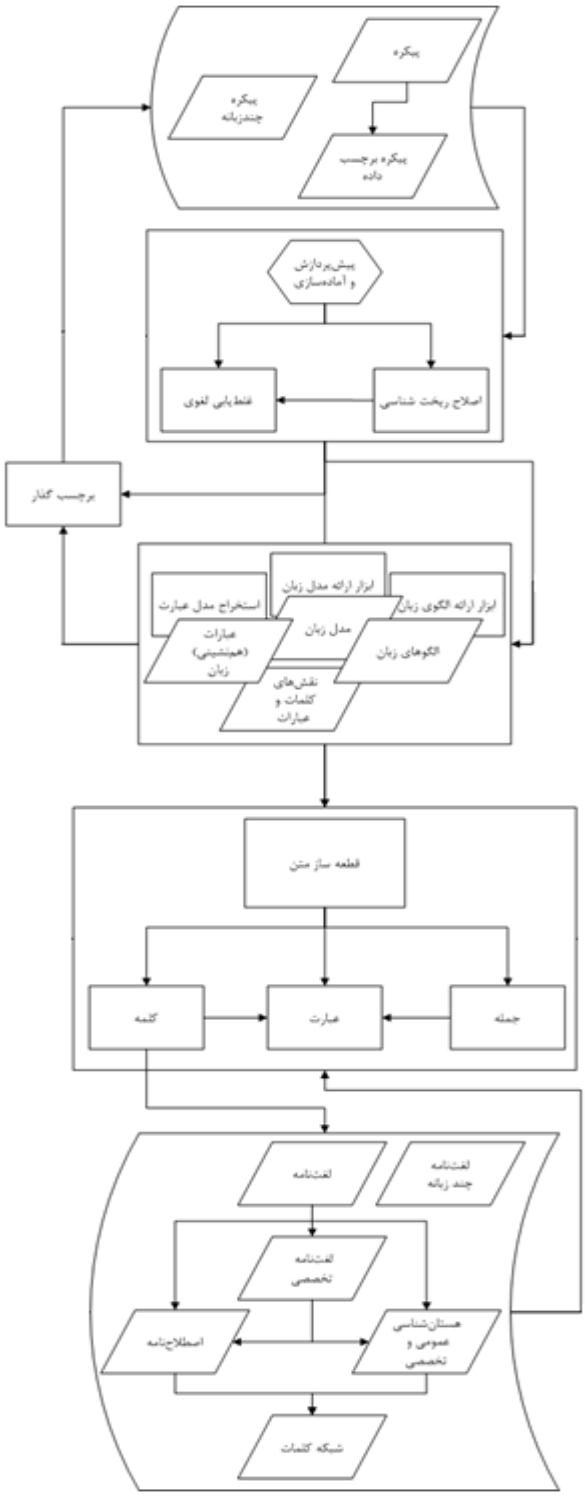
4-2. مرحله دوم

در مرحله دوم و با استفاده از بازخورد الگوهای زبانی مرحله قبل، پیکره‌های متنی جدیدی ایجاد خواهد شد. با استفاده از این پیکره‌ها و واژگان مرحله اول که کارکرد ابزارهای قطعه‌ساز را بهبود می‌بخشد، واژگان مرحله دوم ایجاد می‌شوند و مجددا از این پیکره‌ها الگوهای زبانی استخراج می‌گردد. با استفاده از اصطلاح‌نامه مرحله اول و الگوهای زبانی مرحله دوم، پیکره تخصصی مرحله دوم ایجاد شده و بر اساس واژگان سطح دوم و پیکره‌های عمومی و تخصصی مرحله دوم، اصطلاح‌نامه مرحله دوم ایجاد می‌شود.

4-3. مرحله سوم

در مرحله سوم نیز با چرخه‌ای همانند مرحله دوم، پیکره‌های مرحله سوم با استفاده از واژگان مرحله دوم، الگوهای مرحله دوم و پیکره‌های مرحله دوم ایجاد خواهند شد. سپس واژگان مرحله سوم با استفاده از پیکره‌های مرحله سوم، واژگان مرحله دوم و الگوهای مرحله دوم ایجاد خواهد شد. با استفاده از پیکره‌های مرحله سوم می‌توان الگوهای زبانی مرحله سوم را ایجاد نمود و سپس با استفاده از اصطلاح‌نامه مرحله دوم و الگوهای زبانی مرحله سوم، پیکره تخصصی مرحله سوم ایجاد می‌شود. در ادامه با استفاده از واژگان سطح سوم و پیکره‌های عمومی و تخصصی مرحله سوم، اصطلاح‌نامه مرحله سوم ایجاد می‌شود.

در نهایت با استفاده از الگوهای مرحله سوم و ابزارهای برچسب‌گذاری، به تولید پیکره‌های برچسب داده‌ای از پیکره‌های مرحله سوم اقدام می‌شود.



شکل 1. نمای کلی از چرخه تولید دادگان زیرساختی کاربری خط و زبان فارسی در محیط رایانه‌ای

5. نتیجه‌گیری

پردازش زبان طبیعی در سطوح شش گانه آن شامل: 1) آواشناسی و صدا شناسی ، 2) ریخت‌شناسی، 3) نحو، 4) معناشناسی، 5) عمل‌گرایی، و 6) مباحثه، و کاربردهای اصلی آن شامل: 1) خلاصه‌سازی خودکار، 2) کمک به خواندن زبان‌های طبیعی دیگر، 3) کمک به نوشتن به زبان‌های طبیعی دیگر، 4) استخراج اطلاعات، 5) بازیابی اطلاعات، ۶) ترجمه ماشینی، 7) تشخیص واحدهای اسمی، 8) تولید زبان طبیعی، 9) فهم زبان طبیعی، 10) نویسه‌خوان نوری، 11) تحلیل مرجع‌دارها، 12) سیستم سوال، پاسخ، 13) تشخیص گفتار، 14) مبدل متن به گفتار، 15) نظام‌های مکالمه گفتاری، 16) ساده‌سازی متن، و 17) تایید متن، نیازمند ابزارها و دادگان زیرساختی است. از طرفی در حوزه خط و زبان، پیکره‌های متنی به عنوان نمادی از زبان هستند که می‌باید با تحلیل آن‌ها به استخراج اجزا، قواعد و ساز و کار زبان پی برد. دادگان مهم و زیرساختی حاصل از تحلیل پیکره‌های متنی در پردازش زبان طبیعی و حوزه خط و زبان را می‌توان به موارد زیر تقسیم نمود: 1) پیکره متنی، 2) واژگان زبان، 3) اصطلاح‌نامه، 4) الگوهای زبان، 5) پیکره‌های برچسب داده‌ای، 6) پیکره‌های تخصصی.

پروژه‌های زیرساختی در حوزه خط و زبان فارسی را می‌توان پروژه‌هایی برای ایجاد دادگان زیرساختی در حوزه خط و زبان فارسی در نظر گرفت. از طرفی ابزارهایی نیز جهت حل مسائل زیرساختی مطرح شده در حوزه خط و زبان نیز نیاز هستند تا بتوان اقدام به تحلیل پیکره‌های متنی زبان جهت تولید دادگان زیرساختی نمود. بنابراین پروژه‌های زیرساختی خط و زبان فارسی تلفیقی از ابزارها و دادگان زیرساختی زبان خواهد بود. این ابزارها شامل: 1) ابزارهای قطعه‌بندی متن، 2) برچسب‌گذاری ادات سخن، 3) ابزارهای ابهام‌زدایی، 4) ابزارهای هنجارسازی، تقریبا در تمامی کاربردهای پردازش زبان طبیعی در حوزه خط و زبان مطرح هستند و به عنوان مسائل زیرساختی خط و زبان می‌باید مورد بررسی قرار گیرند. زبان فارسی به عنوان یکی از زبان‌های طبیعی نیز از این قاعده مستثنی نیست.

در این مستند به بررسی چالش‌ها در تولید و ساخت دادگان و ابزارهای زیرساختی و نیز اولویت و سازوکار و مراحل تولید دادگان زیرساختی خط و زبان فارسی پرداختیم.

6. مراجع

[1] J. Hutchins, “Retrospect and prospect in computer-based translation,” in Proceedings of MT Summit VII, 1999, pp. 30–34.

[2] C. D. Manning, and H. Schtze, *Foundations of statistical natural language processing*: MIT Press, 1999.

[3] G. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*: addison-wesley press, 1949.

[4] F. Choi, “Advances in domain independent linear text segmentation,” in Proceedings of the first conference on North American chapter of the Association for Computational Linguistics, 2000, pp. 26-33.

[5] L. Allen, and M. Caldwell, “Modern Logic and Judicial Decision Making: A Sketch of One View,” *Law and Contemporary Problems*, vol. 28, no. 1, pp. 213-270, 1963.