Biannual scientific-research journal of language and linguistics
Spring & Summer 2024, 20 (39), 155-192
https://lsi-linguistics.ihcs.ac.ir/
DOI: 10.30465/LSI.2025.47627.172

Original Research
Received: 05/Dec/2023 *** Accepted: 17/Nov/2024
ISSN: 2423-3847

# Analysis of the Annotated Corpus Developed from Textbooks of Grades 1 to 6

**Masood Ghayoomi**[1] Associate Professor. Faculty of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran

**Elham Salehi**[2] PhD Student. Faculty of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran

**Azam Alijani**[3] PhD Student. Institute for Cognitive Science Studies, Tehran, Iran

## Abstract

In the comprehensive scientific roadmap of our country, Iran, the promotion of Persian language as a science language among other international science languages is taken into consideration. One of the ways to reach this goal is suggested as expanding the usage of the Persian language. To achieve the goal, which can be raised in the field of Persian language policy making, it is necessary to understand more about the linguistic content properties and the basic concepts that are taught in the textbooks to students. The description of these features can be considered when preparing the language content. In this research, a corpus of textbooks from grades 1 to 6 (the primary school period) is developed that contains around 208,000 words and then it is annotated. These courses include Farsi, Experimental Sciences, Social Studies and Heavenly Gifts. All the sentences of different courses are written in plain text files, separated by grade and course, and after normalization in the pre-processing process, they are annotated automatically at four levels: broad transliteration, lemmatization, part-of-speech and syntactic constituency parsing. The results of this research can

1. M.Ghayoomi@ihcs.ac.ir (Corresponding Author)
3. e.salehi@ihcs.ac.ir
3. azamalijani96@gmail.com
**How to cite:** Ghayoomi, M., Salehi, E., & Alijani, A. (2024). An Analysis of the Annotated Corpus Developed from Textbooks of Grades 1 to 6. *Language and Linguistics*, 20(39), 155-192. doi: 10.30465/LSI.2025.47627.172

help to know more about the content of textbooks and to be useful in the fields of education and policy making in language planning.

**Keywords**: Corpus, Corpus Linguistics, Education, Policy making in education.

## 1. Introduction

In the comprehensive scientific roadmap of our country, Iran, the promotion of Persian language as a science language among other international science languages is taken into consideration. One of the ways to reach this goal is suggested as expanding the usage of the Persian language. To achieve the goal, which can also be raised in the field of language policy making, it is necessary to understand more about the basic linguistic properties taught in the textbooks to students. Moreover, the description of these features can be considered when preparing the language content.

### *Research Question(s)*

**1.** What are the basic linguistic properties taught in the textbooks to students?
**2.** Which aspects of these linguistic properties can be considered when preparing the language content?

## 2. Literature Review

This paper aims at describing the linguistic properties of the primary school textbooks. Therefore, the literature goes in different directions. We limit the literature to Persian corpus development and its annotation at phonological, morphological, and syntactic levels. To the best of our knowledge, Assi (1997) developed the primary Persian corpus, called “Persian Linguistic DataBase”. It contained contemporary Persian and Part-of-Speech (PoS) tags (Assi and Hajiabdolhoseini, 2000). Afterwards, Ghayoomi (2004), Daroodi et al. (2004), Bijankhan (2004), AleAhmad et al. (2009), and Bijankhan et al. (2011), Eghbalzade et al. (2012), Sabouri (2022) and Ghayoomi (2022) developed several corpora for Persian mostly from online archive of news pages. The two last corpora contained more than 14 billion of words, separately. It is possible to annotate a corpus with linguistic information, including phonological label (Eslami et al. 2004), lemmatization (Tashakori et l., 2002), PoS tag (Assi and Hajiabdolhoseini, 2000), and syntactic parse tree for both constituency tree structure (Ghayoomi, 2012) and dependency structure (Seraji et al., 2012). Among the available corpora for Persian, the Bijankhan Corpus (Bijankhan, 2004) contained all of these annotation levels at once. The annotations have been done by different researchers.

## 3. Method

In the school structuralism, de Saussure (1916) believed that language contains two levels: one is “form”, and the other one is “meaning”. The “form” which is represented by the orthography can be processed. The syntagmatic relations of the word forms create sentences. In 1960s, the necessity of using linguistic corpora in language studies became aware to anyone. Natural language processing methods paved the way to process a language automatically. This caused to process a large portion of data more quickly than doing it manually.

To describe the linguistic properties of Persian towards language planning, we developed a corpus of textbooks from grades 1 to 6 (the primary school period) that contained around 208,000 words and it was annotated automatically. The courses included Farsi, Experimental Sciences, Social Studies, and Heavenly Gifts. All the sentences of different courses are written in plain text files, separated by the grade and the course, and after normalization in the pre-processing process, they are annotated automatically at four levels: broad transliteration, lemmatization, part-of-speech and syntactic constituency parsing.

## 4. Results

The results of this research can help to know more about the properties of the textbooks’ content and to be useful in the fields of education and policy making in language planning. We annotated the corpus phonologically based on the lexicon developed by Eslami et al. (2004) which contained a list of words and their relevant pronunciation. Then, we extracted the syllable patterns according to the pronunciation. According to the statistical information in Table 1, Grades 4 to 6 have almost equal distribution of words’ syllables. The syllable patterns in all grades have almost equal distribution. Moreover, the syllable patterns of the courses Farsi and Experimental Science contain %28 of the pattern distribution; while the courses Social Studies and Heavenly Gifts contain 24 and 20 percents of the pattern distribution, respectively.

**Table 1**

*Statistical Information of words’ syllables and syllable patterns*

| Grades | # of words’ syllables | # of CVCC pattern | # of CVC pattern | # of CV pattern |
|---|---|---|---|---|
| Grade 1 | 0.04 | 4.40 | 26.88 | 67.62 |
| Grade 2 | 0.11 | 5.04 | 27.71 | 65.93 |
| Grade 3 | 0.14 | 4.97 | 27.98 | 66.11 |
| Grade 4 | 0.22 | 4.99 | 28.65 | 65.19 |
| Grade 5 | 0.24 | 4.51 | 29.29 | 64.99 |
| Grade 6 | 0.24 | 4.45 | 29.44 | 65.05 |
| All | 1 | 4.72 | 28.72 | 65.43 |

The corpus has been PoS tagged automatically by using the Marmot toolkit (Müller et al., 2013) trained with the Bijankhan Corpus (2004). According to reported results in Table 2, the category "noun" contains %34 of the lexicon and it is the most frequent word in the course Experimental Sciences. The category "verb" stays in the second stage with %14.65 distribution. This category has been frequently used in the Farsi course, especially in Grade 5. It should be mentioned that the number of nouns is 2.5 times higher than verbs and this difference exists almost in all grades.

**Table 2**
*Statistical Information of words' PoS tags*

| Grades | Noun | Verb | Punctuation | Preposition | Coordination | Adjective | Pronoun | Adverb | Post-position " ra" | Determiner | Number | Classifier | Interjection |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 1 | 1.5 | 0.63 | 0.44 | 0.44 | 0.25 | 0.24 | 0.16 | 0.12 | 0.13 | 0.16 | 0.38 | 0.01 | 0.00 |
| Grade 2 | 3.90 | 1.86 | 1.69 | 1.14 | 0.88 | 0.66 | 0.51 | 0.37 | 0.37 | 0.31 | 0.25 | 0.04 | 0.01 |
| Grade 3 | 4.76 | 2.30 | 1.75 | 1.51 | 1.11 | 0.89 | 0.66 | 0.52 | 0.45 | 0.42 | 0.20 | 0.02 | 0.01 |
| Grade 4 | 7.19 | 3.25 | 2.46 | 2.31 | 1.61 | 1.45 | 0.88 | 0.73 | 0.65 | 0.59 | 0.41 | 0.02 | 0.01 |
| Grade 5 | 8.48 | 3.40 | 2.63 | 2.66 | 1.98 | 1.59 | 0.99 | 0.65 | 0.74 | 0.66 | 0.46 | 0.03 | 0.01 |
| Grade 6 | 8.18 | 3.22 | 2.44 | 2.63 | 2.24 | 1.59 | 0.93 | 0.67 | 0.52 | 0.66 | 0.39 | 0.03 | 0.01 |
| All | 34.02 | 14.65 | 11.42 | 10.69 | 8.07 | 6.42 | 4.13 | 3.04 | 2.86 | 2.79 | 2.09 | 0.14 | 0.05 |

The PoS tagged words were categorized into two groups, either as functional words or content words. According to reported results in Table 3, %62 of the lexicon of this corpus are content words and the rest, about %38, functional words. The statistical distributions of content words are %38, %16, %14, and %12 for Farsi, Experimental Sciences, Social Studies and Heavenly Gifts, respectively. Grade five contains the highest percentage of content and functional words.

**Table 3**
*Statistical Information of functional and content words*

| Grades | Content Words | Functional Words |
|---|---|---|
| Grade 1 | 2.67 | 1.52 |
| Grade 2 | 7.33 | 4.69 |
| Grade 3 | 9.09 | 5.46 |
| Grade 4 | 13.54 | 8.06 |
| Grade 5 | 15.15 | 9.16 |
| Grade 6 | 14.63 | 8.88 |
| All | 62.42 | 3758 |

The sentences are constituently parsed and the tree depth of the sentences, the number of internal nodes between the leaf node and the root node of the tree, was computed. According to reported results, in Table 4, in average, the sentences contain 29 nodes in the parse tree analysis of each sentence in the corpus. Experimental Sciences and Social Studies have the lowest and the highest number of grades in the corpus.

**Table 4**
*Statistical Information of relative average tree depth*

| Grades | Average Tree Depth |
|---|---|
| Grade 1 | 20.78 |
| Grade 2 | 24.77 |
| Grade 3 | 25.31 |
| Grade 4 | 42.95 |
| Grade 5 | 31.44 |
| Grade 6 | 34.09 |
| All | 29.40 |

In addition, we extracted 50 most frequent words from the corpus and studied which words are used in the different grades. It was found out that some words appear in the corpus from a specific grade, such as the word /?agar/ 'if' that is used in conditional sentences from Grade three. We further extracted 1000 frequent words from the corpus. It was found out that the lexicon of the grades four to six is almost the same and the authors of the teaching content in the textbooks put efforts to use the words in different contexts to make them stick in the memory of the students.

## 5. Conclusion

This research aimed at developing a corpus from primary school textbooks to deepen our wisdom on the content of the texts to use the knowledge in language policy making and language teaching to move towards the goal to make Persian as one of the science languages.

Based on the annotations and the extracted statistics from the corpus, we concluded that the general principles at providing the contents is minded but double thinking is required for some issues in detail. Linguistic and lexical complexity of the textbooks in Grade five is high and it can be transited to Grade six. This transition causes students to learn more complex materials by cognitive growth of students. Also, we found that 1000-word forms can be considered as the basic lexicon of grades four to six that the students have to learn.

**Acknowledgments**
This research is funded by the Institute for Cognitive Science Studies for the project number 11243.

مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۹/۰۵ *** تاریخ پذیرش: ۱۴۰۳/۰۸/۲۷

ISSN: ۲۳۲۲-۳۸۴۷

دوفصلنامه علمی- پژوهشی زبان و زبان‌شناسی
دوره ۲۰، شماره ۳۹، بهار و تابستان ۱۴۰۳، ۱۵۵- ۱۹۲
lsi-linguistics.ihcs.ac.ir
DOI: 10.30465/LSI.2025.47627.172

# تحلیلی بر پیکرهٔ برچسب‌گذاری‌شدهٔ حاصل از محتوای کتاب‌های درسی پایه‌های اول تا ششم ابتدایی

**مسعود قیومی** دانشیار، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

**الهام صالحی** دانشجوی دکتری، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

**اعظم علیجانی** دانشجوی دکتری، مؤسسه آموزش عالی علوم شناختی، تهران، ایران

## چکیده

در سند نقشه جامع علمی کشور، به ارتقای جایگاه زبان فارسی در مقام زبان علم درمیان زبان‌های بین‌المللی علمی اشاره شده‌است. یکی‌از راه‌کارهای رسیدن به این هدف کلان، گسترش زبان فارسی ذکر شده‌است. برای رسیدن به این هدف که می‌تواند در حوزهٔ سیاست‌گذاری‌های مربوط به زبان فارسی مطرح گردد، به درک بیشتر از محتوای زبانی نیاز است. از این منظر، توصیف ویژگی‌های کتاب‌های درسی دانش‌آموزان که با مفاهیم پایه آموزش می‌بینند اهمیت زیادی پیدا می‌کند. این ویژگی‌ها می‌تواند هنگام تهیه محتوای زبانی مد نظر قرار گیرد. در این پژوهش، پیکره‌ای از متون درسی کلاس‌های اول تا ششم ابتدایی با حجمی در حدود ۲۰۸هزار واژه تهیه شده و سپس برچسب‌گذاری شده‌است. این دروس شامل فارسی، علوم، مطالعات اجتماعی و هدیه‌های آسمانی است. تمام جملات دروس مختلف به تفکیک پایه و درس، در فایل‌های متنی ساده، حروف‌نگاری شده و پس‌از هنجارسازی در فرایند پیش‌پردازش، در چهار سطح آوایی، بن‌واژه‌ای، مقولهٔ دستوری و تجزیهٔ سازه‌ای نحوی برچسب‌گذاری شده و بررسی شده‌است. نتایج حاصل از این پژوهش می‌تواند به شناخت بیشتر از محتوای کتاب‌های درسی کمک کند و در حوزهٔ آموزش و سیاست‌گذاری در این حوزه مفید باشد.

**کلیدواژه:** پیکره، زبان‌شناسی پیکره‌ای، آموزش، سیاست‌گذاری در آموزش.

**۱- مقدمه**

ارتقای جایگاه زبان فارسی به‌عنوان هویت ملی در مقام زبان علم درمیان زبان‌های بین‌المللی علمی موضوعی است که در سند نقشه جامع علمی کشور مطرح شده‌است و یکی‌از راه‌کارهای رسیدن به این هدف، گسترش زبان فارسی ذکر شده‌است. این رویکرد سبب شده‌است نوعی سیاست‌گذاری زبانی در اسناد بالادستی صورت پذیرد و وظایف متعددی برای دستگاه‌های متولی تعریف گردد. رجبی و احمدوند (۱۴۰۱) به بررسی سیاست‌گذاری زبانی و جایگاه زبان فارسی در سیاست‌های زبانی جمهوری اسلامی ایران پرداخته و به این نتیجه رسیده‌اند که مسائل و مشکلات پیشِ روی زبان فارسی، از جنس مسائل بدخیم بوده و شیوهٔ اجرای سیاست‌های مربوط به زبان فارسی بر رویکردی دستوری و بالا-به-پایین استوار است که همین امر منجر به ناکامی سیاست‌گذاری روی زبان ملی و نیل به اهداف خود شده‌است. برای اجرای درست سیاست‌ها و حل یک مسئله، ابتدا به آگاهی نیاز داریم.

یکی‌از شیوه‌هایی که می‌تواند به شناخت ما از زبان و برنامه‌ریزی برای آن کمک کند، وجود شناخت از ویژگی‌های زبانی و نیاز زبانی است که لانگ[1] (۲۰۰۵) به آن اشاره کرده‌است. برای فراهم‌آوردن زمینه‌های این شناخت، به جمع‌آوری دادهٔ زبانی و تهیه پیکرهٔ زبانی نیاز است. تهیهٔ این پیکره می‌تواند برای بررسی مسئلهٔ «واکاوی نیازها»[2] مفید باشد. واکاوی نیازها اصطلاحی است که به گفتهٔ لانگ (۲۰۰۵) نزدیک به ۳۰ سال است که در حوزهٔ آموزش زبان مطرح شده‌است. غریبی (۱۳۹۱) در چارچوب مسئلهٔ واکاوی نیازها، به بررسی واکاوی نیازهای عمومی زبانی فارسی‌آموزان پرداخته و عنوان کرده‌است که هیچ دورهٔ آموزشی یا کتاب درسی براساس نیازهای عمومی فارسی‌آموزان نوشته نشده‌است. یکی‌از دلایل آن، عدم آگاهی از ویژگی‌هایی است که باید در تنظیم محتوا در نظر گرفته شود. در راستای بهبود آموزش و سیاست‌گذاری در حوزهٔ آموزش، پاهنگ و همکاران (۱۳۹۶) با انجام تحقیقی پیمایشی-توصیفی، کیفیت مدارس را از منظر اهداف آموزشی و محتوای تدریس، منابع، روش تدریس و غیره بررسی کرده و به این نتیجه رسیده‌اند که محتوای تدریس و منابع فارسی کیفیت نامطلوبی دارد.

در این پژوهش تلاش می‌شود ضمن تهیهٔ یک پیکرهٔ زبانی در چارچوب زبان‌شناسی پیکره‌ای از کتاب‌های درسی پایه‌های اول تا ششم ابتدایی، با بهره‌گیری از زبان‌شناسی رایانشی، به‌صورت الگوریتمی این پیکره در سطح صرف، نحو و آوا برچسب‌گذاری گردد و واکاوی نیازها که بخشی از برنامه‌ریزی آموزشی نیز هست مورد بررسی قرار گیرد؛ زیرا بسیاری از مسائل

1. M. Long
2. needs analysis

آموزش، مانند استفاده از داده‌های طبیعی و مسائلی از این دست، نیازمند داده‌های نیازسنجی است که در این پژوهش داده‌های حاصل از کتاب‌های درسی مورد توجه قرار گرفته‌است. قطعاً محتوای کتاب‌های درسی یکی از مسائل پراهمیتی است که بر یادگیری مفاهیم اولیه و پایهٔ دانش‌آموزان و همچنین کیفیت آموزش اثرگذار است. یکی از کاربردهای مهم و کاربردی این پیکره، کمک به شناخت کلی متخصصان هنگام تهیهٔ محتوای کتاب‌های درسی است.

## ۲- پیشینهٔ مطالعاتی

برای زبان فارسی نیز همانند زبان‌های دیگر، پیکره‌های زبانی نوشتاری و گفتاری تهیه شده‌است. با این وجود زبان فارسی همچنان جزء زبان‌های با منابع محدود دسته‌بندی می‌شود. در این قسمت، شناخته‌شده‌ترین پیکره‌های نوشتاری تهیه‌شده برای زبان فارسی توضیح مختصر داده می‌شود.

عاصی (۱۹۹۷؛ ۱۳۷۶) پیکره‌ای بنام «پایگاه دادهٔ زبان فارسی» با حجم حدود سه میلیون واژه، متشکل از متون نوشتاری و مقداری گفتاری فارسی معاصر، را تهیه کرده‌است. شکل گسترش‌یافته این پایگاه، به بیش از ۷۰ میلیون واژه رسیده‌است. پایگاه دادهٔ گسترش‌یافته، علاوه‌بر فارسی معاصر، فارسی میانهٔ قرون ۵ تا ۷ را در بر دارد (عاصی و قندی، ۱۳۹۴). هدف اصلی تهیهٔ این پایگاه، استخراج اطلاعات آماری از پیکره و همچنین فرهنگ‌نگاری است. شایان ذکر است که حجم کمی از واژه‌های این پایگاه به‌طور نیمه‌خودکار و دستی نشانه‌گذاری شده‌است. در این نشانه‌گذاری، برچسب آوایی، بن‌واژه‌ای، و مقولهٔ دستوری هر واژه مشخص شده‌است. مقولات دستوری واژه‌ها براساس مجموعهٔ برچسب معرفی‌شده توسط عاصی و حاجی‌عبدالحسینی (۲۰۰۰) مشخص شده‌است.

قیومی (۱۳۸۳) و درودی و همکاران (۲۰۰۴) پیکره‌هایی را از متون روزنامه همشهری به‌ترتیب با حجم‌های ۶/۵ و ۳۷ میلیون واژه برای کاربرد در مدل‌سازی آماری زبان تهیه کرده‌اند. بین این دو مجموعه، همپوشی وجود ندارد و مکمل یکدیگر است.

آل‌احمد و همکاران (۲۰۰۹) پیکره‌ای از روزنامه همشهری به‌نام «پیکرهٔ همشهری» تهیه کرده و براساس ویژگی‌های «همایش بازیابی متن»[1] معیارسازی کرده‌اند. با اضافه‌کردن اطلاعات مربوط به ۶۵ جستجو و ۶۵۰۰ دادهٔ محک داوری، امکان استفاده از این داده برای کاربرد در حوزهٔ بازیابی اطلاعات[2] را فراهم آورده‌اند.

---

۱. Text Retrieval Conference (TREC).

2. information retrieval

بی‌جن‌خان و همکاران (۲۰۱۱) یک پیکرهٔ متنی، به‌نام «پیکره»، با حجم حدود صد میلیون واژه را با هدف مدل‌سازی آماری زبان برای زبان فارسی تهیه کرده‌اند. در این پیکره سعی شده‌است که معیارهای تهیهٔ پیکره، مانند نمایندهٔ زبان بودن، متوازن‌بودن و عدم جهت‌گیری در متون مدنظر قرار گیرد. مقولهٔ دستوری قسمتی از این داده، درحدود ۱۰ میلیون واژه، به‌صورت نیمه‌خودکار تعیین شده‌است. قسمتی از این دادهٔ برچسب‌گذاری‌شده، ۵/۲ میلیون واژه، به‌نام «پیکرهٔ بی‌جن‌خان» (بی‌جن‌خان، ۱۳۸۳)، در دسترس است.

پیکرهٔ پرسیکا که توسط اقبال‌زاده و همکاران (۲۰۱۲) معرفی شده‌است حاوی متون خبری برگرفته از خبرگزاری ایسنا است که در یازده مقولهٔ موضوعی شامل ورزشی، اقتصادی، فرهنگی، مذهبی، تاریخی، سیاسی، علمی، اجتماعی، آموزشی، حقوق قضایی و بهداشت طبقه‌بندی شده‌است. همچنین به‌منظور قابل‌استفاده‌شدن این پیکره در کاربردهای مختلف پردازش زبان طبیعی و داده‌کاوی، پیش‌پردازش‌هایی بر روی این پیکره انجام گرفته‌است.

صبوری و همکاران (۲۰۲۲) در پژوهش خود به معرفی پیکرهٔ «ناب» پرداخته‌اند. این پیکره شامل حدود ۱۳۰ گیگابایت متن تمیزشدهٔ فارسی است که متشکل از ۲۵۰ میلیون پاراگراف و ۱۵ میلیارد واژه است. پیکرهٔ متنی ناب به‌صورت کاملاً متن باز در اختیار همگان قرار دارد.

قیومی (۱۴۰۱) پیکرهٔ بزرگ خبری فارسی دیگری را تهیه کرده‌است که قابلیت به‌روزشوندگی خودکار دارد. در فرایند تهیه این پیکره، یک پیوستار پویا برای جمع‌آوری داده وجود دارد و مقطعی نیست. این پیکره از خزش ۲۴ وبگاه خبری و با حجمی بالغ بر ۱۴ میلیارد واژه گردآوری شده‌است. این پیکره برای ارزیابی ساختار هرم وارونهٔ خبر و یافتن همبستگی معنایی میان عنوان و بخش‌های مختلف خبر استفاده شده‌است.

در برچسب‌گذاری آوایی، بن‌واژه‌ای، مقولات دستوری واژه‌ها و ساخت سازه‌ای عبارات پژوهش‌هایی انجام شده‌است که به‌صورت گذرا به مرور آن‌ها می‌پردازیم.

### ۲-۱- برچسب‌زنی آوایی

اسلامی و همکاران (۱۳۸۳) یک واژگان با اندازه ۵۵هزار واژه تهیه کرده‌اند که متشکل از ۴۴هزار مدخل از پیکرهٔ متنی فارسی (بی‌جن‌خان و همکاران، ۲۰۱۱) و ۱۱هزار مدخل از فرهنگ معاصر فارسی امروز (صدری افشار و همکاران، ۱۳۸۱) است. این واژگان حاوی اطلاعاتی چون صورت نوشتاری، ساخت هجایی، مقوله دستوری، تکیه و بسامد واژه است. آنها در چارچوب پژوهش خود، مجموعهٔ دادهٔ «واژگان زایای فارسی» را تهیه کرده‌اند. این واژگان علاوه‌بر آوانویسی واژه‌ها، حاوی تأکید در هر هجا نیز است.

### ۲-۲- بن‌واژه‌سازی

با نگاهی به پژوهش‌های انجام‌شده در حوزهٔ صرف رایانشی می‌توان دریافت که حجم زیادی از منابع به موضوع ریشه‌یابی[1] پرداخته‌است که با بن‌واژه‌سازی[2] کمی متفاوت است؛ ازجمله پژوهش‌های تشکری و همکاران (۲۰۰۲)، مختاری‌پور و جهان‌پور (۲۰۰۶)، شریف‌لو و شمس‌فرد (۲۰۰۸) و جدیدی‌نژاد و همکاران (۲۰۱۰). گاهی میان ریشه‌یابی و بن‌واژه‌سازی تداخل صورت گرفته‌است. با این‌حال، به پژوهش‌های زیر که پایبند به استفاده از اصطلاح بن‌واژه‌سازی بوده‌است می‌توان توجه داشت:

تشکری و میبدی (۱۳۸۰) تلاش کرده‌اند با استفاده از روش حذف پسوند و پیشوندها، به‌طور قاعده‌مند، به بن‌واژهٔ واژه‌ها برسند.

موسوی میانگاه (۲۰۰۶) سامانه‌ای را برای بن‌واژه‌سازی زبان فارسی تهیه کرده‌است. در روش پیشنهادی وی، از روش جدول جستجو استفاده شده و ابتدا پیکره‌ای حاوی فهرستی از ریشه واژه‌ها، شکل تصریفی، پیشوندها و پسوندها به‌صورت دستی تهیه شده و با رویکرد احتمالاتی بهبود داده شده‌است.

فرزانه‌فر (۱۳۸۹) دو مدل برای بن‌واژه‌سازی زبان فارسی تهیه کرده‌است. در مدل اول از ماشین گزاره‌ای محدود[3] استفاده شده‌است که مدلی ایستا است و در مدل دوم از درخت تصمیم[4] استفاده شده‌است که مدلی پویا است.

شمس‌فرد و همکاران (۲۰۱۰) مجموعهٔ ابزاری را تهیه کرده‌اند که با استفاده از آن می‌توان متن را با توجه به دستور خط فرهنگستان زبان و ادب فارسی معیارسازی کرد. همچنین، این ابزار توانایی تصحیح و تقطیع واژگانی، بن‌واژه‌سازی، تحلیل واژگانی و برچسب‌گذاری مقوله دستوری را دارد.

دانش و همکاران (۲۰۱۱) با استخراج n-نگاشت و نمایه‌سازی[5] آن از پیکرهٔ همشهری (آل‌احمد و همکاران، ۲۰۰۹)، یک بن‌واژه‌ساز را برای کاربرد در بازیابی اطلاعات تهیه کرده‌اند.

بابادی و همکاران (۱۳۹۱) تلاش کرده‌اند با استفاده از روش‌های یادگیری ماشین به بررسی بن‌واژه‌سازی بپردازند. در این روش، از پسوندها و پیشوندهایی که دارای الگویی برای ساخت واژهٔ جدید از ریشه است استفاده می‌شود. در ادامه، نیز، برای یافتن بن‌واژهٔ واژه‌های جمع مکسر، از شبکهٔ عصبی استفاده کرده‌اند.

---

1. stemming
2. lemmatization
3. finite state automata
4. decision tree
5. indexing

دولامیک[1] و ساووی[2] (۲۰۰۹) اقدام به تهیهٔ ریشه‌یاب در فارسی کرده‌اند و از آن برای بازیابی اطلاعات استفاده کرده‌اند.

قیومی (۱۳۹۸) در یک فرایند دومرحله‌ای موفق شده‌است ابتدا پیکرهٔ بی‌جن‌خان (۱۳۸۳) را به‌صورت قاعده‌مند بن‌واژه‌سازی نماید و سپس از دادهٔ به‌دست‌آمده برای آموزش بن‌واژه‌ساز آماری مبتنی‌بر یادگیری ماشینی استفاده نماید.

### ۳-۲- برچسب‌زنی مقولات دستوری

به‌نظرمی‌رسد اولین تلاش برای برچسب‌زنی واژه در زبان فارسی به‌صورت نیمه‌خودکار توسط عاصی و حاجی‌عبدالحسینی (۲۰۰۰) در قالب مدل معرفی‌شده توسط شوتز[3] (۱۹۹۵) انجام شده‌است. آن‌ها ۴۳ برچسب معرفی کرده و از داده پایگاه دادهٔ زبان فارسی (عاصی، ۱۹۹۷؛ ۱۳۷۶) استفاده کرده‌اند.

ارومچیان و همکاران (۲۰۰۶) و امیری و همکاران (۲۰۰۷) در پژوهش خود از مدل احتمالات بیشینه[4] استفاده کرده و مدل خود را با پیکرهٔ بی‌جن‌خان (۱۳۸۳) که دارای ۵۸۶ برچسب است آموزش داده‌اند.

تشرفی و همکاران (۲۰۰۷) از برچسب‌زن تی.ان.تی.[5] (برنتز[6]، ۲۰۰۰) استفاده کرده و آن را با پیکرهٔ بی‌جن‌خان آموزش داده‌اند.

شمس‌فرد و فدایی (۲۰۰۸) یک روش ترکیبی که از تحلیل صرفی استفاده می‌کند را برای برچسب‌زنی ابداع کرده‌اند. تعداد برچسب‌های معرفی‌شده آن‌ها ۲۵ برچسب بوده و از پیکرهٔ همشهری (آل‌احمد و همکاران ، ۲۰۰۹) برای پژوهش خود استفاده کرده‌اند.

محسنی (۱۳۸۶) یک تحلیلگر صرفی برای برچسب‌زنی تهیه کرده‌است. در این پژوهش، تعداد برچسب‌های پیکرهٔ بی‌جن‌خان از ۵۸۶ برچسب به ۱۰۵ برچسب کاهش یافته‌است.

سراجی (۲۰۱۱) نیز یک برچسب‌زن زبان مجارستانی را با پیکرهٔ بی‌جن‌خان آموزش داده و از آن برای برچسب‌زنی متن فارسی استفاده کرده‌است.

---

1. L. Dolamic
2. J. Savoy
3. H. Schütze
4. maximum likelihood estimation
5. TnT
6. T. Brants

زاگوت[1] و همکاران (۲۰۱۱) از سامانهٔ مِلت[2] (دنیس[3] و زاگوت، ۲۰۰۹) برای برچسب‌زنی استفاده کرده‌اند. آن‌ها در این پژوهش برچسب‌های موجود در پیکرهٔ بی‌جن‌خان را با توجه به نیاز تغییر داده‌اند.

علایی‌ابوذر (۱۳۹۹) در پژوهشی با عنوان «بررسی امکان افزایش صحت یک ابزار برچسب-دهی به اجزای کلام در فارسی» ابزار برچسب‌زنی «هضم»[4] را مورد مطالعه و بررسی قرار داده‌است.

محمدی و همکاران (۲۰۲۳) با استفاده از الگوریتم «هوش ازدحامی»[5]، روشی را برای برچسب‌گذاری واژه‌های فارسی با کارآیی بالا پیشنهاد داده‌اند. این دسته از الگوریتم‌ها که از رفتار جمعی موجودات طبیعی الهام گرفته‌است در سال‌های اخیر به‌عنوان ابزاری برای حل مسائل پیچیده و غیرقطعی مورد توجه قرار گرفته‌است. بهینه‌سازی کلونی مورچه‌ها[6] که از رفتار جستجوی مورچه‌ها و فرایند تخمگذاری آن‌ها الهام گرفته شده‌است یکی از این الگوریتم‌ها است که در این پژوهش در فرایند برچسب‌گذاری مقولات دستوری پیشنهاد شده‌است. این روش به دقت ۹۶/۸۷ درصد دست یافته‌است.

ملانوروزی و همکاران (۲۰۲۳) با استفاده از یک مدل «یادگیری انتقالی بین‌زبانی»[7] و مدل‌های ازپیش‌آموزش‌دیده‌شده[8] به برچسب‌دهی مقولات دستوری فارسی پرداخته‌اند. همچنین، از داده‌های اطلس جهانی ساختار زبانی، برای یافتن تشابه میان ساختار زبانی فارسی و سایر زبان‌های جهان پرداخته‌اند تا از این ویژگی برای هدف مورد نظر استفاده نمایند. از نظر ساخت زبانی، دو زبان کرمانجی و تاگالوگ، زبان‌های مشابه به فارسی شناسایی شد. همچنین در این پژوهش از اطلاعات ۳۱ زبان مشابه به فارسی برای تحلیل احساسات استفاده شده‌است.

### ۲-۴- تجزیهٔ نحوی

تجزیهٔ نحوی جملات به دو دستهٔ سازه‌ای و وابستگی تقسیم می‌شود. در تجزیهٔ سازه‌ای، تجزیهٔ نحوی به‌صورت درخت‌های سازه‌ای سلسله‌مراتبی ارائه می‌گردد که منطبق با دستور ساخت سازه‌ای چامسکی[9] (۱۹۵۷) است. در تجزیهٔ نحوی وابستگی، تجزیهٔ نحوی به‌صورت وابستگی

1. B. Sigot
2. Melt
3. P. Denis
4. https://www.roshan-ai.ir/hazm/docs/index.html
5. Swarm Intelligence
6. Ant colony optimization
7. Cross-lingual transfer learning
8. pre-trained
9. N. Chomsky

بین واژه‌ها ارائه می‌گردد که ریشه در آرای دستور وابستگی ارائه‌شده توسط تنیر[1] (۱۹۵۳؛ ۱۹۵۹؛ ۱۹۸۰) دارد.

در تجزیهٔ نحوی، معمولاً از روش آماری مبتنی‌بر یادگیری ماشینی استفاده می‌شود. برای آموزش ماشین و تهیهٔ مدل پردازشی به یک پیکرهٔ برچسب‌خوردهٔ حاوی ساختار درختی چندین جمله که دادگان درختی[2] نامیده می‌شود نیاز است.

اولین دادگان درختی سازه‌ای با تعداد ۱۰۲۶ جمله در چارچوب دستور ساخت سازه‌ای هسته‌بنیان (پولارد[3] و ساگ[4]، ۱۹۹۴) برای فارسی توسط قیومی (۲۰۱۲a,b) تهیه شده‌است و این پیکره برای آموزش تجزیه‌گر استنفورد[5] (کلاین[6] و منینگ[7]، ۲۰۰۳)، برکلی[8] (پترو[9] و همکاران، ۲۰۰۶) و بیتپر[10] (اشمیت[11]، ۲۰۰۴) استفاده شده‌است. قیومی و کوهن[12] (۲۰۱۴) این پیکره را به‌طور خودکار به دستور وابستگی تبدیل کرده و از آن برای آموزش تجزیه‌گرهای وابستگی مالت[13] (نیوه[14] و همکاران، ۲۰۰۶) و میت[15] (بونت[16]، ۲۰۰۹) استفاده کردند.

سراجی و همکاران (۲۰۱۲) یک دادگان درختی به‌نام «دادگان درختی اوپسالا[17]» با تعداد ۶۰۰۰ جمله را تهیه کرده و از آن برای آموزش تجزیه‌گرهای مالت (نیوه و همکاران، ۲۰۰۶) و ام.اس.تی.[18] (مک‌دونالد[19] و همکاران، ۲۰۰۵) استفاده کرده‌اند.

رسولی و همکاران (۲۰۱۳) در تلاشی دیگر، یک دادگان درختی وابستگی تهیه کرده‌اند که حاوی حدود ۳۰هزار جمله است.

---

1. L. Tesnière
2. treebank
3. C. J. Pollard
4. I. A. Sag
5. Stanford Parser
6. D. Klein
7. C.D. Manning
8. Berkeley Parser
9. S. Petrov
10. BitPar
11. H. Schmit
12. J. Kuhn
13. Malt Parser
14. J. Nivre
15. Mate
16. B. Bohnet
17. Uppsala
18. Maximum Spanning Tree (MST)
19. R. McDonald

طباطبایی و صراف‌زاده (۱۳۹۶) در چارچوب دستور سازه‌ای، یک دادگان درختی بزرگ که شامل ۱۵۷هزار جمله است را به‌صورت نیمه‌خودکار تهیه کرده و برای آموزش تجزیه‌گر برکلی (پترو و همکاران، ۲۰۰۶) استفاده کرده‌اند.

دهقان و همکاران (۱۳۹۶) دادگان درختی وابستگی رسولی و همکاران (۲۰۱۳) را به‌طور خودکار به دستور سازه‌ای تبدیل کرده و از آن برای آموزش تجزیه‌گر استنفورد (کلاین و منینگ، ۲۰۰۳) و برکلی (پترو و همکاران، ۲۰۰۶) استفاده کرده‌اند.

## ۳- چارچوب نظری

در مکتب ساختگرایی، سوسور[1] (۱۹۱۶) دو سطح برای زبان قائل است. یک سطح «صورت» است که نشانه‌های زبانی ملموس در این سطح طبقه‌بندی می‌شود. سطح دیگر «معنا» است که انتزاعی بوده و با مفهوم در ارتباط است. صورت زبانی به دو صورت آواهای زبانی در گفتار یا خط در نوشتار تجلی پیدا می‌کند. این نشانه‌های ملموس، در محور همنشینی عبارات و جملات را با هم می‌سازد. در حوزهٔ زبان‌شناسی رایانشی، این دو صورت قابلیت پردازش دارد. ازآنجاکه رایانه نشانه‌های ملموس را می‌شناسد، باید هم صورت و هم معنا به‌صورت نشانه‌های قابل پردازش بازنمایی یابد تا برای رایانه قابل استفاده گردد. در پردازش‌های رایانشی صورت آوایی، امواج صوتی به‌عنوان دادهٔ ورودی یک الگوریتم پردازشی در نظر گرفته می‌شود، درحالی‌که در پردازش نوشتاری، متن نگارش‌شده، دادهٔ ورودی الگوریتم پردازشی است. شایان ذکر است که می‌توان آواهای زبانی را در قالب صورت نوشتاری آوانگاری نمود و به‌عنوان دادهٔ ورودی به الگوریتم پردازشی داد. صورت معنایی نیز می‌تواند به روش‌های مختلفی، ازجمله بردار، بازنمایی گردد.

اهمیت و ضرورت زبان‌شناسی پیکره‌ای که در دههٔ ۱۹۶۰ معرفی شد بر همگان پوشیده نیست و در قسمت مقدمه به اهمیت این موضوع برای سیاست‌گذاری در حوزهٔ زبان اشاره شد. بسیاری از پژوهش‌های زبان‌شناسی تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. پیکره، مجموعه‌ای از متون نوشتاری و گفتاری آوانویسی‌شده است که می‌توان آن را مبنایی برای تحلیل و توصیف زبانی قرار داد (عاصی و ترابی،۱۳۹۱). اتکینز[2] و همکاران (۱۹۹۲) انواع پیکره را معرفی کرده و از دیدگاه‌های مختلف بررسی کرده‌اند. براساس این تقسیم‌بندی، پیکره‌ها به

1. F. de Saussure
2. S. Atkins

انواع «متن کامل»، «نمونه‌ای»، «نظارتی»، «بسته»، «باز»، «هم‌زمانی»، «درزمانی»، «مرکزی»، «پوسته‌ای»، «هسته‌ای»، و «پیرامونی» طبقه‌بندی می‌شود.

هدف اصلی زبان‌شناسی رایانشی درک و تولید زبان طبیعی است (جورافسکی[1] و مارتین[2]، ۲۰۲۳). برای درک زبان نیاز است تمامی اطلاعات زبان‌شناختی در حوزه‌های آواشناسی، صرف، نحو، معنا و کاربرد در یک پیکرۀ زبانی نشانه‌گذاری گردد. بنابراین اهمیت تهیۀ پیکره‌های زبانی دوچندان می‌گردد. با رویکرد پردازش رایانشی و دیدگاه سوسور، می‌توان با پردازش صورت زبانی در قالب امواج صوتی یا خط، به مفهوم رسید و امکان یافتن روابط بین واژه‌ها در محور همنشینی یا جانشینی را میسر ساخت.

برخی از کاربردهای پیکره در پردازش زبان طبیعی و درک، بازشناسی گفتار، تبدیل متن به گفتار و برعکس، تدوین فرهنگ‌ها، ترجمۀ ماشینی، گویش‌شناسی و سایر پژوهش‌های زبان‌شناختی است. در استفاده از پیکره در مطالعات زبان‌شناسی سه رویکرد رایج وجود دارد: الف) رویکرد پیکره‌بنیاد: در این رویکرد نظریه مطرح می‌شود و برای سنجش نظریه از پیکره استفاده می‌شود؛ ب) رویکرد برگرفته از پیکره: در این رویکرد از پیکره برای بازیابی اطلاعات استفاده می‌شود؛ ج) رویکرد پیکره‌یار: این رویکرد ترکیبی است از به کارگیری روش‌های کمی و کیفی (علایی‌ابوذر و همکاران، ۱۴۰۰).

در پژوهش حاضر، با دیدگاه سوسور درمورد صورت زبانی، به توصیف ویژگی‌های موجود در پیکرۀ زبانی حاصل از محتوای کتاب‌های درسی پایه‌های اول تا ششم می‌پردازیم تا شناختمان نسبت‌به ویژگی‌های زبانی این گونۀ زبانی افزایش یابد و از توصیف ویژگی‌های آن در راستای نیل به اهداف برنامه‌ریزی زبانی مربوط به زبان فارسی که در بخش مقدمه به آن اشاره شد استفاده گردد.

## ۵- تهیۀ پیکرۀ زبانی

در طراحی کلی پیکره، معیارهایی از قبیل نوع متن استفاده‌شده در تهیۀ پیکره، تعداد متون، انتخاب متون خاص، طول نمونه‌های متون و مواردی از این دست وجود دارد که هر کدام مستلزم تصمیم‌گیری دربارۀ نمونه‌گیری[3] است. نمونه‌گیری در واقع عمل انتخاب متون باتوجه‌به ژانر و هدف مورد نظر است. داده‌های پیکرۀ مورد نظر در این پژوهش از کتاب‌های درسی فارسی، علوم، مطالعات اجتماعی و هدیه‌های آسمانی گردآوری شده‌است. بنابراین، گونۀ زبانی این متون، نوشتاری و رسمی بوده و حوزۀ متون علمی می‌باشد.

---

1. D. Jurafsky
2. H. Martin
3. sampling

جمع‌آوری داده به‌منظور تهیهٔ پیکرهٔ مد نظر به این صورت است که تمام جملات دروس مختلف به تفکیک پایه، در فایل‌های متنی ساده حروف‌نگاری شده و پس‌از هنجارسازی در فرایند پیش‌پردازش، برای برچسب‌گذاری آماده می‌گردد. در فرایند حروف‌نگاری، فقط به قسمت‌هایی از محتوای کتاب که شامل متن پیوسته است توجه شده‌است. بنابراین سؤالات در انتهای دروس، تمارین، جداول و مانند آن در تهیهٔ پیکره کنار گذاشته شده‌است. جدول ۱ حاوی اطلاعات آماری استخراج‌شده از پیکرهٔ خام تهیه‌شده است. این پیکره در چهار سطح آوایی، بن‌واژه، مقولات دستوری و ساخت سازه‌ای برچسب‌گذاری می‌شود که در بخش بعدی توضیح داده خواهد شد.[۱]

**جدول ۱: آمار استخراج‌شده از پیکرهٔ خام تهیه‌شده**

| پایه | درس | تعداد جمله | تعداد واژه | تعداد واژه‌های یکتا | تعداد بن‌واژه | تعداد حروف | تنوع واژگانی صورت‌واژه | طول متوسط جمله | تنوع واژگانی بن‌واژه | طول متوسط واژه |
|---|---|---|---|---|---|---|---|---|---|---|
| اول | فارسی | ۴۰۰ | ۳۹۱۸ | ۹۷۷ | ۷۲۲ | ۱۳۸۷۷ | ۰/۲۵ | ۹/۸۰ | ۰/۷۴ | ۳/۵۴ |
| | علوم | ۴۵۱ | ۴۸۱۱ | ۱۱۳۲ | ۸۰۰ | ۱۸۶۳۰ | ۰/۲۴ | ۱۰/۶۷ | ۰/۷۱ | ۳/۸۷ |
| | کل | ۸۵۱ | ۸۷۲۹ | ۱۷۷۴ | ۱۲۳۶ | ۳۲۵۰۷ | ۰/۲۰ | ۱۰/۲۶ | ۰/۷۰ | ۳/۷۲ |
| دوم | فارسی | ۹۷۱ | ۱۲۹۱۲ | ۲۱۹۴ | ۱۴۰۵ | ۴۲۰۴۶ | ۰/۱۷ | ۱۳/۳۰ | ۰/۶۴ | ۳/۲۶ |
| | علوم | ۷۵۰ | ۸۱۵۷ | ۱۴۲۵ | ۹۹۵ | ۳۰۷۵۶ | ۰/۱۷ | ۱۰/۸۸ | ۰/۷۰ | ۳/۷۷ |
| | هدیه‌ها | ۳۴۴ | ۳۹۴۴ | ۹۷۵ | ۶۵۶ | ۱۳۶۴۵ | ۰/۲۵ | ۱۱/۴۷ | ۰/۶۷ | ۳/۴۶ |
| | کل | ۲۰۶۵ | ۲۵۰۱۳ | ۳۳۷۴ | ۲۰۹۷ | ۸۶۴۴۷ | ۰/۱۳ | ۱۲/۱۱ | ۰/۶۲ | ۳/۴۶ |

۱ . منظور از «اجتماعی» و «هدیه‌ها» در جداول، درس‌های «مطالعات اجتماعی» و «هدیه‌های آسمانی» است.

**ادامه جدول ۱:**

| پایه | درس | تعداد جمله | تعداد واژه | تعداد واژه‌های یکتا | تعداد نوواژه | تعداد حروف | تنوع واژگانی صورت‌واژه | طول متوسط جمله | تنوع واژگانی نوواژه | طول متوسط واژه |
|---|---|---|---|---|---|---|---|---|---|---|
| سوم | فارسی | ۷۷۷ | ۱۰۰۷۲ | ۲۲۲۴ | ۲۹۸۰ | ۴۵۲۵۷ | ۰/۲۲ | ۱۲/۹۶ | ۱/۳۴ | ۴/۴۹ |
| | علوم | ۶۱۵ | ۶۷۶۵ | ۱۲۳۰ | ۱۷۴۰ | ۵۳۱۳۵ | ۰/۱۸ | ۱۱/۰۰ | ۱/۴۱ | ۷/۸۵ |
| | هدیه‌ها | ۶۳۳ | ۸۰۱۴ | ۱۶۹۹ | ۱۰۹۹ | ۲۹۷۵۷ | ۰/۲۱ | ۱۲/۶۶ | ۰/۶۵ | ۳/۷۱ |
| | اجتماعی | ۴۲۵ | ۵۴۳۶ | ۱۳۵۹ | ۱۴۹۰ | ۱۹۶۷۵ | ۰/۲۵ | ۱۲/۷۹ | ۱/۱۰ | ۳/۶۲ |
| | کل | ۲۴۵۰ | ۳۰۲۸۷ | ۴۱۸۷ | ۲۵۸۰ | ۱۴۷۸۲۴ | ۰/۱۴ | ۱۲/۳۶ | ۰/۶۲ | ۴/۸۸ |
| چهارم | فارسی | ۹۰۱ | ۱۲۵۸۰ | ۲۷۳۲ | ۱۷۶۰ | ۴۵۲۵۷ | ۰/۲۲ | ۱۳/۹۶ | ۰/۶۴ | ۳/۶۰ |
| | علوم | ۱۲۰۶ | ۱۳۸۸۱ | ۱۹۴۳ | ۱۳۶۳ | ۲۴۹۱۸ | ۰/۱۴ | ۱۱/۵۱ | ۰/۷۰ | ۱/۸۰ |
| | هدیه‌ها | ۶۵۸ | ۸۸۹۳ | ۱۹۶۴ | ۱۳۰۷ | ۵۳۱۳۵ | ۰/۲۲ | ۱۳/۵۲ | ۰/۶۷ | ۵/۹۷ |
| | اجتماعی | ۶۲۸ | ۹۵۹۳ | ۲۰۷۷ | ۱۴۵۴ | ۲۹۷۵۷ | ۰/۲۲ | ۱۵/۲۸ | ۰/۷۰ | ۳/۱۰ |
| | کل | ۲۱۴۰ | ۲۶۴۹۴ | ۵۷۸۲ | ۳۵۶۶ | ۷۰۲۰۸ | ۰/۲۲ | ۱۲/۳۸ | ۰/۶۲ | ۲/۶۵ |
| پنجم | فارسی | ۷۳۰ | ۱۲۰۶۸ | ۲۸۶۰ | ۱۸۵۰ | ۴۳۹۰۸ | ۰/۲۴ | ۱۶/۵۳ | ۰/۶۵ | ۳/۶۴ |
| | علوم | ۱۰۶۶ | ۱۲۶۹۳ | ۱۸۰۰ | ۱۳۵۳ | ۴۷۰۸۲ | ۰/۱۴ | ۱۱/۹۱ | ۰/۷۵ | ۳/۷۱ |
| | هدیه‌ها | ۶۷۴ | ۱۰۷۲۸ | ۲۴۹۱ | ۱۶۶۸ | ۳۷۷۹۴ | ۰/۲۳ | ۱۵/۹۲ | ۰/۶۷ | ۳/۵۲ |
| | اجتماعی | ۸۳۰ | ۱۵۱۰۸ | ۲۸۳۰ | ۲۰۴۶ | ۵۶۵۲۵ | ۰/۱۹ | ۱۸/۲۰ | ۰/۷۲ | ۳/۷۴ |
| | کل | ۳۳۰۰ | ۵۰۵۹۷ | ۶۵۵۴ | ۴۱۲۱ | ۱۸۵۳۰۹ | ۰/۱۳ | ۱۵/۳۳ | ۰/۶۳ | ۳/۶۶ |

**ادامه جدول ۱:**

| پایه | درس | تعداد جمله | تعداد واژه | تعداد واژه‌های یکتا | تعداد بن‌واژه | تعداد حروف | تنوع واژگانی صورت‌واژه | طول متوسط جمله | تنوع واژگانی بن‌واژه | طول متوسط واژه |
|---|---|---|---|---|---|---|---|---|---|---|
| ششم | فارسی | ۶۳۸ | ۱۰۵۱۶ | ۲۶۷۸ | ۱۸۵۰ | ۱۹۳۱۹ | ۰/۲۵ | ۱۶/۴۸ | ۰/۶۹ | ۱/۸۴ |
| | علوم | ۶۶۰ | ۹۶۸۲ | ۱۸۷۸ | ۱۳۵۳ | ۳۷۶۷۱ | ۰/۱۹ | ۱۴/۶۷ | ۰/۷۲ | ۳/۸۹ |
| | هدیه‌ها | ۶۴۱ | ۱۱۰۵۴ | ۲۴۱۳ | ۱۶۶۸ | ۳۰۶۳۲ | ۰/۲۲ | ۱۷/۲۴ | ۰/۶۹ | ۲/۷۷ |
| | اجتماعی | ۹۶۰ | ۱۷۲۴۱ | ۳۲۴۲ | ۲۳۲۶ | ۶۶۸۵۰ | ۰/۱۹ | ۱۷/۹۶ | ۰/۷۲ | ۳/۸۸ |
| | کل | ۲۸۹۹ | ۴۸۴۹۳ | ۶۷۷۳ | ۴۳۹۹ | ۱۵۴۴۷۲ | ۰/۱۴ | ۱۶/۷۳ | ۰/۶۵ | ۳/۱۹ |
| اول تا ششم | فارسی | ۴۴۱۷ | ۶۲۰۶۱ | ۷۱۶۶ | ۴۴۳۵ | ۲۱۸۱۰۷ | ۰/۱۲ | ۱۴/۰۵ | ۰/۶۲ | ۳/۵۱ |
| | علوم | ۴۷۴۸ | ۵۵۹۸۴ | ۴۴۸۶ | ۲۸۴۷ | ۲۱۲۵۳۱ | ۰/۰۸ | ۱۱/۷۹ | ۰/۶۳ | ۳/۸۰ |
| | هدیه‌ها | ۲۹۵۰ | ۴۲۶۲۹ | ۵۱۳۳ | ۳۰۸۵ | ۱۴۷۹۹۳ | ۰/۱۲ | ۱۴/۴۵ | ۰/۶۰ | ۳/۴۷ |
| | اجتماعی | ۲۸۴۳ | ۴۷۳۷۵ | ۵۷۹۳ | ۳۸۳۵ | ۱۷۹۷۱۸ | ۰/۱۲ | ۱۶/۶۶ | ۰/۶۶ | ۳/۷۹ |
| | کل | ۱۴۹۵۸ | ۲۰۸۰۴۹ | ۱۳۳۷۴ | ۷۸۸۷ | ۷۵۸۳۴۹ | ۰/۰۶ | ۱۳/۹۱ | ۰/۵۹ | ۳/۶۵ |

براساس آمار گزارش‌شده در جدول ۱، پیکرهٔ حاصل از کتاب‌های درسی مورد نظر، متشکل از حدود ۲۰۸هزار صورت‌واژه است که به‌ترتیب درس فارسی حدود ۳۰٪، درس علوم حدود ۲۷٪، درس مطالعات اجتماعی حدود ۲۳٪ و درس هدیه‌های آسمانی حدود ۲۰٪ از واژه‌های پیکره را در بر گرفته‌است. این توزیع آماری براساس پایهٔ تحصیلی از این قرار است که پایهٔ اول ۴/۲۰ ٪، پایهٔ دوم ۱۲/۰۲٪، پایهٔ سوم ۱۴/۵۶٪، پایهٔ چهارم ۱۲/۷۴٪، پایهٔ پنجم ۲۴/۳۲٪ و پایهٔ ششم ۲۳/۳۱٪ از واژگان را در بر گرفته‌است.

طول متوسط واژه‌های پیکره حدود ۴ حرف است و این ویژگی کم‌وبیش در دروس مختلف وجود دارد. طول متوسط واژه‌ها در پایهٔ سوم حدود ۵ واژه و در کلاس چهارم حدود ۳ واژه است. طول متوسط جمله پیکره حدود ۱۴ واژه است؛ منتها طول جملات در دروس مختلف و پایه‌های مختلف متفاوت است به این صورت که جملات درس مطالعات اجتماعی در مقایسه با سایر دروس طولانی‌تر (با طول متوسط حدود ۱۷ واژه) و جملات درس علوم کوتاهتر (با طول متوسط حدود ۱۲ واژه) است. کوتاهی جملات درس علوم بیانگر این نکته است که در محتوای علمی، معمولاً جملات کوتاه است ولی محتوای علوم انسانی معمولاً طولانی است. تعداد واژه‌ها در پایه‌های اول تا چهار کمتر از این میزان و پایه‌ها پنجم و ششم بیشتر از این میزان است. طول متوسط جمله بیان می‌کند که در سطوح بالای تحصیلی، جملات طولانی‌تر شده و از ساخت‌های زبانی پیچیده‌تری استفاده می‌شود؛ درحالی‌که، در پایه‌ها اول تا چهارم تلاش می‌شود از ساخت‌های زبانی ساده‌تر استفاده گردد.

تنوع واژگانی از دو جنبهٔ صورت‌واژه یا بن‌واژه بررسی شد. تنوع واژگانی براساس صورت‌واژه در کتاب‌های درسی نسبتاً پایین بود به این مفهوم که از صورت‌واژه‌های متنوع کمتری برای بیان محتوا استفاده شده‌است. تنوع واژگانی در درس مطالعات اجتماعی بیشتر از سایر دروس بود و برعکس در درس علوم کمتر بود. پایه‌های اول و چهارم بیشترین و پایه‌های دوم و پنجم کمترین تنوع واژگانی صورت‌واژه‌ها را داشت. درحالی‌که تنوع واژگانی براساس بن‌واژه نسبتاً بالا بود. تنوع واژگانی براساس بن‌واژه در پایهٔ اول بیشترین (۰/۷۰)، پایهٔ ششم در جایگاه دوم (۰/۶۵)، پایهٔ پنجم در جایگاه سوم (۰/۶۳) و سایر پایه‌ها با عدد ۰/۶۲ در جایگاه بعدی قرار داشت. وجود تنوع واژگانی بالا براساس بن‌واژه در پایهٔ اول بیانگر این نکته است که در این پایه چندان از تصریف واژگانی استفاده نمی‌شود و تلاش می‌شود واژه‌های جدید در محتوا استفاده گردد؛ درحالی‌که، با افزایش پایهٔ تحصیلی، تنوع واژگانی رو به کاهش می‌نهد و تصریف واژگانی و استفاده تکراری از واژه‌ها انجام می‌پذیرد.

## ۵- برچسب‌گذاری پیکرهٔ تهیه‌شده

### ۵-۱- تحلیل هجایی واژه‌ها

یکی از سطوح تحلیل داده مورد نظر، تحلیل هجایی واژه‌ها است. برای این هدف از فرهنگ واژگان زایای تهیه‌شده توسط اسلامی و همکاران (۱۳۸۳) استفاده کرده‌ایم. در این پیکره، آوانویسی واژه‌ها و تقطیع هجایی واژه‌ها انجام پذیرفته‌است. با کدنویسی انجام‌شده در این پژوهش، آواها براساس مصوت و غیرمصوت با V و C جایگزین شد و سپس به تقطیع هجایی به الگوهای هجایی CV، CVC و CVCC اقدام کردیم. در این پژوهش، باتوجه‌به کسرهٔ اضافهٔ

شخیص‌داده‌شده براساس برچسب نحوی مقولهٔ دستوری واژه که در ادامه توضیح داده خواهد شد، یک هجا به هجاهای موجود اضافه کرده‌ایم.

در جدول ۲ توزیع آماری هجای واژه‌ها و الگوهای هجایی به تفکیک درس و پایه گزارش شده‌است. پایه‌های اول تا سوم حدود ۳۰ درصد از هجای واژه‌های موجود در پیکره را به خود اختصاص داده‌است و توزیع نسبی هجای واژه‌ها در پایه‌های چهارم تا ششم تقریباً متوازن است. توزیع نسبی هجاها در پایه‌های پنجم و ششم به‌طور مساوی و جداگانه ۲۴درصد از مجموع هجاهای پیکره را شامل شده‌است. در میان الگوهای هجایی فارسی، هجای CV پرکاربردترین الگو در مقایسه با دو الگوی دیگر است. توزیع هجاها در دو درس فارسی و علوم در پایه‌های اول تا ششم ۲۸ درصد است که در مقایسه با دو درس دیگر بیشتر است. درس مطالعات اجتماعی ۲۴ درصد و هدیه‌های آسمانی ۲۰ درصد از مجموع هجاهای پیکره را شامل شده‌است.

**جدول ۲: توزیع آماری هجای واژه‌ها و الگوهای هجایی**

| پایه | درس | تعداد کل هجاهای کلمات | تعداد الگوی CVCC | تعداد الگوی CVC | تعداد الگوی CV |
|---|---|---|---|---|---|
| اول | فارسی | ۷۵۵۱ | ۳۸۲ | ۱۹۱۳ | ۵۱۵۶ |
| | علوم | ۹۹۷۸ | ۳۹۰ | ۲۷۹۹ | ۶۶۹۷ |
| | کل (نسبی) | ۰/۰۴ | ۴/۴۰ | ۲۶/۸۸ | ۶۷/۶۲ |
| دوم | فارسی | ۲۲۵۲۳ | ۱۳۴۴ | ۶۲۵۲ | ۱۴۵۹۸ |
| | علوم | ۱۶۶۷۳ | ۵۶۶ | ۴۶۴۱ | ۱۱۲۶۷ |
| | هدیه‌ها | ۷۳۱۲ | ۴۳۲ | ۱۹۹۴ | ۴۷۹۹ |
| | کل (نسبی) | ۰/۱۱ | ۵/۰۴ | ۲۷/۷۱ | ۶۵/۹۳ |
| سوم | فارسی | ۱۸۹۰۰ | ۱۱۵۸ | ۵۲۸۲ | ۱۲۲۷۹ |
| | علوم | ۱۳۷۹۸ | ۴۶۱ | ۳۸۰۵ | ۹۳۶۸ |
| | هدیه‌ها | ۱۴۷۲۸ | ۸۰۰ | ۴۲۴۳ | ۹۵۴۹ |
| | اجتماعی | ۱۰۶۱۵ | ۴۶۶ | ۲۹۰۷ | ۷۱۷۳ |
| | کل (نسبی) | ۰/۱۴ | ۴/۹۷ | ۲۷/۹۸ | ۶۶/۱۱ |
| چهارم | فارسی | ۲۳۵۳۳ | ۱۳۹۷ | ۷۱۴۸ | ۱۴۷۵۹ |
| | علوم | ۲۸۷۵۶ | ۱۱۵۶ | ۷۶۷۲ | ۱۹۴۵۷ |
| | هدیه‌ها | ۱۶۰۳۱ | ۸۶۲ | ۴۷۴۵ | ۱۰۲۷۴ |
| | اجتماعی | ۱۹۵۳۱ | ۹۶۵ | ۵۶۰۶ | ۱۲۷۸۳ |
| | کل (نسبی) | ۰/۲۲ | ۴/۹۹ | ۲۸/۶۵ | ۶۵/۱۹ |

**ادامه جدول ۲:**

| پایه | درس | تعداد کل هجاهای کلمات | تعداد الگوی CVCC | تعداد الگوی CVC | تعداد الگوی CV |
|---|---|---|---|---|---|
| پنجم | فارسی | ۲۳۰۷۷ | ۱۳۱۰ | ۷۰۳۰ | ۱۴۵۶۱ |
| | علوم | ۲۵۳۷۰ | ۸۹۱ | ۷۱۹۲ | ۱۶۷۶۹ |
| | هدیه‌ها | ۲۰۲۳۴ | ۹۶۰ | ۵۹۲۰ | ۱۳۰۷۰ |
| | اجتماعی | ۳۰۹۰۲ | ۱۳۲۹ | ۹۰۲۴ | ۲۰۳۱۸ |
| | کل (نسبی) | ۰/۲۴ | ۴/۵۱ | ۲۹/۲۹ | ۶۴/۹۹ |
| ششم | فارسی | ۱۹۶۶۵ | ۱۰۵۳ | ۶۲۵۹ | ۱۲۱۵۸ |
| | علوم | ۲۰۵۰۱ | ۷۵۳ | ۵۶۱۶ | ۱۳۸۶۶ |
| | هدیه‌ها | ۲۱۲۶۸ | ۱۰۱۱ | ۶۱۴۵ | ۱۳۸۶۴ |
| | اجتماعی | ۳۶۱۱۱ | ۱۵۲۴ | ۱۰۶۹۷ | ۲۳۵۶۳ |
| | کل (نسبی) | ۰/۲۴ | ۴/۴۵ | ۲۹/۴۴ | ۶۵/۰۵ |
| اول تا ششم | فارسی | ۱۱۵۲۴۹ | ۶۶۴۴ | ۳۳۸۸۴ | ۷۳۵۱۱ |
| | علوم | ۱۱۵۰۷۶ | ۴۲۱۷ | ۳۱۷۲۵ | ۷۷۴۲۴ |
| | هدیه‌ها | ۷۹۵۷۳ | ۴۰۶۵ | ۲۳۰۴۷ | ۵۱۵۵۶ |
| | اجتماعی | ۹۷۱۵۹ | ۴۲۸۴ | ۲۸۲۳۴ | ۶۳۸۳۷ |
| | کل (نسبی) | ۱۰۰ | ۴/۷۲ | ۲۸/۷۲ | ۶۵/۴۳ |

### ۵-۲- بن‌واژه‌سازی

برای بن‌واژه‌سازی، از مدل آماری معرفی‌شده توسط مولر[1] و همکاران (۲۰۱۵) که توسط قیومی (۱۳۹۸) برای زبان فارسی تهیه شده‌است استفاده کرده‌ایم. برای آموزش این مدل، از پیکرهٔ توسعه‌یافتهٔ بی‌جن‌خان (۱۳۸۳) توسط قیومی (۱۳۹۸) استفاده شده‌است. طبق بررسی‌های انجام شده توسط قیومی (۱۳۹۸)، درصد دقت مدل آماری مورد نظر برای بن‌واژه‌سازی فارسی ۹۸/۰۸ درصد است و درصد خطای آن ۱/۹۲ درصد اعلام شده‌است.

### ۵-۳- برچسب‌زنی مقولات دستوری و تجزیهٔ سازه‌ای

برچسب‌زنی مقولات دستوری به عمل انتساب برچسب‌های دستوری به واژه‌ها و نشانه‌های تشکیل‌دهندهٔ یک متن گفته می‌شود؛ به این شکل که این برچسب‌ها نشان‌دهندهٔ نقش واژگان و نشانه‌ها در جمله است. برخی از پژوهشگران حوزهٔ زبان‌شناسی پیکره‌ای، وارد مقولهٔ

1. T. Müller

نشانه‌گذاری و برچسب‌زنی نمی‌شوند و معتقدند نشانه‌گذاری نمی‌تواند بدون خطا باشد (لیچ[1]، ۲۰۰۴)؛ اما گروه دیگر نشانه‌گذاری که برای غنی‌سازی پیکرهٔ خام استفاده شود را مفید می‌دانند. برای مثال، برچسب‌گذاری مقولات دستوری انجام‌شده برروی پیکرهٔ براون[2] به‌شکل وسیعی مورد استفاده افراد مختلف قرار گرفت (علایی‌ابوذر و همکاران، ۱۴۰۰). در پژوهش حاضر، با کمک الگوریتم معرفی‌شده توسط مولر و همکاران (۲۰۱۳) داده‌های پیکرهٔ تهیه‌شده برچسب‌گذاری دستوری شده‌است. برای آموزش این مدل، از پیکرهٔ بی‌جن‌خان (۱۳۸۳) استفاده شده‌است. در جدول ۳ آمار مقولات دستوری تخصیص‌داده‌شده به واژگان پیکرهٔ حاصل از کتاب‌های درسی گزارش شده‌است.

براساس نتایج گزارش‌شده در جدول ۳، مقولهٔ دستوری اسم بالاترین میزان کاربرد، درحدود ٪۳۴ از کل واژگان موجود در پیکرهٔ حاصل از کتاب‌های درسی، را شامل شده‌است. از میان دروس مختلف، کاربرد این مقولهٔ دستوری، به‌طورکلی، در درس علوم بیشترین است و این مقوله در پایهٔ ششم بیشترین کاربرد را داشته‌است. کاربرد فعل، با ۱۴/۶۵٪ کاربرد، رتبهٔ دوم را دارا است. از میان دروس مختلف، کاربرد این مقولهٔ دستوری به‌طورکلی در درس فارسی بیشترین است و در پایهٔ پنجم از این مقوله بیشتر استفاده شده‌است. شایان ذکر است که نسبت کاربرد اسم به فعل حدود 2/5 برابر است و این نسبت تقریباً در تمامی پایه‌ها وجود دارد. کاربرد حرف اضافه در رتبهٔ سوم قرار دارد. کاربرد این مقوله در پایهٔ اول کمتر از ۱٪ است؛ ولی با افزایش پایهٔ تحصیلی به کاربرد این مقوله افزوده شده‌است. از میان دروس مختلف، کاربرد این مقولهٔ دستوری، به‌طورکلی، در درس علوم بیشترین است و در پایهٔ ششم از این مقوله دستوری بیشتر استفاده شده‌است. کاربرد حرف ربط در رتبهٔ پنجم قرار دارد. کاربرد این مقوله در پایه‌های اول و دوم کمتر از ۱٪ است و با افزایش پایه، استفاده از این مقوله افزایش یافته‌است. کاربرد کم این مقوله در پایه‌های اول و دوم بیانگر این است که ساختار عمدهٔ جملات در این دو پایه ساده است و از جملات مرکب کمتر استفاده می‌شود. کاربرد صفت، ضمیر، قید، نشانهٔ «را»، حرف تعریف و عدد در مرتبه ششم تا دهم قرار دارد. قرار گرفتن در این رتبه به این معنا است که استفاده از این مقولات در دورهٔ ابتدایی چندان متداول نیست و ساخت‌های زبانی نسبتاً ساده است. کاربرد صفت در پایه‌های اول تا سوم کمتر از ۱٪ است و در درس فارسی از صفت بیشتر استفاده شده‌است. کاربرد ضمیر، قید، نشانهٔ «را»، حرف تعریف و عدد در تمامی پایه‌ها کمتر از ۱٪ است که بیانگر این نکته است که استفاده از این مقولات دستوری در

1. G. Leech
2. Brown corpus

ساخت‌های زبانی به‌کاررفته در جملات کتاب‌های درسی چندان متداول نیست. سایر مقولات، همچون شاخص و شبه‌جمله، نیز، جزء موارد بسیار کم‌کاربرد است که کاربرد این مقوله در پیکرهٔ تهیه‌شده کمتر از ۱٪ است.

**جدول ۳: آمار مقولات دستوری تخصیص‌داده‌شده به واژگان پیکره**

| پایه | درس | اسم | فعل | علائم سجاوندی | حرف اضافه | حرف ربط | صفت | قسم | قید | نشانهٔ مفعول | تعریف حرف | عدد | شاخص | شبه‌جمله |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| اول | فارسی | ۱۴۶۵ | ۶۰۶ | ۴۴۵ | ۳۲۱ | ۲۳۴ | ۲۰۹ | ۱۲۵ | ۹۴ | ۱۲۷ | ۱۳۵ | ۹۵ | ۳ | ۲ |
| | علوم | ۱۶۶۰ | ۶۹۵ | ۴۸۰ | ۵۸۹ | ۲۹۶ | ۲۸۳ | ۲۱۵ | ۱۴۸ | ۱۵۲ | ۲۰۴ | ۶۹۵ | ۱۵ | ۰ |
| | نسبی(٪) | ۱/۵۰ | ۰/۶۳ | ۰/۴۴ | ۰/۴۴ | ۰/۲۵ | ۰/۲۴ | ۰/۱۶ | ۰/۱۲ | ۰/۱۳ | ۰/۱۶ | ۰/۳۸ | ۰/۰۱ | ۰/۰۰ |
| دوم | فارسی | ۳۷۶۰ | ۲۱۳۶ | ۲۲۰۷ | ۱۰۰۶ | ۱۰۸۳ | ۶۹۰ | ۴۹۹ | ۴۷۰ | ۳۱۶ | ۳۱۸ | ۲۹۳ | ۱۲ | ۲۳ |
| | علوم | ۳۰۲۳ | ۱۰۸۶ | ۷۶۶ | ۹۷۹ | ۵۰۹ | ۴۷۵ | ۳۶۸ | ۱۸۰ | ۳۲۵ | ۲۵۰ | ۱۷۹ | ۱۳ | ۰ |
| | هدیه‌ها | ۱۳۲۷ | ۶۴۰ | ۵۴۹ | ۳۷۹ | ۲۴۹ | ۲۰۲ | ۱۹۸ | ۱۱۲ | ۱۲۸ | ۶۷ | ۵۱ | ۶۷ | ۳ |
| | نسبی(٪) | ۳/۹۰ | ۱/۸۶ | ۱/۶۹ | ۱/۱۴ | ۰/۸۸ | ۰/۶۶ | ۰/۵۱ | ۰/۳۷ | ۰/۳۷ | ۰/۳۱ | ۰/۲۵ | ۰/۰۴ | ۰/۰۱ |
| سوم | فارسی | ۳۰۹۵ | ۱۷۶۱ | ۱۰۶۵ | ۹۸۱ | ۸۶۷ | ۷۱۶ | ۵۲۲ | ۳۵۸ | ۲۷۰ | ۱۹۲ | ۱۴۲ | ۵ | ۱۳ |
| | علوم | ۲۴۸۴ | ۹۰۲ | ۶۷۸ | ۷۹۴ | ۴۱۴ | ۳۹۷ | ۲۴۵ | ۱۶۴ | ۲۷۸ | ۲۴۱ | ۱۴۰ | ۱۷ | ۰ |
| | هدیه‌ها | ۲۵۰۶ | ۱۲۹۳ | ۱۱۸۹ | ۸۱۳ | ۶۰۰ | ۴۶۳ | ۳۷۱ | ۲۸۳ | ۲۳۳ | ۱۷۲ | ۷۰ | ۵ | ۵ |
| | اجتماعی | ۱۸۲۷ | ۸۲۴ | ۷۱۶ | ۵۵۸ | ۴۲۱ | ۲۷۳ | ۲۳۵ | ۲۷۳ | ۱۴۹ | ۲۶۲ | ۵۹ | ۷ | ۰ |
| | نسبی(٪) | ۴/۷۶ | ۲/۳۰ | ۱/۷۵ | ۱/۵۱ | ۱/۱۱ | ۰/۸۹ | ۰/۶۶ | ۰/۵۲ | ۰/۴۵ | ۰/۴۲ | ۰/۲۰ | ۰/۰۲ | ۰/۰۱ |
| چهارم | فارسی | ۳۶۸۰ | ۲۲۸۲ | ۱۲۷۱ | ۱۲۵۵ | ۱۱۱۳ | ۸۵۲ | ۶۲۹ | ۵۶۶ | ۴۰۱ | ۳۱۱ | ۱۸۷ | ۹ | ۱۲ |
| | علوم | ۵۰۲۶ | ۱۸۰۲ | ۱۳۷۵ | ۱۵۹۱ | ۸۰۳ | ۱۰۰۰ | ۵۰۰ | ۳۰۹ | ۵۲۵ | ۴۴۸ | ۴۳۶ | ۱۸ | |
| | هدیه‌ها | ۲۷۷۶ | ۱۳۹۵ | ۱۳۶۸ | ۸۸۹ | ۶۷۶ | ۵۰۸ | ۳۷۳ | ۳۳۷ | ۲۴۷ | ۲۰۲ | ۹۳ | ۳ | ۸ |
| | اجتماعی | ۳۴۸۴ | ۱۲۸۴ | ۱۱۱۳ | ۱۰۶۳ | ۷۶۳ | ۶۶۵ | ۳۳۲ | ۲۹۷ | ۱۸۰ | ۲۶۰ | ۱۳۲ | ۷ | ۱ |
| | نسبی(٪) | ۷/۱۹ | ۳/۲۵ | ۲/۴۶ | ۲/۳۱ | ۱/۶۱ | ۱/۴۵ | ۰/۸۸ | ۰/۷۳ | ۰/۶۵ | ۰/۵۹ | ۰/۴۱ | ۰/۰۲ | ۰/۰۱ |
| پنجم | فارسی | ۳۸۹۸ | ۲۰۵۶ | ۷۶۷ | ۱۳۶۸ | ۱۲۳۷ | ۷۸۱ | ۶۴۶ | ۴۱۷ | ۴۰۰ | ۳۵۳ | ۱۱۱ | ۸ | ۶ |
| | علوم | ۳۵۲۷ | ۱۵۴۷ | ۱۲۱۸ | ۱۴۱۴ | ۷۲۴ | ۸۵۶ | ۴۳۴ | ۳۱۹ | ۵۳۰ | ۴۲۰ | ۴۹۳ | ۴۱ | ۱ |
| | هدیه‌ها | ۴۵۴۴ | ۱۶۶۵ | ۱۵۶۱ | ۱۰۴۵ | ۸۷۲ | ۶۵۴ | ۴۵۸ | ۳۱۱ | ۳۱۶ | ۲۲۶ | ۱۷۱ | ۷ | ۱۴ |
| | اجتماعی | ۵۶۸۱ | ۱۸۰۷ | ۱۹۲۷ | ۱۷۱۵ | ۱۲۷۸ | ۱۰۱۰ | ۵۱۴ | ۳۰۱ | ۲۹۰ | ۳۷۳ | ۱۸۴ | ۱۰ | ۱ |
| | نسبی(٪) | ۸/۴۸ | ۳/۴۰ | ۲/۶۳ | ۲/۶۶ | ۱/۹۸ | ۱/۵۹ | ۰/۹۹ | ۰/۶۵ | ۰/۷۴ | ۰/۶۶ | ۰/۴۶ | ۰/۰۳ | ۰/۰۱ |
| ششم | فارسی | ۳۲۷۶ | ۱۷۷۹ | ۸۶۲ | ۱۱۵۸ | ۱۱۰۴ | ۶۳۷ | ۵۶۴ | ۳۶۵ | ۲۵۳ | ۲۹۴ | ۱۶۶ | ۱۰ | ۸ |
| | علوم | ۳۵۸۱ | ۱۲۹۸ | ۱۵۸۷ | ۱۱۴۴ | ۱۱۴۴ | ۷۴۷ | ۳۱۷ | ۲۵۷ | ۲۶۷ | ۲۹۸ | ۲۳۵ | ۱۸ | ۰ |
| | هدیه‌ها | ۳۶۷۲ | ۱۵۴۸ | ۸۲۸ | ۱۲۱۳ | ۹۲۲ | ۶۴۰ | ۴۹۵ | ۳۰۰ | ۲۴۷ | ۲۶۰ | ۱۳۸ | ۵ | ۱۰ |
| | اجتماعی | ۶۴۸۳ | ۲۰۷۲ | ۱۷۹۰ | ۱۹۶۱ | ۱۴۸۱ | ۱۲۹۱ | ۵۶۲ | ۴۷۰ | ۳۱۰ | ۵۱۸ | ۲۶۹ | ۲۰ | ۰ |
| | نسبی(٪) | ۸/۱۸ | ۳/۲۲ | ۲/۴۴ | ۲/۶۳ | ۲/۲۴ | ۱/۵۹ | ۰/۹۳ | ۰/۶۷ | ۰/۵۲ | ۰/۶۶ | ۰/۳۹ | ۰/۰۳ | ۰/۰۱ |
| اول تا ششم | فارسی | ۹/۲۲ | ۵/۱۰ | ۳/۱۸ | ۲/۹۳ | ۲/۷۱ | ۱/۸۷ | ۱/۴۳ | ۱/۰۹ | ۰/۸۵ | ۰/۷۷ | ۰/۴۸ | ۰/۰۲ | ۰/۰۳ |
| | علوم | ۹/۲۸ | ۳/۵۲ | ۲/۹۳ | ۳/۱۳ | ۱/۸۷ | ۱/۸۱ | ۱/۰۰ | ۰/۶۶ | ۱/۰۰ | ۰/۸۹ | ۱/۰۵ | ۰/۰۶ | ۰/۰۰ |
| | هدیه‌ها | ۷/۱۳ | ۳/۱۴ | ۲/۶۴ | ۲/۰۹ | ۱/۶۰ | ۱/۱۹ | ۰/۹۱ | ۰/۶۵ | ۰/۵۶ | ۰/۴۵ | ۰/۲۵ | ۰/۰۴ | ۰/۰۲ |
| | اجتماعی | ۸/۴۰ | ۲/۸۸ | ۲/۶۷ | ۲/۵۵ | ۱/۹۰ | ۱/۵۶ | ۰/۷۹ | ۰/۶۴ | ۰/۴۵ | ۰/۶۸ | ۰/۳۱ | ۰/۰۲ | ۰/۰۰ |
| | نسبی(٪) | ۳۴/۰۲ | ۱۴/۶۵ | ۱۱/۴۲ | ۱۰/۶۹ | ۸/۰۷ | ۶/۴۲ | ۴/۱۳ | ۳/۰۴ | ۲/۸۶ | ۲/۷۹ | ۲/۰۹ | ۰/۱۴ | ۰/۰۵ |

مقولات دستوری را از منظر دیگری بررسی کرده‌ایم و واژه‌ها را به واژه‌های محتوایی و نقشی تقسیم کرده‌ایم. واژه‌های محتوایی به واژگانی گفته می‌شود که بار معنایی جملات را به عهده دارد و دارای مفهوم مستقلی است، مانند اسم، صفت، فعل، قید و طبقه بازی را تشکیل

می‌دهد که می‌توان یک واژه به این مجموعه اضافه یا کم کرد. واژه‌های نقشی برای برقراری ارتباط میان سایر واحدهای زبانی که فاقد مفهوم مستقل است به کار می‌رود که شامل مواردی چون حرف اضافه، حرف تعریف، حرف ربط و شاخص، است. در جدول ۴ توزیع آماری این واژه‌ها به تفکیک دروس و پایه‌ها گزارش شده‌است.

**جدول۴: توزیع آماری واژه‌های محتوایی و نقشی**

| پایه | درس | واژه‌های محتوایی | واژه‌های نقشی | نسبت واژه‌های محتوایی به واژگان | نسبت واژه‌های نقشی به واژگان |
|---|---|---|---|---|---|
| اول | فارسی | ۲۵۴۹ | ۱۳۶۸ | ۱/۲۳ | ۰/۶۶ |
| | علوم | ۳۰۱۶ | ۱۷۹۴ | ۱/۴۵ | ۰/۸۶ |
| | کل | ۵۵۶۵ | ۳۱۶۲ | ۲/۶۷ | ۱/۵۲ |
| دوم | فارسی | ۷۶۱۵ | ۵۲۹۶ | ۳/۶۶ | ۲/۵۵ |
| | علوم | ۵۱۴۸ | ۳۰۰۸ | ۲/۴۷ | ۱/۴۵ |
| | هدیه‌ها | ۲۴۹۱ | ۱۴۵۲ | ۱/۲۰ | ۰/۷۰ |
| | کل | ۱۵۲۵۴ | ۹۷۵۶ | ۷/۳۳ | ۴/۶۹ |
| سوم | فارسی | ۶۴۷۱ | ۳۶۰۰ | ۳/۱۱ | ۱/۷۳ |
| | علوم | ۴۲۱۱ | ۲۵۵۳ | ۲/۰۲ | ۱/۲۳ |
| | هدیه‌ها | ۴۹۰۰ | ۳۱۱۳ | ۲/۳۶ | ۱/۵۰ |
| | اجتماعی | ۳۳۳۳ | ۲۱۰۲ | ۱/۶۰ | ۱/۰۱ |
| | کل | ۱۸۹۱۵ | ۱۱۳۶۸ | ۹/۰۹ | ۵/۴۶ |
| چهارم | فارسی | ۸۰۳۱ | ۴۵۴۸ | ۳/۸۶ | ۲/۱۹ |
| | علوم | ۸۶۹۳ | ۵۱۸۷ | ۴/۱۸ | ۲/۴۹ |
| | هدیه‌ها | ۵۳۷۳ | ۳۵۱۹ | ۲/۵۸ | ۱/۶۹ |
| | اجتماعی | ۶۰۷۱ | ۳۵۲۱ | ۲/۹۲ | ۱/۶۹ |
| | کل | ۲۸۱۶۸ | ۱۶۷۷۵ | ۱۳/۵۴ | ۸/۰۶ |
| پنجم | فارسی | ۷۸۱۶ | ۴۲۵۱ | ۳/۷۶ | ۲/۰۴ |
| | علوم | ۷۸۶۰ | ۴۸۳۲ | ۳/۷۸ | ۲/۳۲ |
| | هدیه‌ها | ۶۵۱۹ | ۴۲۰۸ | ۳/۱۳ | ۲/۰۲ |
| | اجتماعی | ۹۳۳۱ | ۵۷۷۶ | ۴/۴۹ | ۲/۷۸ |
| | کل | ۳۱۵۲۶ | ۱۹۰۶۷ | ۱۵/۱۵ | ۹/۱۶ |

**ادامه جدول۴:**

| پایه | درس | واژه‌های محتوایی | واژه‌های نقشی | نسبت واژه‌های محتوایی به واژگان | نسبت واژه‌های نقشی به واژگان |
|---|---|---|---|---|---|
| ششم | فارسی | ۶۶۴۰ | ۳۸۷۵ | ۳/۱۹ | ۱/۸۶ |
| | علوم | ۶۲۲۴ | ۳۸۷۵ | ۲/۹۹ | ۱/۸۶ |
| | هدیه‌ها | ۶۶۷۲ | ۴۳۸۱ | ۳/۲۱ | ۲/۱۱ |
| | اجتماعی | ۱۰۸۹۸ | ۶۳۴۲ | ۵/۲۴ | ۳/۰۵ |
| | کل | ۳۰۴۳۴ | ۱۸۴۷۳ | ۱۴/۶۳ | ۸/۸۸ |
| اول تا ششم | فارسی | ۳۹۱۲۲ | ۲۲۹۳۸ | ۱۸/۸۰ | ۱۱/۰۳ |
| | علوم | ۳۵۱۵۲ | ۲۰۸۳۱ | ۱۶/۹۰ | ۱۰/۰۱ |
| | هدیه‌ها | ۲۵۹۵۵ | ۱۶۶۷۳ | ۱۲/۴۸ | ۸/۰۱ |
| | اجتماعی | ۲۹۶۳۳ | ۱۷۷۴۱ | ۱۴/۲۴ | ۸/۵۳ |
| | کل | ۱۲۹۸۶۲ | ۷۸۱۸۳ | ۶۲/۴۲ | ۳۷/۵۸ |

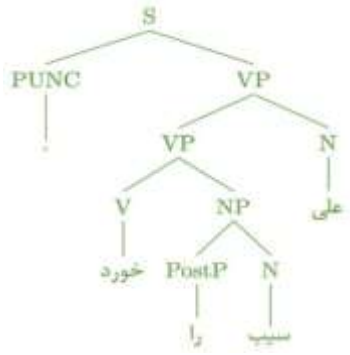
با توجه به نتایج گزارش‌شده در جدول ۴، در مجموع، واژه‌های محتوایی ۶۲٪ از واژگان این پیکره را تشکیل داده‌است که درصد قابل توجهی است؛ از طرفی، واژه‌های نقشی حدود ۳۸٪ از واژگان این پیکره را شامل می‌شود. از مجموع ۶۲٪ واژه‌های محتوایی در میان دروس مختلف، درس فارسی بالاترین میزان کاربرد واژه‌های محتوایی (حدود ۱۸٪) را داراست و علوم، مطالعات اجتماعی و هدیه های آسمانی به ترتیب ۱۶٪، ۱۴٪ و ۱۲٪ را به خود اختصاص داده‌است. نکتهٔ جالب این که در استفاده از واژه‌های نقشی، دروس مختلف به همین شکل رفتار کرده و درس فارسی بیشترین واژه‌های نقشی را دارد.

میزان استفاده از واژه‌های محتوایی و نقشی در پایهٔ اول، به‌ترتیب، حدود ۳٪ و ۱/۵٪ است که حاکی از کاربرد ساخت‌های ساده‌تر زبانی است. در پایهٔ پنجم، حدود ۱۵٪ و ۹٪ از پیکره، به‌ترتیب، متشکل از واژه‌های محتوایی و نقشی است که درمیان پایه‌ها بیشترین است. همچنین، کاربرد واژگان محتوایی در درس علوم و فارسی پایهٔ چهارم نسبت‌به سایر پایه‌ها در این درس بیشتر بود و کاربرد این دسته از واژه‌ها در درس علوم بیشتر بود.

در ادامهٔ تحلیل نحوی پیکرهٔ تهیه‌شده، فاصلهٔ هر واژه در یک جمله تا ریشهٔ درخت سازه‌ای جمله که با اصطلاح «عمق درخت» معرفی می‌کنیم را محاسبه می‌کنیم. برای این هدف، ابتدا با کمک تجزیه‌گر نحوی سازه‌ای فارسی که توسط قیومی (۲۰۱۴) تهیه شده‌است درخت سازه‌ای جملات ترسیم شده و سپس با تقطیع درخت هر جمله و یافتن گره‌های میانی مابین برگ‌های درخت تا ریشهٔ درخت، عمق درخت را محاسبه می‌کنیم. با تجمیع عمق واژه‌ها تا ریشه، عمق جمله محاسبه می‌شود. هرقدر عمق درخت جمله بیشتر باشد به این مفهوم است

که گره‌های میانی بیشتر بوده و جمله از پیچیدگی بیشتری برخوردار است. برای محاسبهٔ عمق جمله، مثال (۱) را در نظر بگیرید:

(۱) علی سیب را خورد.




در مثال ۱، می‌توان رابطهٔ میان هر گره برگ تا ریشه را به‌صورت جدول ۵ بازنمایی نمود. براساس تجمیع تعداد گره‌ها، عمق جمله در این مثال ۱۹ است.

**جدول ۵: بازنمایی مسطح درخت سازه‌ای مثال (۱)**

| واژه | فاصلهٔ واژه تا ریشه | تعداد گره‌ها |
|---|---|---|
| علی | (S (VP ( N *))) | ۳ |
| سیب | (S (VP (VP (NP (N *))))) | ۵ |
| را | (S (VP (VP (NP (PostP *))))) | ۵ |
| خورد | (S (VP (VP (V *))) | ۴ |
| ۰ | (S (PUNC *)) | ۲ |

عمق جملات به تفکیک دروس و پایه در جدول ۶ گزارش شده‌است. براساس نتایج گزارش‌شده، عمق جملات در پایه‌های اول تا سوم بین ۲۰ تا ۲۵ گره است؛ ولی در پایهٔ چهارم، افزایش یک و نیم برابری در عمق جمله تا ۴۲ گره را شاهد هستیم که بیانگر افزایش پیچیدگی در جملات است. در پایه‌های پنجم و ششم تعداد گره‌های میانی به ۳۱ و ۳۴ گره کاهش یافته‌است. اگر پایه‌های پنجم و ششم را با یکدیگر مقایسه کنیم، شاهد افزایش حدود ۳ گره را در جملات پایه ششم هستیم. براساس نتایج جدول ۶ می‌توان دید کلی‌تر از پیچیدگی جملات را به‌دست می‌آورد. براساس این نتایج، تعداد متوسط گره‌های جملات ۲۹ گره است و ازمیان دروس مختلف، جملات درس علوم کمترین و جملات درس مطالعات اجتماعی بیشترین تعداد گره‌های میانی را دارد که این نتیجه در راستا با طول متوسط جمله قرار دارد.

**جدول ۶: فاصلهٔ واژه تا ریشهٔ درخت (عمق جمله)**

| پایه | درس | فاصله واژه تا ریشهٔ جمله | توزیع نسبی عمق درخت |
|---|---|---|---|
| اول | فارسی | ۷۹۴۳ | ۱۹/۸۶ |
| | علوم | ۹۷۴۱ | ۲۱/۶۰ |
| | کل | ۱۷۶۸۴ | ۲۰/۷۸ |
| دوم | فارسی | ۲۶۴۴۲ | ۲۷/۲۳ |
| | علوم | ۱۶۶۹۶ | ۲۲/۲۶ |
| | هدیه‌ها | ۸۰۱۶ | ۲۳/۳۰ |
| | کل | ۵۱۱۵۴ | ۲۴/۷۷ |
| سوم | فارسی | ۲۰۶۰۸ | ۲۶/۵۲ |
| | علوم | ۱۳۹۲۱ | ۲۲/۶۴ |
| | هدیه‌ها | ۱۶۳۸۹ | ۲۵/۸۹ |
| | اجتماعی | ۱۱۱۰۰ | ۲۶/۱۲ |
| | کل | ۶۲۰۱۸ | ۲۵/۳۱ |
| چهارم | فارسی | ۲۵۹۴۰ | ۲۸/۷۹ |
| | علوم | ۲۸۳۲۳ | ۲۳/۴۹ |
| | هدیه‌ها | ۱۸۲۹۷ | ۲۷/۸۱ |
| | اجتماعی | ۱۹۳۵۰ | ۳۰/۸۱ |
| | کل | ۹۱۹۱۰ | ۴۲/۹۵ |
| پنجم | فارسی | ۲۴۸۵۱ | ۳۴/۰۴ |
| | علوم | ۲۶۰۵۳ | ۲۴/۴۴ |
| | هدیه‌ها | ۲۲۱۳۵ | ۳۲/۸۴ |
| | اجتماعی | ۳۰۷۱۵ | ۳۷/۰۱ |
| | کل | ۱۰۳۷۵۴ | ۳۱/۴۴ |
| ششم | فارسی | ۲۱۶۵۳ | ۳۳/۹۴ |
| | علوم | ۱۹۶۸۹ | ۲۹/۸۳ |
| | هدیه‌ها | ۲۲۶۶۲ | ۳۵/۳۵ |
| | اجتماعی | ۳۴۸۱۷ | ۳۶/۲۷ |
| | کل | ۹۸۸۲۱ | ۳۴/۰۹ |
| اول تا ششم | فارسی | ۱۲۷۴۳۷ | ۲۸/۴۰ |
| | علوم | ۱۱۴۴۲۳ | ۲۴/۰۴ |
| | هدیه‌ها | ۷۹۴۸۳ | ۳۰/۴۷ |
| | اجتماعی | ۸۴۸۸۲ | ۳۴/۷۰ |
| | کل | ۴۰۶۲۲۵ | ۲۹/۴۰ |

### ۳-۴- کلمات پربسامد پیکره

جدول ۷ تعداد ۵۰ صورت‌واژهٔ پربسامد که از واژگان حاصل از تجمیع پایه‌های مختلف به‌دست آمده‌است گزارش شده‌است. این واژگان با واژگان جداگانهٔ حاصل از هر پایه مقایسه شده‌است و نتیجهٔ مقایسه در این جدول قابل‌مشاهده است. ازمیان این ۵۰ واژهٔ پربسامد، ۲۸ واژه (۵۶٪) در تمام پایه‌ها مشترک بود که می‌تواند به‌عنوان واژه‌های پایه معرفی گردد. پایهٔ اول کمترین و پایهٔ پنجم بیشترین پوشش را با واژه‌های پربسامد این واژگان داشت.

نکتهٔ دیگر این که در بعضی از پایه‌ها واژه‌هایی وجود داشت که در سایر پایه‌ها وجود نداشت و یا از یک پایه به بعد به‌عنوان واژهٔ پربسامد تلقی شده‌است. برای مثال، حرف اضافهٔ «بر» از پایهٔ سوم به بعد جزء واژگان پربسامد لحاظ شده‌است. مثال بعدی واژهٔ «مردم» است که از پایهٔ چهارم به بعد جزء واژگان پربسامد است. «ایران» واژه‌ای است که در پایه‌های اول، پنجم و ششم جزء واژه‌های پربسامد است و در پایه‌های دوم تا چهارم نیست. مثال دیگر واژهٔ «چند» است که به‌عنوان پرسش‌واژه در پایهٔ اول پرکاربردترین است و در سایر پایه‌ها کاربرد ندارد. واژهٔ «شد» که معمولاً در ساخت مجهول کاربرد دارد از پایهٔ دوم و واژهٔ «اگر» که در ساخت شرطی به‌کار می‌رود از پایهٔ سوم وارد واژگان شده‌است.

**جدول ۷: مقایسهٔ واژه‌های پربسامد واژگان در پایه‌های مختلف**

| ردیف | واژه‌های پربسامد | پایهٔ اول | پایهٔ دوم | پایهٔ سوم | پایهٔ چهارم | پایهٔ پنجم | پایهٔ ششم | مشترک پایه‌ها |
|---|---|---|---|---|---|---|---|---|
| | | ۳۶ (۷۲٪) | ۳۸ (۷۶٪) | ۴۳ (۸۶٪) | ۴۳ (۸۶٪) | ۴۴ (۸۸٪) | ۴۲ (۸۴٪) | ۲۸ (۵۶٪) |
| ۱ | و | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲ | را | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳ | به | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۴ | از | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۵ | در | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۶ | که | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۷ | با | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۸ | است | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۹ | این | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۰ | آن | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۱ | برای | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |

**ادامه جدول ۷:**

| ردیف | واژه‌های پربسامد | پایهٔ اول | پایهٔ دوم | پایهٔ سوم | پایهٔ چهارم | پایهٔ پنجم | پایهٔ ششم | مشترک پایه‌ها |
|---|---|---|---|---|---|---|---|---|
| | | ۳۶ (۷۲٪) | ۳۸ (۷۶٪) | ۴۳ (۸۶٪) | ۴۳ (۸۶٪) | ۴۴ (۸۸٪) | ۴۲ (۸۴٪) | ۲۸ (۵۶٪) |
| ۱۲ | خود | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۳ | کنید | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۴ | یک | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۵ | چه | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۱۶ | او | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۱۷ | هم | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۱۸ | بود | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۱۹ | ما | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۰ | تا | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۱ | گفت | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۲۲ | آن-ها | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۳ | من | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۴ | کرد | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۲۵ | هر | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۶ | آب | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۷ | شما | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۸ | می‌کنند | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۲۹ | یا | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳۰ | می‌شود | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳۱ | روی | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳۲ | دارد | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳۳ | بر | ۰ | ۰ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۳۴ | می‌کند | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |
| ۳۵ | روز | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ | ۰ |
| ۳۶ | شد | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۳۷ | مردم | ۰ | ۰ | ۰ | ۱ | ۱ | ۱ | ۰ |
| ۳۸ | زندگی | ۱ | ۰ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۳۹ | دیگر | ۰ | ۰ | ۱ | ۱ | ۰ | ۱ | ۰ |

**ادامه جدول ۷:**

| ردیف | واژه‌های پربسامد | پایهٔ اول | پایهٔ دوم | پایهٔ سوم | پایهٔ چهارم | پایهٔ پنجم | پایهٔ ششم | مشترک پایه‌ها |
|---|---|---|---|---|---|---|---|---|
| | | ۳۶ (۷۲٪) | ۳۸ (۷۶٪) | ۴۳ (۸۶٪) | ۴۳ (۸۶٪) | ۴۴ (۸۸٪) | ۴۲ (۸۴٪) | ۲۸ (۵۶٪) |
| ۴۰ | شده | ۰ | ۰ | ۱ | ۱ | ۱ | ۱ | ۰ |
| ۴۱ | نه | ۱ | ۱ | ۰ | ۰ | ۱ | ۱ | ۰ |
| ۴۲ | زیر | ۰ | ۰ | ۱ | ۱ | ۱ | ۰ | ۰ |
| ۴۳ | ایران | ۱ | ۰ | ۰ | ۰ | ۱ | ۱ | ۰ |
| ۴۴ | چند | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ |
| ۴۵ | اگر | ۰ | ۰ | ۱ | ۰ | ۱ | ۱ | ۰ |
| ۴۶ | تو | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ | ۰ |
| ۴۷ | کار | ۰ | ۱ | ۰ | ۱ | ۰ | ۰ | ۰ |
| ۴۸ | دو | ۱ | ۰ | ۰ | ۰ | ۱ | ۰ | ۰ |
| ۴۹ | انجام | ۱ | ۱ | ۱ | ۰ | ۰ | ۰ | ۰ |
| ۵۰ | استفاده | ۱ | ۰ | ۰ | ۱ | ۰ | ۰ | ۰ |

برای مطالعهٔ دقیق‌تر واژه‌های پربسامد، بین واژه‌های محتوایی و نقشی تمایز ایجاد کردیم و یک فهرست از ۲۰۰ واژهٔ نقشی پربسامد و فهرست‌های دیگری از ۲۰۰، ۵۰۰ و ۱۰۰۰ صورت‌واژهٔ محتوایی پربسامد تهیه کردیم و واژه‌های مشترک با صورت‌واژه‌های جدول ۷ را در پایه‌های مختلف بررسی کردیم. در جدول ۸ تعداد واژه‌های پربسامد محتوایی و نقشی در پایه‌های مختلف گزارش شده‌است.

**جدول ۸: مقایسهٔ واژه‌های پربسامد محتوایی و نقشی در پایه‌های مختلف**

| نوع واژه | واژه‌های پربسامد | پایهٔ اول | پایهٔ دوم | پایهٔ سوم | پایهٔ چهارم | پایهٔ پنجم | پایهٔ ششم |
|---|---|---|---|---|---|---|---|
| نقشی | ۲۰۰ | ۶۷ | ۱۱۹ | ۱۱۰ | ۱۲۳ | ۱۲۱ | ۱۲۵ |
| محتوایی | ۲۰۰ | ۱۹۰ | ۱۹۹ | ۱۹۹ | ۲۰۰ | ۱۹۹ | ۲۰۰ |
| | ۵۰۰ | ۴۱۱ | ۴۸۲ | ۴۹۲ | ۴۹۶ | ۴۹۳ | ۴۹۶ |
| | ۱۰۰۰ | ۶۷۵ | ۸۸۵ | ۹۳۶ | ۹۷۹ | ۹۷۸ | ۹۶۴ |

براساس نتایج گزارش‌شدهٔ مربوط به واژه‌های نقشی، در پایهٔ اول، واژه‌های نقشی محدودی موجود است. این میزان در پایه‌های دوم و سوم و همچنین در پایه‌های چهارم تا ششم تقریباً یکسان است. همچنین براساس نتایج گزارش‌شدهٔ مربوط به واژه‌های محتوایی، با احتساب ۲۰۰

صورت‌واژهٔ پربسامد، پایهٔ اول حاوی ۹۵٪ این واژه‌ها بود و این واژه‌ها تقریباً در سایر پایه‌ها به‌کار رفته‌است. با افزایش واژه‌های محتوایی به ۵۰۰ واژه، ۸۲/۲٪ از واژه‌ها در پایهٔ اول، ۹۶/۴٪ از واژه‌ها در پایهٔ دوم و در سایر پایه‌ها تقریباً تمامی واژه‌های محتوایی پربسامد دیده شده‌است. با افزایش مجدد صورت‌واژه‌های محتوایی، پایهٔ اول ۶۷/۵٪، پایهٔ دوم ۸۸/۵٪، پایهٔ سوم ۹۳/۶٪ و سایر پایه‌ها حدود ۹۷٪ از ۱۰۰۰ واژهٔ بربسامد را شامل شده‌است. نکتهٔ قابل توجه این است که از پایهٔ سوم تا ششم دایرهٔ واژگانی دانش‌آموزان در حال تکمیل است و از پایهٔ چهارم تا ششم تقریباً این دایرهٔ واژگانی ثابت است و تلاش می‌شود واژه‌ها در بافت‌های زبانی متفاوت به‌کار برده شود تا در ذهن دانش‌آموز تثبیت گردد.

## ۵- جمع‌بندی

در این پژوهش، به تهیهٔ یک پیکرهٔ زبانی از کتاب‌های درسی پرداختیم تا با استفاده از آن بتوانیم شناختمان را نسبت‌به محتوای درسی کتاب‌های آموزشی پایه‌های اول تا ششم بیشتر کرده و ضمن آگاهی از نیازها و اقدام جهت رفع کاستی‌ها، امکان استفاده از دستاوردهای پژوهشی آن را در حوزهٔ آموزش زبان و سیاست‌گذاری در حوزهٔ زبان مهیا نماییم. محتوای این پیکره، شامل دروس مختلف فارسی، علوم، مطالعات اجتماعی و هدیه‌های آسمانی پایه‌های اول تا ششم ابتدایی بود که پس‌از جمع‌آوری داده و تهیهٔ پیکرهٔ زبانی هدف، وارد مرحلهٔ برچسب‌گذاری در سطوح آوایی و بررسی الگوی هجایی واژه‌ها ، بن‌واژه‌ای، تعیین مقولات دستوری واژه‌ها، تجزیهٔ نحوی جملات و محاسبهٔ عمق جمله شدیم که تمامی مراحل به‌صورت کاملاً خودکار انجام پذیرفت.

باتوجه‌به چشم‌انداز سند علمی کشور که زبان فارسی به عنوان زبان علمی در رشته‌های علمی مختلف انتخاب و استفاده شده و همچنین باتوجه‌به شناخت از مسائل زبانی که برای برنامه‌ریزی کاربرد دارد به این جمع‌بندی رسیدیم که در نگاه کلان به محتوای درسی، اصول کلی در تنظیم محتوا از نظر ساخت نحوی، الگوهای آوایی و انتخاب نوع واژه‌ها رعایت شده‌است اما در جزئیات، کمی جای تأمل وجود دارد و نیاز است بعضی از محتوای زبانی پایه‌ها بازبینی شود. برای مثال، قسمتی از ویژگی‌هایی که از نظر حجم محتوا، واژگان و پیچیدگی‌های زبانی که در پایهٔ پنجم وجود دارد به پایهٔ ششم منتقل شود تا همزمان با فرایند رشد شناختی دانش‌آموزان، یادگیری محتوای درسی پیچیده‌تر اتفاق بیفتد. همچنین در محتوای درسی پایه‌های چهارم تا ششم، حدود ۱۰۰۰ صورت‌واژه می‌تواند به‌عنوان واژگان پایه تلقی گردد و تلاش شود این واژه‌ها در دروس مختلف سه پایهٔ متأخر تکرار شود تا در ذهن دانش‌آموز تثبیت گردد.

**تشکر و قدردانی**

این پژوهش براساس طرح با کد پیگیری ۱۱۲۴۳ توسط ستاد علوم و فناوری‌های شناختی ایران مورد حمایت قرار گرفته‌است.

**منابع**

اسلامی، محرم؛ شریفی آتشگاه، مسعود؛ علیزاده لمجیری، صدیقه؛ و زندی، طاهره (۱۳۸۳). «واژگان زایای زبان فارسی». *مجموعه مقالات اولین کارگاه زبان فارسی و رایانه*. دانشگاه تهران، ۶-۱۱.

بی‌جن‌خان، محمود (۱۳۸۳). «نقش پیکره زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار». *مجله زبان‌شناسی*.سال نوزدهم، شماره ۳۸، ۴۸-۶۷.

بابادی، امین؛ غیاث‌نژادعمران، پویا؛ و قاسم‌ثانی، غلامرضا (۱۳۹۱). «استفاده از یادگیری ماشین در ریشه‌یابی کلمات فارسی». *مجموعه مقالات هجدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران*، دانشگاه صنعتی شریف، دانشکده مهندسی رایانه، تهران، ایران.

پاهنگ، نظام‌الدین؛ مهدیون، روح اله؛ و یاریقلی، بهبود (۱۳۹۶). «بررسی کیفیت مدارس و شناسایی عوامل موثر بر آن: پژوهشی ترکیبی». *دوفصلنامه علمی پژوهشی مدیریت مدرسه*. ۵ (۱): ۱۷۳-۱۹۳.

تشکری، مسعود؛ و میبدی، محمدرضا (۱۳۸۰). «طراحی یک ریشه‌یاب خودکار برای واژگان فارسی». *مجموعه مقالات هفتمین کنفرانس ملی سالانه انجمن کامپیوتر ایران*، دانشگاه صنعتی شریف، دانشکده مهندسی رایانه، تهران، ایران.

دهقان، محمدحسین؛ ملاعباسی، محمد؛ فیلی، هشام؛ و شاکری، آزاده (۱۳۹۶). «تولید درخت‌بانک سازه‌ای زبان فارسی به روش نیمه‌خودکار». در *مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی*، نشر نویسه پارسی، صص: ۶۳-۸۲.

رجبی، ابوالفضل؛ و احمدوند، شجاع (۱۴۰۱). «سیاست‌گذاری زبانی و جایگاه زبان فارسی در سیاست‌های زبانی جمهوری اسلامی ایران»، *نشریه مطالعات ملی*، ۲۳ (۱): ۵۵-۷۷.

صدری افشار، غلامحسین؛ حکمی، نسترن؛ و حکمی، نسرین (۱۳۸۱). *فرهنگ فارسی*. تهران: فرهنگ معاصر.

طباطبایی، شهره؛ و صراف رضایی، ایمان (۱۳۹۶). «پیکره سازه: درختبانک بزرگ زبان فارسی در دستور سازه‌ای». در *مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی*، نشر نویسه پارسی، صص: ۴۱-۶۱.

عاصی، مصطفی (۱۳۷۶). «پایگاه داده‌های زبان فارسی». *مجموعه مقالات سومین کنفرانس زبان‌شناسی*، دانشگاه علامه طباطبایی و پژوهشگاه علوم انسانی و مطالعات فرهنگی.

عاصی، مصطفی؛ و قندی، سعیده (۱۳۹۴). «پایگاه داده‌های زبان فارسی و پیکره تاریخی آن». در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، نشر نویسه پارسی، ۱۹۳-۲۲۰.

عاصی، مصطفی؛ و ترابی، منیره (۱۳۹۱). «ارائه و معرفی پیکره‌ای برای فارسی‌آموزان خارجی». مجموعه مقالات دانشگاه علامه طباطبایی، جلد دوم، ۵۰۵-۵۱۶، ش ۲۸۱.

علایی‌ابوذر، الهام؛ پاک‌نیت، نصراله؛ حجت‌پناه، علی‌اصغر؛ زالی، مجتبی؛ و آقالویی آغمیونی، محمدهادی (۱۴۰۰). «معرفی یک پیکرهٔ متنی تخصصی: پیکرهٔ پژوهش‌نامه». *نشریه پژوهش‌های زبان‌شناسی تطبیقی*، ۱۱(۲۲): ۲۷۱-۲۸۹.

علایی‌ابوذر، الهام (۱۳۹۹). «بررسی امکان افزایش صحت یک ابزار برچسب‌دهی به اجزای کلام در فارسی». *نشریه پژوهش‌های زبان‌شناسی تطبیقی*، ۱۰(۱۹): ۹۵-۱۱۰.

غریبی، افسانه (۱۳۹۱). «واکاوی نیازهای عمومی زبانی فارسی‌آموزان». *پژوهش‌نامهٔ آموزش زبان فارسی به غیرفارسی زبانان*. ۱(۱): ۶۱-۷۸.

فرزانه‌فر، حامد (۱۳۸۹). *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی اصفهان، دانشکده برق و رایانه، تهران، ایران.

قیومی، مسعود (۱۳۸۳). *پیش‌بینی واژه در پردازش رایانه‌ای زبان فارسی*. پایان‌نامه کارشناسی ارشد دانشگاه آزاد اسلامی، واحد تهران مرکز.

قیومی، مسعود (۱۳۹۸). «گذار از بن‌واژه‌سازی قاعده‌مند به آماری در فارسی». *مجموعه مقالات پنجمین همایش ملی زبان‌شناسی رایانشی*. تهران: نشر نویسه پارسی.

قیومی، مسعود (۱۴۰۱). «ارزیابی ساختار هرم وارونه در پیکرهٔ بزرگ خبری فارسی: تحلیل گفتمان خبری براساس همبستگی میان عنوان و محتوای خبر». *مجله زبان و زبان‌شناسی*. ۱۸(۳۵): ۲۱-۴۵.

محسنی، مهدی. (۱۳۸۶) *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه علم و صنعت، دانشکده رایانه، تهران، ایران.

AleAhmad, A.; Amiri, H.; Darrudi, E.; Rahgozar, M.; and Oroumchian, F. (2009). "Hamshahri: A standard Persian text collection". *Knowledge- Based Systems*, 22, 382–387.

Amiri, H.; Hojjat, H.; and Oroumchian, F. (2007). "Investigation on a Feasible Corpus for Persian POS Tagging". In *Proceedings of the 12th International CSI Computer Conference*, Iran.

Assi, M. (1997). "Farsi linguistic database (FLDB)". *International Journal of Lexicography*, 10(3): 5.

Assi, M.; and HajiAbdolhosseini, M. (2000). "Grammatical tagging of a Persian corpus". *International Journal of Corpus Linguistics*, 5(1):69–82.

Atkins, S.; Clear, J.; and Ostler, N. (1992). "Corpus design criteria". *Literary and Linguistic Computing*, 7(1), 1–16.

Bijankhan, M.; Sheykhzadegan, J.; Bahrani, M.; and Ghayoomi, M. (2011). "Lessons from building a Persian written corpus: Peykare", *Language Resources and Evaluation*, 45: 143–164.

Bohnet, B. (2009). “Efficient parsing of syntactic and semantic dependency structures”. In *Proceedings of the 13th Conference on Computational Natural Language Learning*: Shared Task, pp: 67–72, Stroudsburg, PA, USA.

Brants, T. (2000). “TnT - A statistical part-of-speech tagger”. In *Proceedings of the Association for Neuro-Linguistic Programming and NAACL*, pp. 224–231.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Danesh, M.; Minaei, B.; and Kashefi, O. (2011). “Challenging massive information retrieval in Persian”. *International Journal of Information and Education Technology*, Vol. 1, No. 3.

Darrudi, E.; Hejazi, M.R.; and Oroumchian, F. (2004). “Assessment of a modern Farsi corpus”. In *Proceedings of the 2nd Workshop on Information Technology and its Disciplines*, pp: 73–77, Kish Island, Iran.

Denis, P.; and Sagot, B. (2009). “Coupling an annotated corpus and a morpho-syntactic lexicon for state-of-the-art POS tagging with less human effort”. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, Hong Kong, Chine.

de Saussure, F. (1916). *Cours de Linguistique Generale*, Lausanne, Paris: Payot.

Dolamic, L.; and Savoy, J. (2009). “Persian language, Is stemming efficient?”. In *Proceedings of the 20th International Conference on Database and Expert Systems Applications*, eds. Tjoa, A. M. and Wagner, R.; IEEE Computer Society, pp. 388–392.

Eghbalzadeh, H.; Hosseini, B.; Khadivi, S.; and Khodabakhsh, A. (2012). “Persica: A Persian corpus for multipurpose text mining and natural language processing”. In *Proceedings of the 6th International Symposium on Telecommunications*. IEEE. Tehran.

Ghayoomi, M. (2012a). “Bootstrapping the development of an HPSG-based treebank for Persian”. *Linguistic Issues in Language Technology*, CSLI Publications, 7 (19).

Ghayoomi, M. (2012b). “From grammar rule extraction to treebanking: A bootstrapping approach”. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp:1912–1919, Istanbul, Turkey.

Ghayoomi, M. (2014). *From HPSG-based Persian Treebanking to Parsing: Machine Learning for Data Annotation*. PhD Dissertation. Freie Universität Berlin, Berlin, Germany.

Ghayoomi, M.; and Kuhn, J. (2014). “Converting an HPSG-based treebank into its parallel dependency-based treebank”. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 802–809, Reykjavik, Iceland.

Jadidinejad, A.; Mahmoudi, F.; and Dehdari, J. (2010). “Evaluation of Perstem: A simple and efficient stemming algorithm for Persian”. In *Proceedings of the Multilingual Information Access Evaluation I. Text Retrieval Experiments*, eds. Peters, C.; Nunzio, G. D.; Kurimo, M.; Mandl, T.; Mostefa, D.; Peñas, A.; and Roda, G.; Heidelberg, Germany: Springer, vol. 6241 of Lecture Notes in Computer Science, pp. 98–101.

Jurafsky, D.; and Martin, H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Klein, D.; and Manning, C.D. (2003). "Accurate unlexicalized parsing". In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Leech, G. (2004). "Adding linguistic annotation," chapter 2, Edited by Wynne, M., *Developing Linguistic Corpora: A Guide to Good Practice*. AHDS. Literature, Languages and Linguistics. The Oxford Text Archive.

Long, M. H. (2005). *Second Language Needs Analysis*. Cambridge: Cambridge University Press.

McDonald, R.; Pereira, F.; Ribarov, K.; and Hajič, J. (2005). "Non-Projective dependency parsing using spanning tree algorithms". In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp: 523-530, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mokhtaripour, A.; and Jahanpour, S. (2006). "Introduction to a new Farsi stemmer". In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 826–827, New York, NY, USA: ACM.

Mohammadi, A.; Hajiaghajani, S.; and Bahrani, M. (2023). "ACO-tagger: A novel method for part-of-speech tagging using Ant Colony optimization". ArXiv: 2303.16760. Cornell University.

Mollanorozy, S.; Tanti, M.; and Nissim, M. (2023). "Cross-lingual transfer learning with Persian". In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual Natural Language Processing*, pp 89-95.

MosaviMiangah, T. (2006). "Automatic lemmatization of Persian words". *Journal of Quantitative Linguistics*, 13, 1–15.

Müller, T.; Cotterell, R.; Fraser, A.; and Schütze, H. (2015). "Joint lemmatization and morphological tagging with Lemming". In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, pp. 2268–2274.

Müller, T.; Schmid, H.; and Schütze, H. (2013). "Efficient higher-order CRFs for morphological tagging". In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA: Association for Computational Linguistics, pp. 322–332.

Nivre, J.; Hall, J.; and Nilsson, J. (2006). "Maltparser: A data-driven parser generator for dependency parsing". In *Proceedings of the 15th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp: 2216-2219.

Oroumchian, F.; Tasharofi, S.; Amiri, H.; Hojjat, H.; and Raja, F. (2006). *Creating a Feasible Corpus for Persian POS Tagging*, Technical Report TR3/06, University of Wollongong in Dubai.

Petrov, S.; Barrett, L.; Thibaux, R.; and Klein, D. (2006). “Learning accurate, compact, and interpretable tree annotation”. In *Proceedings of the 21st International Conference on Computational Linguistics and Association for Computational Linguistics*, pp. 433–440.

Pollard, C. J.; and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rasooli, M. S.; Kouhestani, M.; and Moloodi, A. (2013). “Development of a Persian syntactic dependency treebank”. In *Proceedings of the HLT Conference of the NAACL*, pp. 306–314, Atlanta, Georgia.

Sabouri, S.; Rahmati, E.; Gooran, S.; and Sameti, H. (2022) “Naab: A ready-to-use plug-and-play corpus for Farsi”. In arXiv:2208.13486v1, Cornell University.

Sagot, B.; Walther, G.; Faghiri, P.; and Samvelian, P. (2011). “A new morphological lexicon and a POS tagger for the Persian Language”. In *International Conference in Iranian Linguistics*, Uppsala, Sweden.

Schmit, H. (2004). “Efficient parsing of highly ambiguous context-free grammars with bit vectors”. In *Proceedings of the 20th International conference on Computational Linguistics*. Geneva, Switzerland.

Schütze, H. (1995). “Distributional part-of-speech tagging”. In *Proceedings of the 7th Conference on European Chapter of the Association for Computational Linguistics*, pp: 141–148. Morgan Kaufmann Publishers Inc.

Seraji, M. (2011). “A statistical part-of-speech tagger for Persian”. In *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA*, pp. 340–343, Riga, Latvia.

Seraji, M.; Megyesi, B.; and Nivre, J. (2012). “Bootstrapping a Persian dependency treebank”. *Linguistic Issues in Language Technology*, 7(18).

Shamsfard, M.; and Fadaee, H. (2008). “A hybrid morphology-based POS tagger for Persian”. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, eds. Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odjik, J.; Piperidis, S.; and Tapias, D.; Marrakech, Morocco: European Language Resources Association.

Shamsfard, M.; Jafari, H.S.; and Ilbeygi, M. (2010). “STeP-1: A set of fundamental tools for Persian text processing”. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp: 859-865, Valletta, Malta.

Sharifloo, A.; and Shamsfard, M. (2008). “A bottom up approach to Persian stemming”. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 583–588.

Tashakori, M.; Meybodi, M. R.; and Oroumchian, F. (2002). “Bon: The Persian stemmer”. In *Proceedings of the 1st EuroAsian Conference on Information and Communication Technology*, pp. 487–494, London, UK, UK: Springer-Verlag.

Tasharofi, S.; Raja, F.; Oroumchian, F.; and Rahgozar, M. (2007). “Evaluation of statistical part of speech tagging of Persian text”. In *Proceedings of the International Symposium on Signal Processing and its Applications*, Sharjah, (U.A.E.).

Tesnière, L. (1953). *Esquisse d'une syntaxe structural*. Paris: Librairie C. Klincksieck.

Tesnière, L. (1959). *Éléments de syntaxe structural*. Paris: Librairie C. Klincksieck.
Tesnière, L. (1980). *Grundzüge der strukturalen Syntax*. Stuttgart: Klett-Cotta. Translated by Ulrich Engel.

.

**استناد به این مقاله:** قیومی، مسعود، صالحی، الهام و علیجانی، اعظم. (۱۴۰۳). تحلیلی بر پیکرهٔ برچسب‌گذاری‌شدهٔ حاصل از محتوای کتاب‌های درسی پایه‌های اول تا ششم ابتدایی. *زبان و زبان‌شناسی*، ۲۰(۳۹)، ۱۵۵- ۱۹۲.
doi: 10.30465/LSI.2025.47627.172