



## Persian Author Identification based on Systemic Functional Grammar

Fatemeh Soltanzadeh<sup>1</sup> | Azadeh Mirzaei<sup>2</sup> | Mohammad Bahrani<sup>3</sup> |  
Shahram Modarres Khiabani<sup>4</sup>

1. Ph.D. Candidate, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: f\_soltanzadeh@atu.ac.ir
2. Corresponding Author, Assistant Professor of Linguistics, Department of Linguistics, Faculty of Persian Literature and Foreign Languages, Allameh Tabataba'i University, Tehran, Iran. E-mail: azadeh.mirzaei@atu.ac.ir.
3. Assistant Professor, Department of Computer, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran. E-mail: bahrani@atu.ac.ir.
4. Assistant Professor, Department of English Language and Translation, Islamic Azad University, Karaj, Iran. E-mail: shmodarress@yahoo.com.

---

### Article Info

### ABSTRACT

**Article type:**  
Research Article

**Article history:**  
Received:  
Accepted:

**Keywords:**  
Author Identification,  
Forensic Linguistics,  
Systemic Functional  
Grammar,  
Function Words,  
Conjunctive Adjunct,  
Mood Adjunct,  
Comment Adjunct.

Automated Author identification is one of the important fields in forensic linguistics. In this study, the effectiveness of systemic functional grammar features in Persian authorship attribution was compared with that of function words. First, a corpus composed of documents written by seven contemporary Iranian authors was collected. Second, a list of function words was extracted from the corpus. Moreover, conjunction, modality and comment adjunct system networks were applied to form a lexicon using linguistics resources. Then, the relative frequency of function words in addition to systemic functional features were calculated in each document. Multilayer perceptron classifier, a type of neural network, was used for learning phase which resulted in a desirable accuracy in evaluation phase. The results of the study showed that using function words method is superior to systemic functional approach alone in Persian author identification, however, simultaneous use of the two methods increases the effectiveness in comparison to each alone.

---

**Cite this article:**



© The Author(s).

Publisher: Razi University.

0000000000000000



## تشخیص خودکار هویت نویسنده متن در زبان فارسی بر اساس دستور نقش‌گرای نظام‌مند

فاطمه سلطان‌زاده<sup>۱</sup>، آزاده میرزایی<sup>۲</sup>، محمد بحرانی<sup>۳</sup>، شهرام مدرس خیابانی<sup>۴</sup>

۱. دانشجوی دکتری گروه زبان‌شناسی، دانشکده ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: f\_soltanzadeh@atu.ac.ir
۲. نویسنده مسئول، دانشیار گروه زبان‌شناسی، دانشکده ادبیات فارسی و زبان‌های خارجه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: azadeh.mirzaei@atu.ac.ir
۳. استادیار گروه رایانه، دانشکده آمار، ریاضی و رایانه، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: bahrani@atu.ac.ir
۴. استادیار گروه آموزش مترجمی زبان انگلیسی، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران. رایانامه: shmodarress@yahoo.com

### چکیده

### اطلاعات مقاله

تشخیص خودکار هویت نویسنده متن یکی از مسائل مهم زبان‌شناسی حقوقی تلقی می‌شود. پژوهش حاضر تلاش دارد کارایی ویژگی‌های مبتنی بر مفاهیم دستور نقش‌گرای نظام‌مند هالیدی را با کارایی واژه‌های دستوری در تشخیص هویت نویسنده مقایسه نماید. در ابتدا، پیکره‌ای از آثار هفت نویسنده معاصر ایرانی گردآوری شد. در مرحله دوم، فهرستی از واژه‌های دستوری از پیکره استخراج شد؛ به علاوه یک مجموعه واژگان بر اساس شبکه نظام حروف ربط، شبکه نظام افزوده وجه و شبکه نظام افزوده نگرشی با استفاده از منابع زبانی تهیه شد. سپس بسامد نسبی واژه‌های دستوری و ویژگی‌های مبتنی بر دستور نقش‌گرای نظام‌مند در هر متن محاسبه شد. طبقه‌بند پرسپترون چند لایه، نوعی شبکه عصبی، برای مرحله آموزش سامانه به کار گرفته شد و به دقت مطلوبی در مرحله ارزیابی منجر شد. بررسی نتایج ارزیابی سامانه حاکی از این است که روش محاسبه بسامد واژه‌های دستوری نسبت به روش مبتنی بر دستور نقش‌گرای نظام‌مند در تشخیص هویت نویسنده متون فارسی برتری دارد. در صورتی که ویژگی‌های دستور نقش‌گرای نظام‌مند هالیدی در کنار ویژگی بسامد واژه‌های دستوری به کار روند، کارایی سامانه نسبت به حالتی که تنها از ویژگی بسامد واژه‌های دستوری استفاده شود، ارتقا می‌یابد.

نوع مقاله: مقاله پژوهشی

تاریخ دریافت:

تاریخ پذیرش:

واژه‌های کلیدی:

تشخیص هویت نویسنده،  
زبان‌شناسی حقوقی،  
دستور نقش‌گرای نظام‌مند،  
واژه‌های دستوری،  
افزوده ربطی،  
افزوده وجه،  
افزوده نگرشی.

استناد:



00000000000000000000



## ۱. مقدمه

در جهان امروز با توجه به پیشرفت فناوری و استفاده گسترده از وب، این امکان برای کاربران فراهم است که در فضای مجازی هویت واقعی خود را پنهان سازند، با هویت جعلی خود را معرفی نمایند، ایمیل‌هایی بدون نام نویسنده واقعی ارسال نمایند و احیاناً دست به سرقت ادبی یا علمی بزنند. از این رو لازم است سامانه‌هایی طراحی شوند که امنیت فضای مجازی را تأمین کنند، سرقت ادبی و علمی را کشف نمایند و به صورت خودکار هویت واقعی نویسندگان متون را در این فضا شناسایی کنند. به همین سبب در دهه‌های اخیر تشخیص خودکار هویت نویسنده متن به یکی از مسائل مهم در زبان‌شناسی حقوقی مبدل شده است.

تشخیص خودکار هویت نویسنده متن بر این فرض استوار است که هر فرد یک گویش فردی<sup>۱</sup> و یا سبک خاص و منحصر به فرد در گفتار یا نوشتار دارد. در حقیقت هنگامی که مردم از زبان استفاده می‌کنند، ساخت‌های خاصی از واژگان-دستور<sup>۲</sup> را برمی‌گزینند که سایر افراد ممکن است از آنها کمتر استفاده نمایند. در مرحله بعد هر فرد این ساخت‌ها را به شکل متفاوتی از دیگر افراد ترکیب می‌کند تا پیام خود را منتقل کند و این به معنای منحصر به فرد بودن گویش فردی و سبک گفتار یا نوشتار افراد بشر است (کلتارد،<sup>۳</sup> ۲۰۰۴).

در زبان فارسی چند پژوهش در خصوص تشخیص هویت نویسنده متن انجام شده است (فرهمندپور و همکاران، ۱۳۹۱؛ عارفی و همکاران، ۱۴۰۰). این پژوهش‌ها بیشتر توجه خود را به جنبه‌های فنی و مهندسی کار معطوف نموده و به جنبه‌های زبانی و نظری کمتر اهتمام ورزیده‌اند. پژوهش‌های مذکور عمدتاً از ویژگی‌های تکراری (همچون غنای واژگانی<sup>۴</sup>؛ بسامد گروه‌های نحوی، باهم‌آیی کلمات، بسامد نسبی علائم نگارشی) برای طراحی سامانه‌های تشخیص خودکار هویت نویسنده متن استفاده می‌کنند و تلاشی در جهت طراحی ویژگی‌های زبانی بر مبنای یک نظریه زبانی خاص ننموده‌اند. برای زبان فارسی دو پژوهش در راستای تشخیص خودکار هویت نویسنده متن انجام شده است که ناظر بر جنبه‌های زبانی هستند (دباغ، ۲۰۰۷ و گلشائی، ۱۳۹۸). این پژوهش‌ها از روش محاسبه بسامد واژه‌های دستوری<sup>۵</sup> در تشخیص هویت نویسنده متن استفاده کرده‌اند. محاسبه بسامد واژه‌های دستوری یک روش مهم و معتبر در تشخیص هویت نویسنده

<sup>۱</sup> idiolect

<sup>۲</sup> lexigrammar

<sup>۳</sup> Coulthard

<sup>۴</sup> lexical richness

<sup>۵</sup> function words

است. واژه‌های دستوری به این دلیل که نویسنده کنترل خودآگاه در استفاده از آنها ندارد و در متن به دفعات زیاد به کار می‌روند، می‌توانند ویژگی‌های منحصر به فردی در تشخیص خودکار نویسنده تلقی شوند (سگارا<sup>۱</sup> و همکاران، ۲۰۱۵؛ گلشائی، ۱۳۹۸). دباغ (۲۰۰۷) بر اساس واژه‌های دستوری پربسامد و با استفاده از روش‌های آماری به تفکیک سبک نگارش نظامی گنجوی/ شهریار و عبدالحسین زرین‌کوب/ سیمین دانشور پرداخته است. نتایج این پژوهش نشان می‌دهد که واژه‌های دستوری قادر به تفکیک سبک نویسنده‌ها در نظم و نثر فارسی هستند. گلشائی (۱۳۹۸) با تکیه بر گویش فردی و با استفاده از واژه‌های دستوری زبان فارسی تشخیص هویت نویسنده متن را برای پنج نویسنده معاصر بررسی کرده است. نتایج این پژوهش نشان می‌دهد که نویسنده‌های مختلف واژه‌های دستوری را به طور مشابه به کار نمی‌گیرند. در واقع با وجود اینکه برخی واژه‌های دستوری پربسامد توسط همه نویسنده‌ها به کار گرفته می‌شود اما اولویت به کارگیری آنها توسط نویسندگان مختلف متفاوت است. گرچه دو پژوهش نام‌برده (دباغ، ۲۰۰۷ و گلشائی، ۱۳۹۸) اثبات می‌کنند که واژه‌های دستوری در نثر فارسی قابلیت تفکیک سبک نویسندگان را دارند اما برای واژه‌های دستوری، تقسیم‌بندی معنایی ارائه ننموده‌اند. دباغ (۲۰۰۷) در انتخاب واژه‌های دستوری صرفاً به معیار پربسامد بودن واژه‌های دستوری اکتفا کرده است. گلشائی (۱۳۹۸) نیز به بررسی حداقل کلمات مورد نیاز برای تفکیک سبک نویسندگان بر اساس واژه‌های دستوری می‌پردازد. وی توالی‌های یک تا سه‌واژه‌ای (تک‌نگاشتی<sup>۲</sup>، دونگاشتی<sup>۳</sup>، سه‌نگاشتی<sup>۴</sup>) از واژه‌های دستوری را در تشخیص هویت نویسنده متن مورد ارزیابی قرار می‌دهد و در نهایت واژه‌های تک‌نگاشتی یا همان واژه‌های دستوری تک‌واژه‌ای را کاراترین نوع واژه‌های دستوری می‌داند. علاوه بر این در پژوهش وی، گستره مفهوم واژه‌های دستوری کمی محدود شده و ضمائر و تمامی افعال (ربطی و غیرربطی) حذف شده است.

با توجه به نکات ذکر شده در بررسی نقش واژه‌های دستوری زبان فارسی در تفکیک سبک نویسندگان، نیاز است در انتخاب واژه‌های دستوری به مولفه‌های معنایی توجه بیشتر داشت و برای آنها دسته‌های معنایی ارائه کرد. در نهایت می‌توان انواع مختلف واژه‌های دستوری را مورد آزمون قرار داد و بهترین ترکیب واژه‌های دستوری را از این میان انتخاب نمود. پژوهش حاضر سعی بر آن دارد که این خلأ را در حوزه تشخیص هویت نویسنده متون فارسی پر کند و برای واژه‌های

---

<sup>1</sup> Segarra

<sup>2</sup> unigram

<sup>3</sup> bigram

<sup>4</sup> trigram

دستوری، تقسیم‌بندی معنایی ارائه کند. علاوه بر آن کارایی سامانه تشخیص هویت نویسنده با واژه‌های دستوری را مبنای مقایسه قرار دهد و یک نظریه معتبر معنایی را در تشخیص سبک نویسنده زبان فارسی محک بزند. همچنین به نظر می‌رسد که طراحی ویژگی‌های زبانی جدید که مبتنی بر یک نظریه زبانی مدون باشد، در کنار واژه‌های دستوری، در ارتقای کارایی سامانه‌های تشخیص هویت نویسنده موثر باشد. بنابراین تشخیص خودکار نویسنده متن برای زبان فارسی، مبتنی بر یک نظریه زبانی معتبر و مدون با استفاده از ویژگی‌های زبانی جدید یک ضرورت برای زبان فارسی در جهان پیچیده امروز محسوب می‌شود. در این راستا، در پژوهش حاضر تلاش بر آن است تا با استفاده از ویژگی‌های زبانی مبتنی بر دستور نقش‌گرای نظام‌مند<sup>۱</sup> هالیدی<sup>۲</sup> (هالیدی و متیسن<sup>۳</sup>، ۲۰۱۴)، هویت نویسنده متن به صورت خودکار تشخیص داده شود و تأثیر ویژگی‌های دستور نقش‌گرا در مقایسه با ویژگی بسامد واژه‌های دستوری سنجیده شود.

پرسش‌های پژوهش حاضر عبارتند از: ۱- کدامیک از ویژگی‌های زبانی مبتنی بر دستور نقش‌گرای نظام‌مند هالیدی در تشخیص سبک نویسندگان فارسی زبان تأثیرگذار است؟ ۲- اگر ویژگی‌های زبانی مبتنی بر دستور نقش‌گرای نظام‌مند هالیدی در کنار ویژگی بسامد واژه‌های دستوری به کار روند، در کارایی سامانه نسبت به حالتی که تنها از ویژگی بسامد واژه‌های دستوری استفاده شود، چه تأثیری دارد؟

برای تشخیص خودکار نویسنده متن از ویژگی‌های زبانی در سطوح مختلف همچون ویژگی‌های آوایی، واژگانی، نحوی، معنایی، گفتمانی و ترکیب آنها استفاده می‌شود. پس از تعریف ویژگی‌ها، اسناد موجود در پیکره که هر کدام به یک نویسنده خاص در یک مجموعه بسته تعلق دارند، با بهره‌گیری از ویژگی‌های زبانی تعریف‌شده به روش یادگیری ماشین دسته‌بندی و اسناد مشابه به یک نویسنده واحد منسوب می‌گردد.

اولین تلاش در جهت سنجیدن سبک نگارش به قرن نوزدهم میلادی و بررسی نمایشنامه‌های شکسپیر<sup>۴</sup> باز می‌گردد (مندنهال<sup>۵</sup>، ۱۸۸۷). این پرسش که آیا شکسپیر تمامی آثارش را خود به تنهایی نوشته است، سال‌ها توسط پژوهشگران مورد بحث بود. برای مثال، برخی منتقدان ادبی معتقد بودند که شکسپیر نمایشنامه هنری هشتم را با کمک فرد دیگری نوشته است. این ادعا سال‌ها

<sup>1</sup> systemic functional grammar (SFG)

<sup>2</sup> Halliday

<sup>3</sup> Matthiessen

<sup>4</sup> Shakespeare

<sup>5</sup> Mandenhal

توسط پژوهشگران رایانه و هوش مصنوعی بررسی و در نهایت این نتیجه حاصل شد که وی با همکاری فردی به نام فلچر<sup>۱</sup> نمایشنامه هنری هشتم را به رشته تحریر درآورده است (پلچاک<sup>۲</sup>، ۲۰۲۱). این کار با تکیه بر ویژگی‌های آوایی و الگوهای غالب ریتمیک (توزیع هجاهای تکیه دار و بدون تکیه در یک سطر) انجام شده است.

از شمارش کلمات پرتکرار، میانگین طول واژه‌ها در متن و تعداد تکرار کلمات منحصربه‌فرد و غنای واژگانی نیز به عنوان ویژگی‌های واژگانی در تشخیص خودکار نویسنده استفاده می‌شود (استاماتاتوس<sup>۳</sup>، ۲۰۰۹). برخی پژوهشها از ترکیب چند نوع ویژگی بهره گرفته‌اند. برای مثال، در پژوهش ویراسینگه<sup>۴</sup> و همکاران (۲۰۲۱)، ویژگی بسامد واژه‌های دستوری، بسامد گروه‌های نحوی و غنای واژگانی به کار گرفته شده است. پژوهش نجفی و تاوان (۲۰۲۲) که از بسامد گروه‌های نحوی، موجودیت‌های نامدار<sup>۵</sup> و باهم‌آیی کلمات در کنار یادگیری عمیق<sup>۶</sup> بهره می‌جوید به کارایی خوبی در تعیین هویت نویسنده برای زبان انگلیسی دست می‌یابد. پژوهش مارتینز<sup>۷</sup> و همکاران (۲۰۲۲)، نیز ویژگی بسامد گروه‌های نحوی را با روش شبکه عصبی گرافی<sup>۸</sup> مورد آزمون قرار می‌دهد و به دقت خوبی دست می‌یابد.

دو پژوهش دیگر (آرگامون<sup>۹</sup> و همکاران، ۲۰۰۷؛ آرگامون و کوپل<sup>۱۰</sup>، ۲۰۱۳) در تشخیص خودکار نویسنده متن برای زبان انگلیسی انجام شده است که از مجموعه‌ای از ویژگی‌ها بر اساس دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) بهره می‌جوید. در پژوهش اول آرگامون و همکاران (۲۰۰۷)، از دستور نقش‌گرای نظام‌مند در تشخیص هویت، جنسیت، رده سنی و ملیت نویسنده متن استفاده شده و کارایی این دستور در مقایسه با بسامد واژه‌های دستوری سنجیده شده است. در پژوهش دیگر (آرگامون و کوپل، ۲۰۱۳)، از این دستور در تشخیص

<sup>1</sup> Fletcher

<sup>2</sup> Plecháč

<sup>3</sup> Stamatatos

<sup>4</sup> Weerasinghe

<sup>5</sup> named entities

<sup>6</sup> deep learning

<sup>7</sup> Martinez-Galicia

<sup>8</sup> graph neural networks

<sup>9</sup> Argamon

<sup>10</sup> Koppel

جنسیت، رده سنی، ملیت و ویژگی‌های شخصیتی<sup>۱</sup> نویسنده استفاده شده است و کارایی این دستور در مقایسه با واژه‌های محتوایی<sup>۲</sup> مقایسه شده است. برای این منظور یک واژگان خاص طراحی شده که شامل کلمات و عباراتی است که از طریق شبکه<sup>۳</sup> واژگان و اصطلاح‌نامه‌های<sup>۳</sup> برخط گردآوری شده است. هر مدخل واژگان شامل یکسری ویژگی‌ها بر اساس دستور نقش‌گرای نظام‌مند است. نتایج پژوهش اول (آرگامون و همکاران، ۲۰۰۷) حاکی از این است که این ویژگی‌ها در ترکیب با ویژگی‌های مبتنی بر واژه‌های دستوری بیشترین کارایی را دارند. پژوهش دوم (آرگامون و کوپل،<sup>۴</sup> ۲۰۱۳) نیز گزارش می‌کند که در تشخیص جنسیت، رده سنی و ملیت نویسنده ترکیب واژه‌های محتوایی و ویژگی‌های مبتنی بر دستور نقش‌گرای نظام‌مند بالاترین کارایی را داشته است. همچنین روش دستور نقش‌گرای نظام‌مند در قیاس با مبنای مقایسه خود (واژه‌های محتوایی) کارایی پایین‌تری در تشخیص جنسیت، رده سنی و ملیت نویسنده متن برای زبان انگلیسی داشته است.

## ۲. مبانی نظری

### ۲-۱. دستور نقش‌گرای نظام‌مند

در این پژوهش از مفاهیم دستور نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) در تشخیص خودکار هویت نویسنده متن استفاده می‌شود. این دستور زبان‌ها را به عنوان نظامی از انتخاب‌ها برای بیان معنا مدل‌سازی و تمامی انتخاب‌های ساختاری و واژگانی را بر اساس نقش‌های معنایی آنها بازنمایی می‌کند. این نظریه در پردازش زبان طبیعی<sup>۵</sup> و زبانشناسی رایانشی از دهه ۱۹۶۰ میلادی به کارگرفته شده است و خصوصاً در تولید متن<sup>۶</sup> بسیار کاربرد دارد (متیسن و بتمن،<sup>۷</sup> ۱۹۹۱؛ تیچ،<sup>۸</sup> ۱۹۹۵). دستور نقش‌گرای نظام‌مند هالیدی مجموعه‌ای از محدودیت‌ها را در بیان معنا تعریف می‌کند. این دستور در حقیقت شبکه‌ای از گزینه‌های محتمل است. این شبکه از انتخاب‌ها،

<sup>1</sup> personality

<sup>2</sup> content words

<sup>3</sup> thesaurus

<sup>4</sup> Koppel

<sup>5</sup> natural language processing

<sup>6</sup> text generation

<sup>7</sup> Bateman

<sup>8</sup> Teich



شبکه نظام<sup>۱</sup> نامیده می‌شود. با داشتن این شبکه می‌توان برای توصیف سبک یک متن از ویژگی‌های آماری بهره برد. بدین شکل که بسامد شرطی هر گره را در این شبکه با داشتن گره والد آن محاسبه کرده و آن را به عنوان یک معیار برای توصیف سبک متن به کار برد. این تفسیر بر این اصل استوار است که یک نویسنده که دارای سبک خاص و منحصر به فرد است، برای بیان معنا گزینه‌های خاصی را بر سایر گزینه‌ها ترجیح می‌دهد (آرگامون و همکاران، ۲۰۰۷). بر این اساس، در پژوهش حاضر، از دستور نقش‌گرای نظام‌مند هالیدی به عنوان چارچوب نظری پژوهش و از روش پیشنهادی در پژوهش آرگامون و همکاران (۲۰۰۷) برای تعریف ویژگی‌ها در تشخیص هویت نویسنده استفاده می‌شود.

زبان در دستور نقش‌گرای نظام‌مند به صورت نظامی از معناها در نظر گرفته می‌شود که با صورت همراه می‌شود. واحد مطالعه در دستور نقش‌گرای نظام‌مند بند<sup>۲</sup> است. از لحاظ معنایی بند دارای سه لایه مستقل است که با هم در تعاملند. این لایه‌ها فرانش<sup>۳</sup> نامیده می‌شوند و عبارتند از: ۱- فرانش اندیشگانی<sup>۴</sup> که خود دارای دو بخش فرانش تجربی و منطقی است و به بازنمایی تجربه نویسنده از هستی اختصاص دارد؛ ۲- فرانش بینافردی<sup>۵</sup> که چگونگی تعامل میان شرکت‌کنندگان در یک رویداد گفتمانی را بیان می‌کند و ۳- فرانش متنی<sup>۶</sup> که خواننده و شنونده را قادر می‌سازد که متن را از مجموعه‌ای از جملات تصادفی و نامربوط بازشناسد (هالیدی و متیسن، ۲۰۱۴).

در پژوهش حاضر، افزوده‌ها<sup>۷</sup> در دو فرانش بینافردی و متنی مورد توجه هستند. در این پژوهش، افزوده<sup>۸</sup> ربطی<sup>۸</sup> که در فرانش متنی مطرح است و افزوده<sup>۹</sup> وجهی<sup>۹</sup> (افزوده<sup>۹</sup> وجه<sup>۹</sup> و افزوده<sup>۹</sup>)

<sup>1</sup> system network

<sup>2</sup> clause

<sup>3</sup> metafunction

<sup>4</sup> ideational metafunction

<sup>5</sup> interpersonal metafunction

<sup>6</sup> textual metafunction

<sup>7</sup> adjunct

<sup>8</sup> conjunctive adjunct

<sup>9</sup> modal adjunct

<sup>1</sup> mood adjunct

نگرشی<sup>۱</sup> که در فرانش بینافردی به کار می‌رود، برای تشخیص هویت نویسنده استفاده می‌شود. در این راستا سه نوع شبکه نظام معرفی می‌شود: شبکه نظام حروف ربط<sup>۲</sup>، شبکه نظام وجه‌نمایی<sup>۳</sup> و شبکه نظام افزوده نگرشی. در ابتدا درباره فرانش متنی و شبکه نظام حروف ربط سخن می‌شود. سپس به فرانش بینافردی و دو نوع شبکه نظام وجه‌نمایی و شبکه نظام افزوده نگرشی پرداخته می‌شود.

## ۲-۱-۱. فرانش متنی

در فرانش متنی نحوه چینش سازه‌ها و تحت عنوان ساخت آغازگری<sup>۴</sup> پایان‌بخشی<sup>۵</sup> بررسی می‌شود. آغازگر به عنوان نقطه عزیمت پیام اولین سازه بند است به شرط آنکه مشارک<sup>۶</sup> فرآیند<sup>۷</sup> و یا افزوده حاشیه‌ای باشد (هالیدی و متیسن، ۲۰۱۴). باقی جمله منهای آغازگر، پایان‌بخش است.

افزوده‌های ربطی در فرانش متنی مطرح هستند و بین دوبند یا دو جمله پیوند منطقی برقرار می‌کنند. جایگاه بی‌نشان آنها در ابتدای بند و یا بعد از آغازگر است. این افزوده‌ها که با انسجام متن مرتبطند، در دستور نقش‌گرای نظام‌مند در شبکه نظام حروف ربط سازمان‌بندی می‌شوند. انواع مختلف حروف ربط برای پیوند دادن بندها به کار گرفته می‌شوند. در حقیقت نشان می‌دهند که چگونه یک بند بر اساس بافت پیشین بسط و توسعه می‌یابد. در این شبکه نظام، افزوده ربطی (حرف ربط) به سه دسته کلی افزوده تشریحی<sup>۸</sup>، گسترشی<sup>۹</sup> و تفصیلی<sup>۱۰</sup> تقسیم می‌شود (هالیدی و متیسن، ۲۰۱۴) که در شکل ۱ قابل مشاهده است.

در افزوده تشریحی، افزوده‌ها خود به دو دسته کلی تقسیم می‌شوند: بدلی<sup>۱۱</sup> و واضح‌سازی<sup>۱۲</sup>. در افزوده بدلی، یک عنصر دوباره معرفی یا بیان می‌شود که یا به صورت توضیحی<sup>۱</sup> است و یا به

<sup>1</sup> comment adjunct

<sup>2</sup> conjunction

<sup>3</sup> modality

<sup>4</sup> Theme

<sup>5</sup> Rheme

<sup>6</sup> participant

<sup>7</sup> process

<sup>8</sup> Elaborating

<sup>9</sup> Extending

<sup>1</sup> Enhancing 0

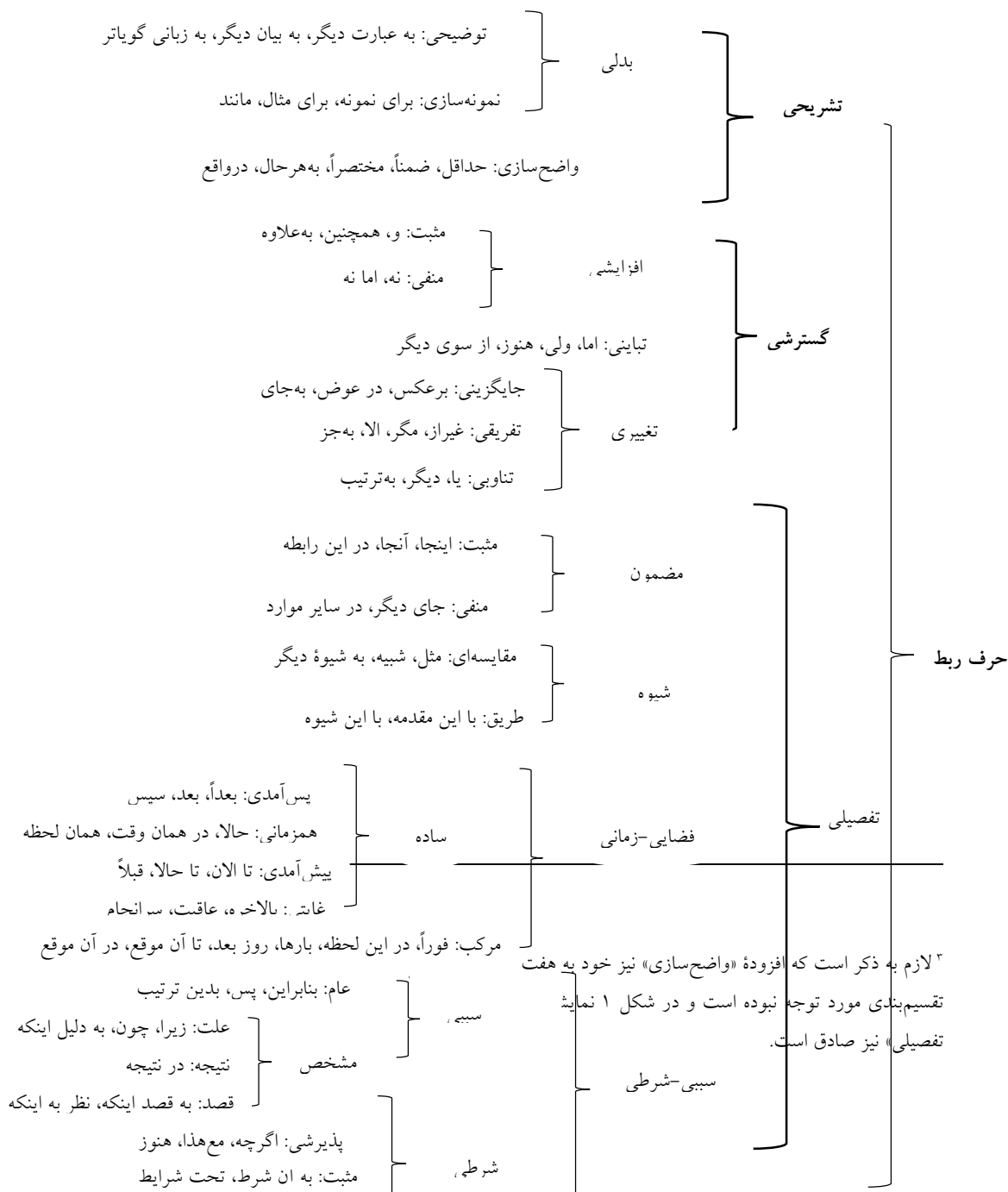
<sup>1</sup> appositive 1

<sup>1</sup> clarification 2

شکل نمونه‌سازی<sup>۲</sup> است. در افزوده واضح‌سازی نیز آن عنصر به صورت واضح‌تری بیان می‌گردند.<sup>۳</sup>

افزوده تشریحی، بدلی، توضیحی:

(۱) آنچه به تو می‌گویم یقین و حتمی است مثل این است که خودم از دهن صاحبکار شنیده باشم. (علوی، ۱۳۸۶)



شکل ۱. شبکه نظام حروف ربط (هالیدی و متیسن، ۲۰۱۴؛ جعفری، ۱۳۸۸: ۱۳۷-۱۳۹)

افزوده تشریحی، بدلی، نمونه‌سازی:

(۲) ما مسلمانیم ارباب مثلاً خود من تا حالا حتی یک بار هم نشده نمازم قضا بشود (گلشیری، ۱۳۷۰).

افزوده گسترشی خود به سه دسته کلی تقسیم می‌شود: ۱- افزوده افزایشی؛ چیزی را به معنا می‌افزاید، ۲- افزوده تغییری؛<sup>۲</sup> به نوعی معنا را تغییر می‌دهد، ۳- افزوده تباینی؛<sup>۳</sup> برخلاف آن چیزی را بیان می‌کند.

افزوده گسترشی، افزایشی، مثبت:

(۳) بچه‌ها هورا کشیدند و کف زدند. (آل احمد، ۱۳۴۶)

افزوده گسترشی، تباینی:

(۴) شب‌ها اگر از فرسنگ‌ها راه به جهت تپه ماهورهای شبانکاره نگاه می‌کردی نور نارنجی پایداری می‌دید که بیش و کم می‌شد اما مرگ نداشت (ابراهیمی، ۱۳۹۹).

افزوده تفصیلی که پیوستگی<sup>۴</sup> را از طریق مقایسه بین عناصر یا بندها به وجود می‌آورد خود به چهار دسته کلی تقسیم می‌شود: مضمون<sup>۵</sup>، شیوه<sup>۶</sup>، فضایی-زمانی<sup>۷</sup>، سببی-شرطی<sup>۱</sup>.

<sup>1</sup> additive

<sup>2</sup> varying

<sup>3</sup> adversative

<sup>4</sup> cohesion

<sup>5</sup> matter

<sup>6</sup> manner

<sup>7</sup> spatio-temporal

افزوده تفصیلی، شیوه، طریق:

(۵) یکی دو مرتبه که مردم ده بیچاره می شدند، کدخدا را پیش خان همسایه می فرستادند و از او کمک می گرفتند و بدین طریق دهکده‌ای به تصرف خانی در می آمد (علوی، ۱۳۹۹).

(۶) راه افتاد طرف در و پیش از اینکه خارج شود ایستاد و بعد با صدای محکمی پرسید: برویانف را چطور کشتی؟ (ساعدی، ۱۳۹۷)

افزوده تفصیلی، فضایی-زمانی، ساده، همزمانی:

(۷) خودت عا<sup>۱</sup>جز نکن ننه حالا کی می خواد بفروشه؟ (محمود، ۱۳۵۳)

افزوده سببی-شرطی از دو نوع کلی سببی و شرطی تشکیل شده است. افزوده سببی خود به دو نوع عام<sup>۲</sup> و مشخص<sup>۳</sup> تقسیم می شود. در نوع مشخص افزوده به طور مشخص به نتیجه<sup>۴</sup> علت<sup>۵</sup> و قصد<sup>۶</sup> اختصاص می یابد.

افزوده سببی-شرطی، سببی، مشخص، علت:

(۸) وکیل رعیت وفادارترین رعیت خویش را به هوسی به زنجیر کشید و کشان به بازار آورد و به دار مجازاتی که هیچ حقش نبود آویخت، فقط به خاطر آنکه خود را در برابر اجانب کوچکتر از او می دید (ابراهیمی، ۱۳۹۹).

افزوده سببی-شرطی، شرطی، منفی:

(۹) با این حرفها نمی خواهم ناراحتت کنم اما تو باید بدانی، باید بفهمی، حتماً کریستین برایت نگفته اگر نه کار به اینجاها نمی کشید (گلشیری، ۱۳۵۰).

## ۲-۱-۲. فرانش بینافردي

در فرانش بینافردي هر بند نوعی تبادل است. تعامل زبانی در این فرانش یک رابطه داد و ستد کالا یا خدمات یا اطلاعات است. در حقیقت در این فرانش پیام از دو نوع گزاره<sup>۷</sup> یا پیشنهاد<sup>۸</sup> است. در صورت دادن یا خواستن اطلاعات، پیام از نوع گزاره و در صورت عرضه یا درخواست

<sup>1</sup> causal-conditional

<sup>2</sup> general

<sup>3</sup> specific

<sup>4</sup> result

<sup>5</sup> reason

<sup>6</sup> purpose

<sup>7</sup> proposition

<sup>8</sup> proposal

کالا یا خدمات<sup>۱</sup> پیام از نوع پیشنهاد است (هالیدی و متیسن، ۲۰۱۴). گوینده در فرانش بینافردی ابزارهایی در دست دارد که با آنها می‌تواند نظر خود را در پیام منعکس کند، میزان قطعیت گزاره را تخمین بزند یا سبب شود که احتمال اجرایی شدن پیشنهاد بالا برود. میزان قطعیت هر گزاره با صورت‌بندی‌هایی که نماینده مفهوم احتمال هستند (حتماً، احتمالاً، شاید، به احتمال زیاد، احتمال می‌رود و ...) و همچنین با قیود تکرار (همیشه، اغلب، گاهی و ...) مشخص می‌شود. میزان قطعیت و یا، به بیانی بهتر، ضمانت اجرایی پیشنهاد با صورت‌بندی‌هایی که نماینده مفهوم اجبار یا التزام<sup>۲</sup> (اجبار، پیشنهاد و اجازه) و تمایل<sup>۳</sup> (تصمیم، توانایی و تمایل) هستند منتقل می‌شود (میرزایی، ۱۳۹۷ الف). در این فرانش محتوای یک بند به دو عنصر وجه<sup>۴</sup> و مانده<sup>۵</sup> تقسیم می‌شود. وجه آن بخش از بند است که از دو قسمت اصلی تشکیل شده است: ۱- فاعل<sup>۶</sup> که یک گروه اسمی است؛ ۲- عنصر زمان‌دار<sup>۷</sup> که بخشی از گروه فعلی است و بیانگر زمان، نمود، وجه، جهت و قطبیت است. عنصر بعدی در سازمان‌بندی نقش‌ها در معنای بینافردی، مانده است که خود از سه عنصر نقشی مختلف تشکیل شده است: الف) محمول؛<sup>۸</sup> ب) متمم؛<sup>۹</sup> ج) افزوده. محمول در تمام بندها به وسیله گروه فعلی تحقق می‌یابد. متمم آن قسمت از مانده است که فاعل نیست ولی توانایی فاعل شدن را دارد و معمولاً به صورت گروه اسمی ظاهر می‌شود. افزوده عنصری از مانده است که توانایی فاعل شدن را ندارد و به طور مشخص به صورت یک گروه قیدی یا حرف اضافه‌ای ظاهر می‌شود (جعفری، ۱۳۸۸: ۱۳۳).

افزوده وجهی در فرانش بینافردی مطرح است و خود به دو دسته تقسیم می‌شود: الف) افزوده‌های وجه که مفاهیمی چون احتمال، اجبار، تمایل، عادت، شدت و زمان را بیان می‌کنند. ب) افزوده‌های نگرشی که تفسیر، قضاوت و اظهار نظر در مورد پیام و گزاره‌ای که بیان می‌شود را ارائه می‌دهد. افزوده‌های وجه در شبکه نظام وجه‌نمایی جای می‌گیرند. آنها نویسنده را قادر می‌سازند که

<sup>1</sup> goods and services

<sup>2</sup> obligation

<sup>3</sup> inclination or readiness

<sup>4</sup> mood element

<sup>5</sup> residue

<sup>6</sup> subject

<sup>7</sup> finite

<sup>8</sup> predicator

<sup>9</sup> complement

حوادث و موجودیت‌ها را در متن بر اساس احتمال، عادت، تمایل و اجبار (التزام) ارزش‌گذاری کند. این شبکه نظام از چهار زیرنظام نوع<sup>۱</sup>، جهت‌گیری<sup>۲</sup>، مقدار<sup>۳</sup> و قطعیت<sup>۴</sup> تشکیل شده است. زیرنظام اول، زیرنظام نوع است که از دو نوع کلی وجه‌سازی<sup>۵</sup> (وجه احتمال - عادت) و تعدیل‌سازی<sup>۶</sup> (وجه تمایل - التزام) تشکیل شده است. وجه‌سازی در ارتباط با گزاره موضوعیت می‌یابد. گزاره‌ها یا تایید می‌شوند یا انکار. نقش نظام وجهیت تحلیل قلمرو عدم قطعیت مابین آری یا خیر است (حسین حمه و همکاران، ۱۴۰۰؛ هالیدی و متیسن، ۲۰۱۴). در وجه احتمال-عادت از یک سو، احتمال وقوع گزاره مطرح است (قطعاً، احتمالاً، شاید، بعیده، غیرممکنه) و از سوی دیگر، میزان تکرار وقوع گزاره (همیشه، معمولاً، گاهی، هرگز). در مورد اول وجه احتمال - عادت از نوع احتمالی است و دومی از نوع عادت است. تعدیل‌سازی در ارتباط با پیشنهاد مطرح می‌شود. همانطور که پیشتر اشاره شد، میزان قطعیت پیشنهاد با صورت‌بندی‌هایی که نماینده مفهوم اجبار یا التزام (اجبار، پیشنهاد و اجازه) و تمایل (تصمیم، توانایی و تمایل) هستند منتقل می‌شود. پس در تعدیل‌سازی نیز دو نوع وجهیت مطرح است: وجه التزامی (مانند باید، ضرورتاً، قهراً و ...) و وجه تمایلی (مانند می‌خواهم، دوست داریم، تصمیم داریم و ...). دومین زیرنظام، زیرنظام جهت‌گیری است. این زیرنظام، دو نوع تقابل را نشان می‌دهد: تقابل شخصی<sup>۷</sup> و بی‌طرفانه<sup>۸</sup> و تقابل صریح<sup>۹</sup> و ضمنی<sup>۱</sup>. برای درک بهتر، این دو نوع تقابل در جدول ۱ با مثال نشان داده شده است.

---

<sup>1</sup> type

<sup>2</sup> orientation

<sup>3</sup> value

<sup>4</sup> polarity

<sup>5</sup> modalization

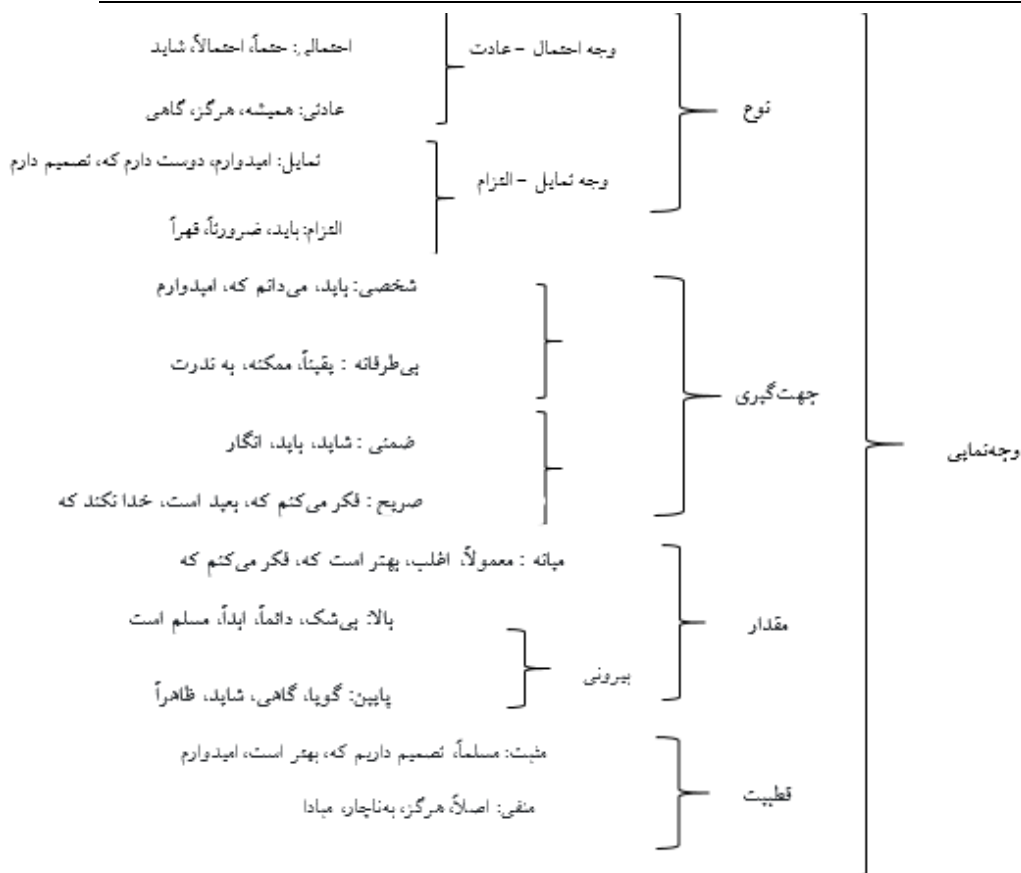
<sup>6</sup> modulation

<sup>7</sup> subjective

<sup>8</sup> objective

<sup>9</sup> explicit

<sup>1</sup> implicit



شکل ۲. شبکه نظام افزوده وجه (هالیدی و متیسن، ۲۰۱۴: ۱۸۲)

جدول ۱. تحلیل نمونه‌ها بر اساس زیرنظام جهت‌گیری (هالیدی و متیسن، ۲۰۱۴: ۱۸۱)

جهت‌گیری	شخصی	بی‌طرفانه
ضمنی	باید	قطعاً
صریح	مطمئنم	مسلم است

در مثال‌های «باید» و «مطمئنم» که از نوع شخصی هستند، نویسنده منبع نظر را ذکر می‌کند؛ در حالی که در «قطعاً» و «مسلم است» منبع آن مشخص نیست و حالت بی‌طرفانه دارند. در مورد «باید»، ذکر منبع نظر به صورت ضمنی و در «مطمئنم» به صورت صریح بیان شده است (هالیدی و متیسن، ۲۰۱۴: ۱۸۱). سومین زیرنظام، زیرنظام مقدار است که دارای سه سطح «بالا»، «پایین» و «میان» است.



چهارمین زیرنظام وجه‌نمایی، زیرنظام قطبیت است. قطبیت تقابل میان مثبت و منفی است. قطبیت را می‌توان به نوعی دو سر پیوستار وجهیت در نظر گرفت (میرزایی، ۱۴۰۰). در ادامه مثال‌هایی از افزوده‌های وجه در جملات پیکره پژوهش حاضر آورده شده است. در جدول ۲ افزوده‌های وجه در این مثال‌ها بنابر شبکه نظام وجه‌نمایی تحلیل شده اند.

(۱۰) فولاد گفت: یعنی ممکنه که تو تب و هذیان از خونه زده باشه بیرون و تو نخلستانا، گوشه‌ای افتاده باشه و سرما خشکش کرده باشه؟ (محمود، ۱۳۸۱).

(۱۱) با این حرف‌ها نمی‌خواهم ناراحت کنم اما تو باید بدانی، باید بفهمی، حتماً کریستین برایت نگفته اگر نه کار به اینجاها نمی‌کشید (گلشیری، ۱۳۵۰).

(۱۲) عمه بزرگ گفت: خسرو خان، از یک شازده بعید است که بادبادک پسر باغبان را بر دارد (گلشیری، ۱۴۰۰).

جدول ۲. تحلیل نمونه‌ها بر اساس شبکه نظام وجه‌نمایی

مثال	نوع (سطح اول)	نوع (سطح دوم)	جهت‌گیری	مقدار	قطبیت
ممکنه	وجه‌سازی	احتمالی	بی‌طرفانه	میانه	مثبت
حتماً	وجه‌سازی	احتمالی	بی‌طرفانه	بالا	مثبت
بعید است	وجه‌سازی	احتمالی	بی‌طرفانه	پایین	منفی
گهگاهی	وجه‌سازی	عادی	بی‌طرفانه	پایین	مثبت
نمی‌خواهم	تعدیل‌سازی	تمایل	شخصی	بالا	منفی
باید	تعدیل‌سازی	التزام	شخصی	بالا	مثبت
تصمیم دارم	تعدیل‌سازی	تمایل	شخصی	بالا	مثبت

(۱۳) این بهت و درنگ گهگاهی از صبح با شیرو بود (دولت‌آبادی، ۱۴۰۱).

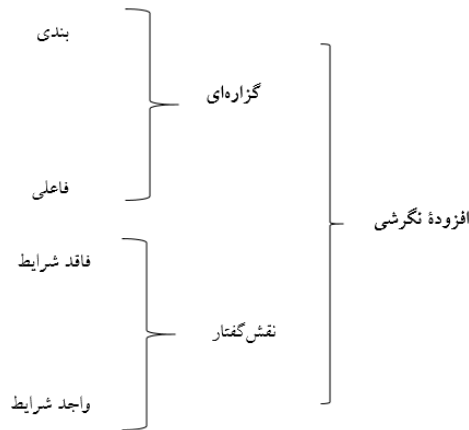
(۱۴) من تصمیم دارم هر شبه بیایم پیش تو، به شرط اینکه باز هم از این چیزها بنویسی (ساعدی، ۱۳۷۷).

شبکه نظام دیگر شبکه افزوده نگرشی است. این افزوده‌ها نظر نویسندگان یا گوینده را بیان می‌کنند و هدف آنها یا محتوای گزاره است یا بخشی از نقش گفتار، از این رو به دو گروه گزاره‌ای<sup>۱</sup> و نقش‌گفتار<sup>۲</sup> تفکیک می‌شوند. در افزوده نگرشی از نوع گزاره‌ای، گوینده یا نویسندگان نظر

<sup>۱</sup> propositional

<sup>۲</sup> speechfunctional

خود را در مورد کل گزاره یا بخشی از آن (فاعل) بیان می‌کند. پس این نوع افزودهای به دو نوع بندی و فاعلی تقسیم می‌شود. افزوده نگرشی از نوع نقش‌گفتار نیز به دو نوع فاقد شرایط<sup>۱</sup> و واجد شرایط<sup>۲</sup> تقسیم می‌شود (هالیدی و متیسن، ۲۰۱۴: ۱۹۰). در شکل ۳ شبکه نظام افزوده نگرشی آمده است.



شکل ۳. شبکه نظام افزوده نگرشی (هالیدی و متیسن، ۲۰۱۴)

در شکل ۴ و ۵، افزوده بندی و فاعلی و در شکل ۶ افزوده نقش‌گفتار نمایش داده شده است.

افزوده گزاره‌ای، بندی، صلاحیتی<sup>۳</sup>، فرض، شایعه<sup>۴</sup>؛

(۱۵) بر فرض هم که خواهرم با این عنکبوت مشغولیتی پیدا کرده باشد، تازه به من چه؟ عنکبوت، عنکبوت است دیگر (آل احمد، ۱۳۵۰).

افزوده گزاره‌ای، بندی، مطلوبیت<sup>۵</sup>، غیرمطلوب:

(۱۶) اگر خود شما دختر داشتید به همچو آدمی شوهرش می‌دادید؟ گفت متأسفانه من دختر ندارم (آل احمد، ۱۳۵۰)

<sup>1</sup> unqualified

<sup>2</sup> qualified

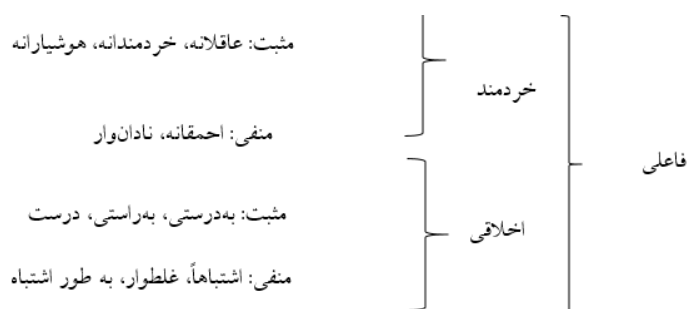
<sup>3</sup> qualificative

<sup>4</sup> hearsay

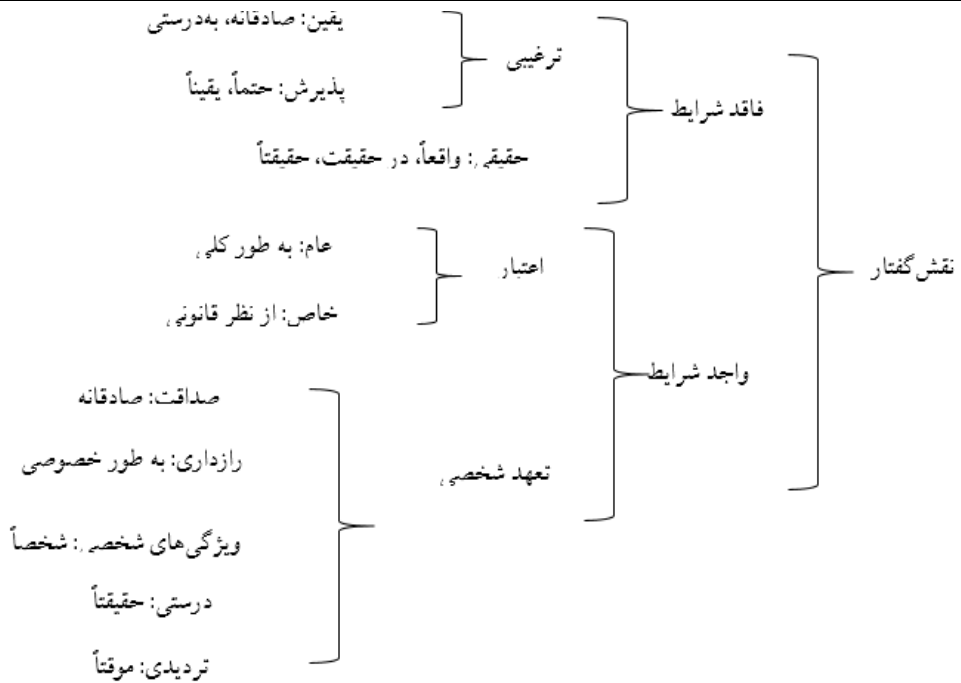
<sup>5</sup> desirability



شکل ۴. شبکه نظام افزوده گزاره‌ای بندی (هالیدی و متیسن، ۲۰۱۴: ۱۹۰)



شکل ۵. شبکه نظام افزوده گزاره‌ای فاعلی (هالیدی و متیسن، ۲۰۱۴: ۱۹۰)



شکل ۶. شبکه نظام افزوده نقش‌گفتار (هالیدی و متیسن، ۲۰۱۴؛ جعفری، ۱۳۸۸: ۱۴۲)

افزوده گزاره‌ای، فاعلی، اخلاقی، منفی:

(۱۷) یک کتابخانه نفیس معماری داشت اما حتی اشتباهاً، یک دیوان حافظ و مولوی در کنار آن مجموعه نفیس، جای نگرفته بود (ابراهیمی، ۱۳۷۴).

افزوده نقش‌گفتار، فاقد شرایط، حقیقی:

(۱۸) شاید حقیقتاً هم اسد نمی‌ترسید و حس ترس سامون بود که او را می‌داشت بپندارد برادرش هم می‌ترسد (دولت‌آبادی، ۱۳۹۵)

افزوده نقش‌گفتار، واجد شرایط، اعتبار، عام:

(۱۹) به‌طور کلی تو کم و کسر نداری؟ (ساعدی، ۱۳۷۷)

#### ۱-۲. واژه‌های دستوری

بنابر تعریف ردفورد<sup>۲</sup> (۲۰۰۴: ۵۶-۴۰)، واژه‌های دستوری حاوی اطلاعاتی در خصوص نقش‌های دستوری منتسب به برخی مفاهیم (چون شخص، شمار، جنس، حالت و مانند آن) هستند و معنای قائم به ذات ندارند. با این توضیح، حرف تعریف، کمی‌نما، ضمیر، فعل کمکی، علامت مصدر و

<sup>1</sup> validity

<sup>2</sup> Radford

متمم‌نما جز مقوله‌های دستوری به حساب می‌آیند. در این پژوهش در تعریف واژه‌های دستوری دو ملاک در نظر گرفته شده است. اگر واژه مورد نظر مسئول انتقال مفهومی دستوری باشد یا اگر خودش به صورت مستقل معنی نداشته باشد، آنگاه آن واژه دستوری است. به این ترتیب حروف اضافه که حوزه معنایی‌شان در همنشینی با واژه‌های دیگر مشخص می‌شود، دستوری محسوب می‌شوند (میرزایی و صفری، ۱۳۹۴: ۲۵۴-۲۵۵). در مجموع، در پژوهش حاضر حرف اضافه، کمی‌نما، ضمیر، فعل کمکی، همکرد فعل، متمم‌نمای «که»، حرف تعریف «یه» و نقش نمای مفعولی «را» کلمات دستوری به شمار می‌آیند.

### ۳. پیکره پژوهش

در این پژوهش به منظور طراحی سامانه‌ای برای تشخیص خودکار هویت نویسنده متون فارسی، پیکره‌ای از آثار هفت نویسنده معاصر ایرانی به نام‌های هوشنگ گلشیری، بزرگ علوی، احمد محمود، محمود دولت‌آبادی، نادر ابراهیمی، جلال آل احمد و غلامحسین ساعدی گردآوری و هنجارسازی شد. منبع مورد استفاده برای گردآوری این داده، پیکره متنی زبان فارسی (بیجن خان و همکاران، ۲۰۱۱)، پایگاه داده‌های زبان فارسی (عاصی، ۱۹۹۷) بوده است. علاوه بر این دو پیکره، آثار مختلف هفت نویسنده در وب جستجو شد و کتب متعلق به آنها که به قالب پی‌دی‌اف بودند، توسط نرم افزار ایبو<sup>۲</sup> به قالب متنی در آمدند. در نهایت کل پیکره هنجارسازی شد و به صورت هماهنگ در آمد. در جدول ۳ آماره‌های این پیکره آمده است. لازم به ذکر است از آنجا که زبان مورد استفاده در اکثر این آثار زبان محاوره است و ابزارهای پیش‌پردازشگر<sup>۳</sup> زبان فارسی در مورد زبان محاوره کارایی نامطلوبی دارند و یا دسترسی به آنها محدود است، پیکره به صورت خام و فاقد برچسب در نظر گرفته شده است. همچنین روش پیشنهادی در این پژوهش مستقل از ابزارهای پیش‌پردازشگر است.

جدول ۳. آماره‌های پیکره داستانی

نام نویسنده	تعداد آثار	مجموع کل واژه‌ها
نادر ابراهیمی	۸	۲۶۱۸۶۶
احمد محمود	۹	۳۶۱۵۶۵
جلال آل احمد	۹	۲۵۴۲۷۶
بزرگ علوی	۷	۳۲۶۸۰۲
هوشنگ گلشیری	۱۲	۲۴۷۷۹۱
محمود دولت آبادی	۸	۳۴۶۱۳۲

<sup>1</sup> normalization

<sup>2</sup> <https://www.eboo.ir/>

<sup>3</sup> preprocessor

۳۲۳۳۵۸	۶	غلامحسین ساعدی
۲۱۲۱۷۹۰	۵۹	مجموع

#### ۴. روش پژوهش

در این پژوهش برای تشخیص خودکار هویت نویسنده متن از رویکرد نقش‌گرای نظام‌مند هالیدی (هالیدی و متیسن، ۲۰۱۴) و روش پیشنهادی در پژوهش آرگامون و همکاران (۲۰۰۷) استفاده شده است. برای این منظور ابتدا یک فهرست شامل واژه‌های دستوری، یک مجموعه واژگان برای افزوده‌های ربطی، وجه و نگرشی بر اساس رویکرد نقش‌گرای نظام‌مند هالیدی و همچنین پیکره‌ای از آثار هفت نویسنده معاصر گردآوری شد (ارجاع به بخش ۳). سپس الگوریتمی طراحی شد که کلماتی از متون پیکره را که در فهرست واژه‌های دستوری و یا مجموعه واژگان تعریف شده‌اند استخراج و بسامد نسبی شاخص‌های تعریف شده برای هر متن را محاسبه می‌کند. بسامد نسبی شاخص‌ها در این پژوهش شامل بسامد نسبی هر یک از واژه‌های دستوری و همچنین بسامد نسبی هر گره در شبکه نظام‌های مبتنی بر دستور نقش‌گرای هالیدی است. در مرحله بعد مقدار هر یک از شاخص‌ها در هر متن به عنوان معیاری برای تمایز قائل شدن میان متون مختلف در نظر گرفته می‌شود. بدین شکل که الگوریتم یادگیری ماشینی، بردار ویژگی‌ها که شامل هر یک از شاخص‌های تعریف شده برای متون است، را دریافت می‌کند و با توجه به تفاوت میان مقادیر این شاخص‌ها بین سبک نویسندگان تمایز قائل می‌شود. در ادامه تمامی این مراحل به تفصیل شرح داده می‌شود.

#### ۴-۱. استخراج کلمات از متن

واژگان شامل کلمات و عباراتی است که به صورت نیمه خودکار<sup>۲</sup> از طریق پیکره یا شبکه واژگان گردآوری شده است. هر مدخل واژگان شامل یکسری ویژگی‌های معنایی بر اساس دستور نقش‌گرای نظام‌مند است. برای گردآوری ویژگی‌های معنایی واژگان از سه منبع پیکره گفتمانی زبان فارسی (میرزایی و صفری، ۲۰۱۸)، فارس‌نت (شمس‌فرد و همکاران، ۲۰۱۰) و پیکره گردآوری شده در پژوهشی در حوزه دستور نقش‌گرای نظام‌مند (جعفری، ۱۳۸۸) بهره گرفته شده است. پیکره گفتمان زبان فارسی جزء معدود پیکره‌های برچسب‌خورده جهان است که اطلاعات گفتمانی را به صورت دستی بر روی داده‌های زبانی قرار داده‌است. پیکره گفتمان زبان فارسی در دو سطح درون‌جمله‌ای و بینا جمله‌ای تهیه شده است. کلیات شیوه برچسب‌گذاری این پیکره که در مجموع

<sup>۱</sup> node

<sup>۲</sup> منظور از نیمه خودکار این است که در ابتدا فهرست کلمات به صورت خودکار گردآوری می‌شود و سپس به صورت دستی توسط زبان‌شناس دسته‌بندی و برچسب‌گذاری می‌شود.

حجمی بالغ بر یک میلیون کلمه دارد، بر اساس استاندارد درخت‌بانک گفتمانی پن است (پراساد<sup>۲</sup> و دیدگران، ۲۰۰۸) که این شیوه‌نامه خود به لحاظ نظری در بحث روابط منطقی از رویکرد نقش‌گرای هالیدی و متیسن در فرانش منطقی پیروی می‌کند. علاوه بر آن، در پیکره گفتمان برای سه مفهوم وجه، قطبیت و قطعیت برچسب‌گذاری صورت گرفته است (شمس‌فرد و بیجن‌خان، ۱۴۰۱: ۷۳). فارس‌نت نیز بزرگترین شبکه واژگان<sup>۳</sup> زبان فارسی است. پژوهش جعفری (۱۳۸۸) به بررسی افزوده‌ها در زبان فارسی بر اساس رویکردهای نقشی و صوری پرداخته و در این راستا پیکره‌ای گردآوری کرده است. در ادامه نحوه جمع‌آوری واژگان برای انواع افزوده‌ها و همچنین فهرست واژه‌های دستوری بیان شده است.

همانطور که پیشتر اشاره شد، افزوده‌های ربطی در فرانش منطقی مطرح هستند و بین دو بند یا دو جمله پیوند منطقی برقرار می‌کنند. برای تهیه فهرست افزوده‌های ربطی، کل حروف ربط موجود در پیکره گفتمان استخراج شد. در پیکره گفتمان به پیروی از درخت‌بانک گفتمانی پن برچسب‌های نماینده معنای منطقی در چهار گروه کلی شامل زمانی، وابستگی، مقایسه و گسترش دسته‌بندی می‌شوند. در این دسته‌بندی هر گروه به دو زیربخش «نوع» و «زیرنوع» تقسیم‌بندی می‌شود (شمس‌فرد و بیجن‌خان، ۱۴۰۱: ۸۰). به دلیل آنکه هر یک از حروف ربط می‌تواند در پیکره بیش از یک برچسب داشته باشد، تنها آن حروفی انتخاب شد که در بیش از شصت درصد موارد دارای یک نوع برچسب معین است. بدین ترتیب ۳۷۲ حرف ربط انتخاب شد. پس از استخراج این فهرست، برچسب هر یک از حروف ربط به صورت دستی بررسی شد و بر اساس شبکه نظام ربط نمایش داده شده در شکل ۱ بازبینی و دسته‌بندی مجدد شد.

برای تهیه فهرست افزوده‌های وجه و نگرشی از پیکره گفتمان، فارس‌نت و پیکره پژوهش جعفری (۱۳۸۸)، استفاده شد. برای تهیه فهرست افزوده‌های وجه، فهرست کلمات دارای سه نوع برچسب «وجه»، «قطبیت» و «قطعیت» از پیکره گفتمان استخراج شد و بر اساس شبکه نظام وجه‌نمایی که در شکل ۲ قابل مشاهده است، دسته‌بندی شد. علاوه بر آن از پیکره پژوهش جعفری برای تهیه فهرست افزوده وجه و نگرشی استفاده شد. برای افزوده‌های نگرشی مستخرج از پیکره جعفری، فهرست کلمات مترادف با آنها با کمک فارس‌نت یافت و به مجموعه واژگان افزوده شد. از آنجا که در برخی آثار نویسندگان پیکره معاصر از زبان محاوره استفاده شده است، برخی عبارات که صورت محاوره‌ای دارند (مانند «معلومه»، «ممکنه»)، نیز برچسب‌گذاری و به مجموعه واژگان

<sup>۱</sup> Penn discourse treebank

<sup>۲</sup> Prasad

<sup>۳</sup> WordNet

افزوده شد. همچنین صورتهای منفی مانند «معلوم نیس» در مقابل «معلوم است» نیز در مجموعه واژگان گنجانده شد. علاوه بر آن صورتهای تصریفی افعالی مانند «فکر کردن»، «حس کردن»، «خواستن»، «دوست داشتن» و غیره نیز در مجموعه واژگان قرار گرفت. در مجموع ۳۲۲ افزوده وجه و ۲۰۱ افزوده نگرشی برچسب گذاری شد.

از آنجا که فهرست واژه‌های دستوری و محتوایی وابسته به حوزه تخصصی یا ژانر هر متن است (میرزایی و صفری، ۱۳۹۴)، برای تهیه فهرست واژه‌های دستوری، از پیکره گردآوری شده در پژوهش حاضر استفاده شد. برای این منظور از روش وزن‌دهی فرکانس کلمه - معکوس فرکانس سند استفاده شد.<sup>۱</sup> بر اساس معیار فرکانس کلمه - معکوس فرکانس سند، ۱۹۷ واژه‌ای که دارای مقادیر کوچکتری بودند، به عنوان واژه دستوری انتخاب شد. تا به اینجا نحوه گردآوری واژگان شرح داده شد. در ادامه نحوه عملکرد سامانه تشریح می‌شود.

در ابتدا، کلماتی از متن که در مجموعه واژگان تعریف شده‌اند استخراج و سپس بر اساس برچسب آن واژه یا عبارت در مجموعه واژگان، برچسب گذاری می‌شوند. به عنوان نمونه، حرف ربط «برای مثال»، در مجموعه واژگان بر اساس نظام حروف ربط به ترتیب سلسه مراتب دارای برچسب «تشریحی»، «بدلی»، «نمونه‌سازی» است. سامانه تمامی موارد وقوع عبارت «برای مثال» را در متن کشف و بر اساس برچسب واژگان برچسب گذاری می‌کند. برای استخراج کلمات از متن سعی بر آن بود که اگر یک کلمه یا عبارت زیر مجموعه عبارت دیگری در مجموعه واژگان باشد (مانند «که» و «به محض آن که»)، برچسب عبارت بزرگتر در نظر گرفته شود.

#### ۴-۲. محاسبه بسامد نسبی شاخص‌ها

پس از استخراج کلمات از متن، بسامد نسبی مقادیر شاخص‌های معنایی برای هر متن محاسبه می‌شود. خروجی این مرحله بردار ویژگی‌های معنایی است که توصیفگر متن است. بر اساس ویژگیهای مبتنی بر دستور نقش‌گرای هالیدی و واژه‌های دستوری هر متن با بردار ویژگی‌های عددی تعریف می‌شود.

<sup>۱</sup> tf-idf

<sup>۲</sup> در این روش میزان تکرار یک کلمه در یک مستند در مقابل تعداد تکرار آن در مجموعه کلیه مستندات سنجیده می‌شود. در این روش به هر کلمه در یک متن عددی تخصیص می‌دهد. این معیار برای کلماتی که تقریباً در تمامی اسناد پیکره یافت شوند، دارای مقادیر کمینه است. این کلمات همان واژه‌های دستوری تلقی می‌شوند (شوتز و همکاران، ۲۰۰۸).



بسامد نسبی هر گره در نظام‌های مبتنی بر دستور نقش‌گرای هالیدی<sup>۱</sup>، بسامد انتخاب گره O1 بر حسب انتخاب O2 است که گره والد آن در شبکه نظام محسوب می‌شود. در حقیقت گره O2 دارای برچسب درشت<sup>۲</sup> و O1 دارای برچسب ریز<sup>۳</sup> است و زیر مجموعه آن برچسب تلقی می‌شود. بسامد نسبی O1 با داشتن گره O2 به عنوان گره والد به این شکل تعریف می‌شود:

$$RF_d(O_1|O_2) = \frac{N_d(O_1, O_2)}{N_d(O_2)} \quad (\text{فرمول ۱})$$

که در آن  $N_d(O_1, O_2)$  تعداد رخداد همزمان O1 و O2 است و  $N_d(O_2)$  تعداد رخداد O2 در هر متن است و  $d$  به معنای متن یا همان سند است (آرگامون و همکاران، ۲۰۰۷: ۸). به زبان ساده‌تر، چه نسبت از تعداد رخداد کلمات در متن که دارای برچسب O2 هستند، دارای برچسب ریزتر O1 نیز هستند.

به عنوان نمونه، در شبکه نظام حروف ربط، گره‌های «تشریحی»، «گسترشی» و «تفصیلی» زیر مجموعه گره حرف ربط محسوب می‌شوند. بسامد نسبی گره «تشریحی» با داشتن گره «حرف ربط» بدین معناست که چه نسبت از حروف ربط به کارگرفته شده در یک متن، از نوع تشریحی هستند. در مورد شبکه نظام وجه‌نمایی و افزوده نگرشی نیز بسامد نسبی بر اساس فرمول ۱ محاسبه می‌شود. برای مثال، در شبکه نظام وجه‌نمایی، گره‌های «بالا»، «پایین» و «میان» زیر مجموعه گره «مقدار» هستند. بسامد نسبی گره «بالا» با داشتن گره «مقدار» بدین معناست که چه نسبت از افزوده‌های وجه در یک متن، دارای مقدار «بالا» هستند.

علاوه بر ویژگی بسامد نسبی، صرفاً برای نظام وجه‌نمایی، یکسری ویژگی‌ها بر اساس ترکیب گره‌ها تعریف می‌شود. برای هر جفت گره در زیر نظام‌های مختلف وجه‌نمایی (مانند «نوع» و «مقدار»)، بسامد نسبی واژه‌هایی که با نوع دو گره برچسب‌خورده‌اند، بر اساس ترکیب گره‌های

<sup>۱</sup> هر شبکه نظام در هر سطح شامل چند گره است. برای مثال شبکه نظام حروف ربط در سطح اول شامل سه گره، افزوده تشریحی، گسترشی و تفصیلی است. در سطح بعد گره تشریحی شامل دو گره بدلی و واضح‌سازی است. در حقیقت گره تشریحی والد دو گره بدلی و واضح‌سازی است. در اینجا گره تشریحی نسبت به دو گره بدلی و واضح‌سازی دارای برچسب درشت و دو گره بدلی و واضح‌سازی دارای برچسب ریزتر هستند.

<sup>۲</sup> parent node

<sup>۳</sup> coarse-grained

<sup>۴</sup> fine-grained

والد آنها تعریف می‌شود. (آرگامون و همکاران، ۲۰۰۷). برای مثال، بسامد نسبی جفت گره «وجه احتمال-عادت» و «بالا» با داشتن جفت گره «نوع» و «مقدار» بیان می‌کند که چه نسبت از کلماتی که دارای «وجه احتمال-عادت» هستند، در زیر نظام مقدار دارای مقدار «بالا» نیز می‌باشند (مانند «حتماً» و «دائماً»).

برای مقایسه کارایی ویژگی‌های مبتنی بر دستور نقش‌گرای هالیدی و ویژگی واژه‌های دستوری، ویژگی بسامد واژه‌های دستوری تعریف می‌شود. برای این منظور، واژه‌های دستوری در هر متن شناخته شده و بسامد نسبی هر یک از آنها به عنوان یک ویژگی محاسبه می‌گردد. برای محاسبه بسامد نسبی واژه‌های دستوری از این فرمول استفاده شده است:

$$\frac{\text{count}(w)}{\sum_{w' \in FW} \text{count}(w')} \quad (\text{فرمول ۲})$$

در این فرمول  $w$  هر یک از واژه‌های دستوری است که بسامد آن نسبت به بسامد کل واژه‌های دستوری در یک متن خاص محاسبه می‌شود (آرگامون و همکاران، ۲۰۰۷: ۱۳). در صورت کسر فرمول ۲ تعداد رخداد یک واژه دستوری خاص و در مخرج آن تعداد رخداد کل واژه‌های دستوری در متن قرار دارد.

#### ۳-۴. یادگیری ماشینی

برای مرحله یادگیری ماشینی از روش یادگیری باناظر<sup>۱</sup> استفاده می‌شود. در این روش به یک سامانه، مجموعه‌ای از جفت‌های ورودی - خروجی ارائه شده و سامانه تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد. یادگیری باناظر نیازمند تعدادی داده ورودی به منظور آموزش سامانه است. همانطور که پیشتر اشاره شد، خروجی مرحله دوم بردار ویژگی‌های معنایی است که توصیفگر متن است. در مورد تشخیص خودکار نویسنده متن، ورودی‌ها بردار ویژگی‌های زبانی متون هستند و خروجی همان نام نویسنده متن است.

در پژوهش حاضر برای دسته‌بندی نویسندگان، از روش یادگیری عمیق استفاده شده است. رویکردهای مبتنی بر یادگیری عمیق در پژوهش‌های نوین در حوزه پردازش زبان طبیعی از جمله تشخیص خودکار هویت نویسنده بسیار مورد توجه هستند (اوجندو<sup>۲</sup> و همکاران، ۲۰۲۰). یکی از

<sup>۱</sup> supervised learning

<sup>۲</sup> Uchendu

انواع طبقه‌بندهای روش یادگیری عمیق، طبقه‌بند پرسپترون چند لایه<sup>۱</sup> است که از آن در این پژوهش بهره گرفته شده است. برای استفاده از این طبقه‌بند از یک ابزار<sup>۲</sup> یادگیری ماشین به زبان پایتون استفاده شده است. الگوریتم یادگیری، بردار ویژگی‌های معنایی هر متن را به عنوان ورودی دریافت می‌کند. بردار ویژگی‌های معنایی می‌تواند شامل انواع ویژگی‌های مبتنی بر دستور نقش‌گرای هالیدی (شبکه نظام حروف ربط، وجه‌نمایی و افزوده نگرشی) و یا ترکیب انواع مختلف آنها باشد. از سوی دیگر این بردار می‌تواند شامل ویژگی بسامد نسبی واژه‌های دستوری باشد. در حالت دیگر در انتخاب نوع بردار ویژگی می‌توان تمامی ویژگی‌های مبتنی بر دستور نقش‌گرای هالیدی و ویژگی بسامد نسبی واژه‌های دستوری را با هم آمیخت. در نهایت الگوریتم یادگیری نام نویسنده متن را به عنوان خروجی بر می‌گرداند. در ادامه کارایی این الگوریتم در حالت‌های مختلف بردار ویژگی ارزیابی و گزارش می‌شود.

#### ۵. یافته‌ها

به منظور ارزیابی کیفیت سامانه طراحی شده در تشخیص خودکار هویت نویسنده متن از روش ارزیابی متقاطع  $K$  تایی<sup>۳</sup> استفاده می‌شود. اگر کل داده‌ها را به طور تصادفی به  $K$  زیرنمونه تفکیک کنیم، می‌توان در هر مرحله از فرایند، تعداد  $K-1$  از این لایه‌ها را به عنوان مجموعه داده آموزشی و یکی را به عنوان مجموعه داده آزمون در نظر گرفت. مشخص است که تعداد تکرارهای فرآیند برابر با  $K$  خواهد بود (رفائیل‌زاده و دیگران، ۲۰۰۹). در حقیقت یادگیری سامانه با کمک داده آموزشی صورت می‌گیرد و با استفاده از داده آزمون ارزیابی می‌شود. در این پژوهش،  $K=5$  در نظر گرفته شده است. نکته قابل توجه در ارزیابی این است که نه تنها داده آموزش و آزمون با یکدیگر هم‌پوشانی ندارند، بلکه برای ارزیابی سامانه تشخیص سبک نگارش یک نویسنده، نباید یک داستان خاص بین داده آموزش و آزمون تقسیم شود. این نکته‌ای است که در برخی پژوهش‌ها مغفول مانده است. به همین دلیل در ارزیابی سامانه، کل آثار هر نویسنده به پنج قسمت تقسیم شد: چهار قسمت برای آموزش و یک قسمت برای آزمون. این تقسیم بندی به گونه‌ای است که داستان‌های موجود در بخش آموزش از داستان‌های موجود قسمت آزمون متمایز هستند. برای محاسبه کارایی سامانه از معیار دقت استفاده شد. دقت در تشخیص خودکار نویسنده عبارت است از نسبت نتایج درست به دست آمده توسط سامانه به کل تعداد نمونه‌های موجود در داده آزمون

<sup>1</sup> multilayer perceptron

<sup>2</sup> <https://scikit-learn.org/>

<sup>3</sup> K-Fold Cross Validation

<sup>4</sup> accuracy

(درونه و شریعتی، ۲۰۱۴). از آنجا که فرآیند ارزیابی پنج بار تکرار شد، در نهایت از نتایج به دست آمده در پنج مرحله برای کل سامانه و همینطور برای هر نویسنده میانگین وزن دار گرفته شد.<sup>۲</sup>

مدل طراحی شده در چند حالت مورد آزمون قرار گرفت: حالت اول بردار ویژگی شامل ویژگی‌های تعریف شده بر اساس دستور نقش‌گرای نظام‌مند است، حالت دوم بردار ویژگی از نوع ویژگی‌های واژه‌های دستوری است و حالت سوم ترکیب حالت اول و دوم است. در هر حالت دقت سامانه به روش ارزیابی متقاطع K تایی سنجیده شد. در جدول ۴ نتایج ارزیابی سامانه به طور خلاصه گزارش شده است. همانطور که مشاهده می‌شود، بهترین نتایج حاصل از ترکیب روش مبتنی بر دستور نقش‌گرای نظام‌مند و بسامد واژه‌های دستوری است. همچنین دقت روش مبتنی بر واژه‌های دستوری از روش دستور نقش‌گرا بسیار بیشتر است. در شکل ۷، نتایج ارزیابی سامانه به طور کامل و برای هر نویسنده به طور مجزا نمایش داده شده است. علاوه بر این میزان کارایی تمامی ویژگی‌های تعریف شده در سامانه بررسی و کاراترین ویژگی‌ها در عملکرد سامانه انتخاب شد. این کار از طریق آزمون F تحلیل واریانس یک طرفه انجام شد.<sup>۳</sup> در این پژوهش مقادیر بردار ویژگی (کل ویژگی‌های مبتنی بر دستور نقش‌گرای نظام‌مند و ویژگی واژه‌های دستوری)، بر اساس نتایج به دست آمده از آزمون F تحلیل واریانس یک طرفه مرتب و ۱۵۰ ویژگی برتر انتخاب شد. بر اساس این معیار، بسامد نسبی واژه‌های دستوری «آن»، «رو» و «شد» به ترتیب دارای بالاترین امتیاز در میان کل ویژگی‌ها هستند.

جدول ۴. نتایج ارزیابی کلی سامانه بر اساس معیار دقت

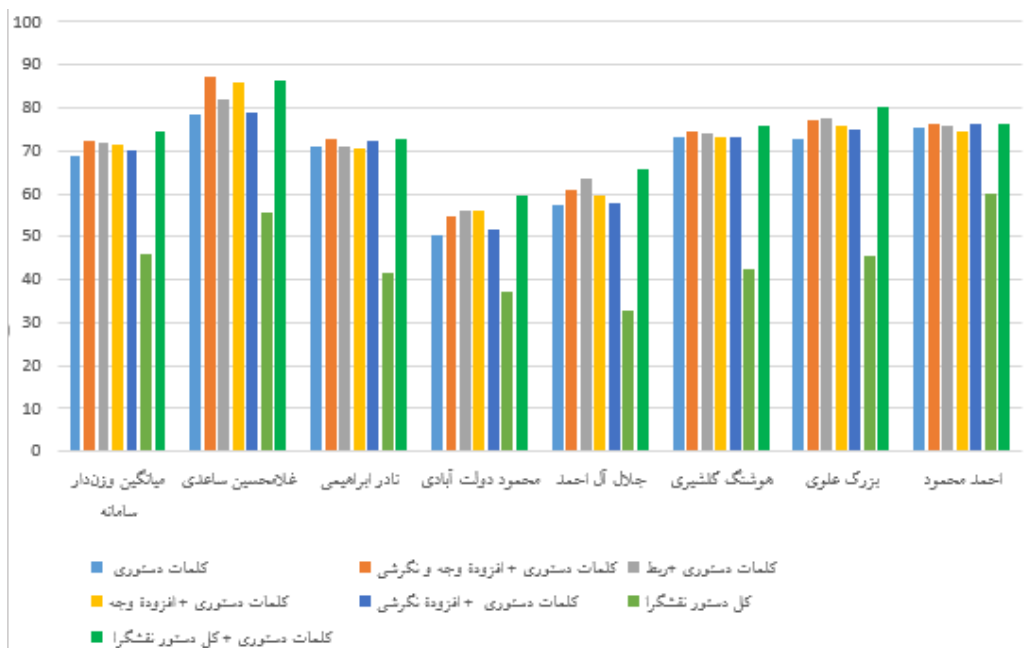
نام مجموعه ویژگی	کلمات دستوری	کلمات دستوری، افزوده وجه و نگرشی	کلمات دستوری و افزوده ربطی	کل دستور نقش‌گرا	کلمات دستوری و کل دستور نقش‌گرا
میانگین وزن دار	۶۸,۹۸	۷۲,۵۳	۷۲,۰۵	۴۶	۷۴,۳۱

<sup>1</sup> weighted mean

<sup>۲</sup> در محاسبه میانگین وزن دار یک مجموعه عامل نابرابر، برای هر یک از عامل‌ها، وزن یا ارزش معینی در نظر گرفته می‌شود و سپس آن عامل در وزن معین ضرب می‌گردد. آن‌گاه جمع این ارقام به دست آمده بر مجموع وزن‌ها تقسیم می‌شود.

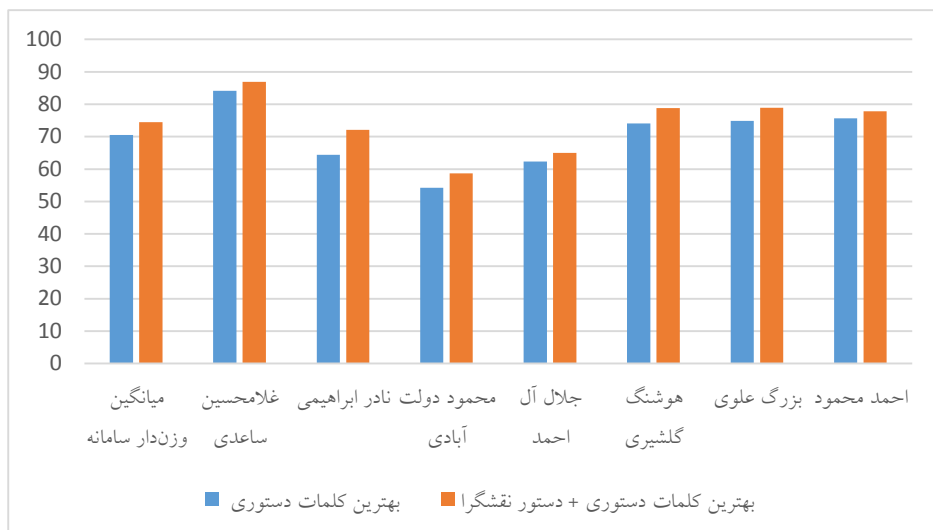
<sup>3</sup> ANOVA

<sup>۴</sup> آزمون تحلیل واریانس یک طرفه برای آزمون مقایسه میانگین یک متغیر کمی در بین بیش از دو گروه مستقل استفاده می‌شود.



شکل ۷. نمودار مقایسه‌ای ارزیابی سامانه بر اساس معیار دقت

پس از انتخاب بهترین ویژگی‌ها کارایی سامانه در دو حالت سنجیده شد: ۱- کلمات دستوری، ۲- کلمات دستوری (بهترین ویژگیها) و ویژگیهای مبتنی بر دستور نقشگرا. این نتایج در جدول ۵ گزارش شده است. همانطور که در شکل ۸ قابل مشاهده است، بهترین نتایج را ترکیب حالت بهترین کلمات دستوری و دستور نقشگرا به دست می‌دهد.



شکل ۸. نمودار مقایسه‌ای ارزیابی سامانه بر اساس معیار دقت (بهترین ویژگی‌ها)

جدول ۵. نتایج ارزیابی سامانه بر اساس معیار دقت (بهترین ویژگی‌ها)

نام مجموعه ویژگی	میانگین وزن‌دار
بهترین کلمات دستوری	۷۰/۴۷
بهترین کلمات دستوری و دستور نقشگرا	۷۴/۴۷

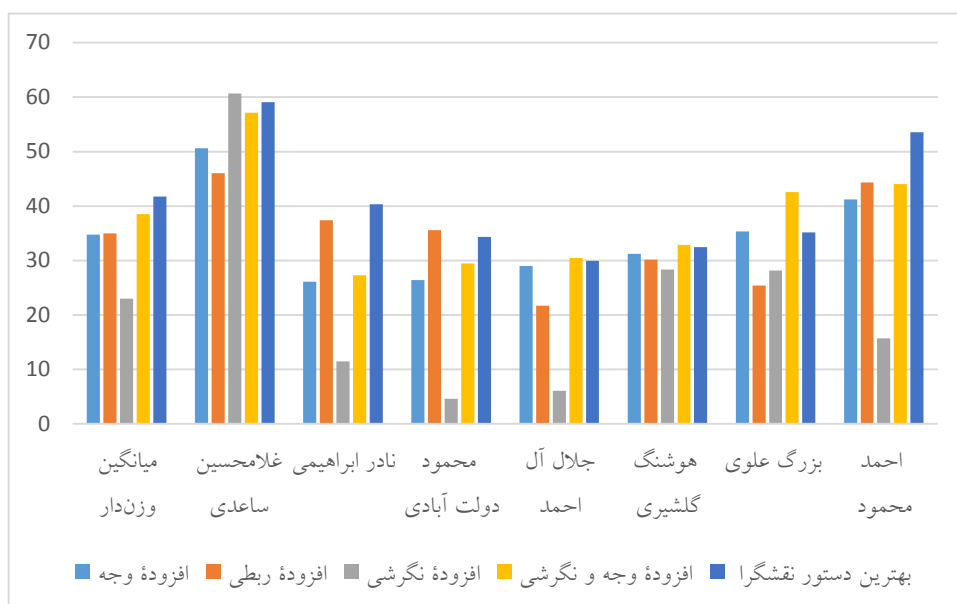
## ۶. بحث و نتیجه‌گیری

در این پژوهش برای تعیین هویت نویسنده متن یک سامانه خودکار به روش یادگیری عمیق معرفی شد. برای طراحی سامانه از ویژگی بسامد واژه‌های دستوری و مفاهیم دستور نقشگرای نظام‌مند استفاده شد. سپس عملکرد سامانه در سه حالت ویژگی‌های مبتنی بر واژه‌های دستوری، ویژگی‌های مبتنی بر دستور نقشگرای نظام‌مند و ترکیب این دو نوع ویژگی مورد سنجش و تحلیل قرار گرفت و بهترین ویژگی‌ها انتخاب شد. نتایج ارزیابی سامانه نشان می‌دهد روش محاسبه بسامد واژه‌های دستوری نسبت به روش مبتنی بر دستور نقشگرای نظام‌مند (شبکه نظام ربط، وجه‌نمایی و افزوده نگرشی) در تشخیص هویت نویسنده متن برتری دارد اما در صورتی که ویژگی‌های زبانی مبتنی بر دستور نقش‌گرای نظام‌مند هالیدی در کنار ویژگی بسامد واژه‌های دستوری به کار روند، کارایی سامانه نسبت به حالتی که تنها از ویژگی بسامد واژه‌های دستوری استفاده شود، افزایش می‌یابد.

به نظر می‌رسد برای این نتیجه که ویژگی بسامد نسبی واژه‌های دستوری کارایی بالاتری نسبت به ویژگی‌های مبتنی دستور نقشگرا در تفکیک سبک نویسندگان داشته، دلایل متفاوتی وجود داشته است. دلیل اول می‌تواند به عدم کنترل خودآگاه نویسنده در استفاده از واژه‌های دستوری مربوط باشد. نویسنده متن واژه‌های دستوری را به طور ناخودآگاه به کار می‌گیرد و برخلاف واژه‌های محتوایی کنترل خودآگاه بر استفاده کردن یا استفاده نکردن از آنها ندارد (گلشایی، ۱۳۹۸). از آنجا که تعیین مرز دقیق بین واژه‌های محتوایی و دستوری عملاً امکان‌پذیر نیست، می‌توان در خصوص واژه‌های دستوری یک پیوستار تعریف کرد. در این پیوستار هرچه قدر به سمت دستوری بودن واژه پیش رویم، کنترل خودآگاه در استفاده کردن از واژه‌ها کاهش می‌یابد. در این پژوهش بر اساس آزمون F تحلیل واریانس یک طرفه، ضمایر «آن»، «این»، «آنها»، «من» و «خود» ویژگی‌های بسیار خوبی در تمایز قائل شدن میان سبک نویسندگان تلقی می‌شوند. نتایج حاصل از پژوهش حاضر در مورد کاراترین ویژگی‌ها در تشخیص هویت نویسنده گواهی بر این مدعاست که ضمایر در این پیوستار به سمت دستوری بودن تمایل دارند و نویسنده کنترل خودآگاه در استفاده از آنها ندارد. بدیهی است که این پیوستار خود می‌تواند موضوع یک پژوهش جدید باشد.

دومین دلیل برای اینکه در تشخیص هویت نویسنده ویژگی‌های مبتنی دستور نقشگرا به طور مستقل موثرتر از واژه‌های دستوری عمل نکرده اند، می‌تواند این باشد که استفاده از واژه‌های دستوری به جای یکدیگر، تفاوت سبکی بیشتری ایجاد می‌کند. برای مثال تفاوت سبک در متونی که از واژه‌های «من» و «اینجانب» استفاده می‌کنند قابل توجه است؛ این در حالی است که برای مثال «مسلماً» و «مسلم است» که در زیرنظام جهت‌گیری از شبکه نظام وجه‌نمایی اولی «ضمنی» و دومی «صریح» تلقی می‌شوند، تفاوت سبکی چندانی ایجاد نمی‌کنند. این مطلب درباره «باید» و «نباید» که در زیرنظام قطبیت با یکدیگر متفاوتند نیز صادق است.

دلیل دیگر برای کارایی بیشتر واژه‌های دستوری نسبت به ویژگی‌های مبتنی دستور نقشگرا، می‌تواند اختلاف درصد حضور واژه‌های دستوری و کلمات مرتبط با سه نوع ویژگی دستور نقشگرا در پیکره باشد. به غیر از افزوده ربطی «و» از شبکه نظام حروف ربط که فراوانی بسیار بالایی در پیکره دارد، درصد حضور واژه‌های دستوری بیشتر از واژه‌های مرتبط با ویژگی‌های مبتنی دستور نقشگراست. برای بررسی بهتر ویژگی‌های مبتنی بر دستور نقشگرای هالیدی، هر یک از این سه شبکه نظام حروف ربط، وجه‌نمایی و افزوده نگرشی هم به طور مجزا و هم در ترکیب با یکدیگر مورد ارزیابی قرار گرفت و نتایج آنها با یکدیگر و با حالت بهترین ویژگی‌های سه نوع نظام مقایسه شد که در شکل ۹ قابل ملاحظه است.



شکل ۹. نمودار مقایسه‌ای کارایی سامانه با انواع ویژگی‌های دستور نقشگرا

همانطور که در نمودار ۹ مشاهده می‌شود، ترکیب بهترین ویژگی‌های مرتبط با هر سه نوع افزوده وجه، نگرشی و ربطی بالاترین دقت را در میان سایر حالات دارد. پس از آن، افزوده وجهی (افزوده وجه و افزوده نگرشی) قرار دارد. گواه دیگر در برتری افزوده وجهی نسبت به افزوده ربطی، درصد حضور ویژگی‌های مرتبط با افزوده وجهی در میان ویژگی‌های منتخب است (۳۹/۰۷ درصد). این درحالی است که درصد حضور ویژگی‌های مرتبط با افزوده ربطی در این فهرست ۱۱/۲۵ درصد است.

تحلیل دقیق‌تر بهترین ویژگی‌های افزوده وجهی نشان می‌دهد که ویژگی‌های مرتبط با زیرنظام‌های «مقدار» و «نوع» از شبکه نظام افزوده وجهی هم به صورت مستقل و هم در ترکیب با سایر زیرنظامها بسیار کارا هستند. زیرنظام‌های «قطبیت» و «جهت‌گیری» در ترکیب با سایر زیرنظامها به بهبود عملکرد سامانه کمک شایانی می‌کنند. در مورد زیرنظام وجه‌سازی (وجه احتمال-عادت) علاوه بر اینکه کل زیرنظام و زیرنظام «وجه احتمالی» در کارایی سامانه بسیار موثر هستند، زیرنظام «وجه عادت» بیشترین تأثیر را دارد. در زیرنظام تعدیل‌سازی (وجه تمایل-التزام) نیز علاوه بر اثرگذاری کل زیرنظام و زیرنظام «وجه التزامی»، زیرنظام «وجه تمایلی» بالاترین کارایی را دارد. در شبکه نظام افزوده نگرشی، زیرنظام گزاره‌ای خصوصاً از نوع «اخلاقی» و «فرض» بیشترین میزان اثربخشی را دارد.

در مورد نظام شبکه حروف ربط، بسامد افزوده‌های ربطی هر یک از سه زیرنظام «تشریحی»، «گسترشی» و «تفصیلی» (به طور مجزا) به کل افزوده‌های ربطی ویژگی‌هایی ارزشمند به شمار می‌روند. بر پایه ویژگی‌های منتخب، زیرنظام «گسترشی» بهترین نتایج را به دست می‌دهد. در این زیرنظام، بسامد نسبی افزوده‌هایی از نوع «افزایشی» و «تباینی» امتیاز بالایی دارند. در زیرنظام «تفصیلی»، زیرنظام «سببی-شرطی» بسیار موثر است. در این زیرنظام، زیرنظام «مشخص» خصوصاً حروف ربط از نوع «علت» و پس از آن «نتیجه» تأثیرگذارند. در زیرنظام «تشریحی»، افزوده‌های ربطی «بدلی» خصوصاً از نوع «نمونه‌سازی»، کارآمد هستند.

نتایج کلی این پژوهش همسو با نتایج مقاله مرجع (آرگامون و همکاران، ۲۰۰۷) است. در پژوهش آرگومان و همکاران (۲۰۰۷)، روش بسامد نسبی واژه‌های دستوری نسبت به روش دستور نقش‌گرا (شبکه نظام حروف ربط، شبکه نظام وجه‌نمایی و شبکه نظام افزوده نگرشی) برتری دارد و ترکیب روش واژه‌های دستوری و دستور نقش‌گرا بالاترین کارایی را دارد. در پژوهش آرگومان و همکاران (۲۰۰۷) روش دستور نقش‌گرا به افزایش کارایی سامانه کمک کرده است که با نتایج کلی پژوهش حاضر هم‌راستاست.



یکی از تفاوت‌های پژوهش حاضر با پژوهش آرگومان و همکاران (۲۰۰۷) این است که در پژوهش آرگومان تعداد واژه‌های دستوری برابر با ۶۷۵ واژه در نظر گرفته شده است؛ این در حالی است که در پژوهش حاضر تعداد کل واژه‌های دستوری برابر با ۱۹۷ واژه است و پس از مرحله یافتن بهترین ویژگی‌ها تعداد ۷۴ واژه از واژه‌های دستوری انتخاب شده است. آرگومان و همکاران (۲۰۰۷) با استفاده از الگوریتم طبقه‌بندی ماشین بردار پشتیبان<sup>۱</sup> و با ۶۷۵ واژه دستوری، تنها با ویژگی بسامد نسبی واژه‌های دستوری به دقتی در حدود ۸۵ درصد در تشخیص هویت نویسنده برای زبان انگلیسی دست یافته‌اند. در این پژوهش به منظور مقایسه نتایج کار با مقاله مرجع از ۶۶۴ واژه دستوری و الگوریتم ماشین بردار پشتیبان استفاده شد و به دقتی معادل با ۸۲٫۷۵ درصد برای زبان فارسی دست یافت که به نتایج پژوهش مرجع نزدیک است. در این مقایسه مشخص شد الگوریتم یادگیری بر روی داده دچار بیش‌برازش<sup>۲</sup> شده است<sup>۳</sup> و این مورد را می‌توان از نقاط ضعف پژوهش مرجع یاد کرد که البته در آن مقاله نیز چند بار به آن اشاره شده است. به همین سبب در پژوهش حاضر به منظور جلوگیری از بیش‌برازش تعداد واژه‌های دستوری بسیار کمتر از مقاله مرجع و برابر با ۱۹۷ کلمه در نظر گرفته شده است. فارغ از اینکه تعداد واژه‌های دستوری چقدر است، نتایج حاصل از افزودن ویژگی‌های مبتنی بر دستور نقشگرا به سامانه همسو با پژوهش آرگومان و همکاران است. در مقاله مرجع کارایی سامانه با افزودن ویژگی‌های مبتنی بر دستور نقشگرا، در حدود ۵ درصد بیشتر از حالتی است که صرفاً از واژه‌های دستوری استفاده شود. این افزایش کارایی برابر با افزایش کارایی پژوهش حاضر در حالت قرارگرفتن ویژگی‌های مبتنی بر دستور نقشگرا در کنار واژه‌های دستوری است (ارجاع به جدول ۴). تفاوت دیگر نتایج مقاله مرجع با پژوهش حاضر این است که ویژگی‌های مبتنی بر دستور نقشگرای هالیدی (به طور مستقل از واژه‌های دستوری) برای زبان فارسی به اندازه زبان انگلیسی در تشخیص هویت نویسنده کارایی نداشته‌اند. علاوه بر تفاوت‌های دو زبان فارسی و انگلیسی، به نظر می‌رسد تفاوت در حجم واژگان تعریف شده برای شبکه نظام‌های دستور نقشگرا از عوامل این تفاوت باشد.

شبکه نظام‌های مختلف دستور نقشگرا (شبکه نظام حروف ربط، شبکه نظام وجه‌نمایی و شبکه نظام افزوده نگرشی) به طور مستقل نتایج خوبی را در تعیین هویت نویسنده به دست نمی‌دهند. در

<sup>۱</sup> SVM

<sup>۲</sup> overfitting

<sup>۳</sup> بیش‌برازش به این معناست که الگوریتم داده‌های آموزشی را بسیار خوب یادگرفته است ولی قابلیت پیش‌بینی خوب داده‌های جدید را ندارد.

حقیقت بهترین نتایج، نتایجی است که از ترکیب ویژگیهای منتخب از هر سه نظام و ویژگی واژه‌های دستوری به دست می‌آید. گرچه رویکرد نقش‌گرای نظام‌مند هالیدی خصوصاً در فرانش بینافردی در مورد وجه، دسته‌بندی دقیقی ارائه می‌کند اما شبکه نظام‌های این دستور به صورت مستقل از واژه‌های دستوری عملکرد خوبی در تشخیص هویت نویسنده متن ندارند و تنها در کنار واژه‌های دستوری به افزایش کارایی سامانه یاری می‌رسانند.

از کاستی پژوهش حاضر، می‌توان به این مساله اشاره کرد که به غیر از تعداد محدودی از افعال در فهرست افزوده وجهی و افعال کمکی، مقوله فعل در طراحی سامانه هویت نویسنده مورد توجه قرار نگرفت. به عنوان کارهای آتی می‌توان حجم واژگان برای افزوده‌های ربطی و همینطور افزوده‌های وجهی خصوصاً با مقوله فعل را گسترش داد و به این شکل به پژوهش‌های حوزه دستور نقش‌گرای نظام‌مند هالیدی کمک نمود. به عنوان کارهای آتی می‌توان الگوریتم استخراج سه نوع ویژگی دستور نقش‌گرا و همینطور واژه‌های دستوری را بر روی سایر آثار داستانی نیز اعمال کرد و بسامد انواع ویژگی‌ها را به دست آورد و از آن در جهت نقد ادبی و سبک‌شناسی برای اهداف زبانی استفاده کرد.

### منابع

- ابراهیمی، نادر. (۱۳۹۹). *بر جاده‌های آبی سرخ*. تهران: روزبهان.
- ابراهیمی، نادر. (۱۳۷۴). *یک عاشقانه آرام*. تهران: روزبهان.
- آل احمد، جلال. (۱۳۴۶). *نفرین زمین*. تهران: انتشارات فردوس.
- آل احمد، جلال. (۱۳۵۰). *پنج داستان*. تهران: انتشارات فردوس.
- جعفری، آریتا. (۱۳۸۸). بررسی افزوده‌ها در زبان فارسی: بر اساس رویکردهای نقشی و صوری. *دستور (ویژه نامه نامه فرهنگستان)*، ۵(۱)، ۱۵۵-۱۲۸.
- حسین حمه، همزه، علی‌اکبری، نسرین، و کریمی، یادگار. (۱۴۰۰). بررسی وجه و وجهیت در کردی سورانی: تحلیلی نقش‌گرا. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۹(۴)، ۱-۲۳.
- دولت‌آبادی، محمود. (۱۳۹۵). *روزگار سپری شده مردم سالخورده*. تهران: نشر چشمه.
- دولت‌آبادی، محمود. (۱۴۰۱). *کلیدر (چاپ ۳۷)*. تهران: فرهنگ معاصر.
- ساعدی، غلامحسین. (۱۳۹۷). *گریبه در شهر*. تهران: انتشارات نگاه.
- ساعدی، غلامحسین. (۱۳۷۷). *آشفته‌حالان بیداریخت*. تهران: انتشارات نگاه.
- شمس‌فرد، مهرنوش. و بیجن‌خان، محمود. (۱۴۰۱). پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی. تهران: انتشارات سمت.

- عارفی، سمیه، بصیری، محمد احسان، و روزمند، امید. (۱۴۰۰). انتخاب ویژگی برای شناسایی نویسنده در متون کوتاه برخط فارسی. *فناوری اطلاعات و ارتباطات ایران*، ۳۵-۵۶.
- علوی، بزرگ. (۱۳۸۶). *ورق پاره‌های زندان*. تهران: انتشارات نگاه.
- علوی، بزرگ. (۱۳۹۹). *گیله‌مرد*. تهران: انتشارات نگاه.
- فرهمندپور، زینب، نیک‌مهر، هومن، منصوری‌زاده، محرم، و طیب‌زاده، امید. (۱۳۹۱). یک سیستم نوین هوشمند تشخیص هویت نویسنده فارسی زبان بر اساس سبک نوشتاری. *محاسبات نرم*، (۲)، ۲۶-۳۵.
- گلشائی، رامین. (۱۳۹۸). واژه‌های دستوری به‌مثابه نشانگرهای گویش فردی: رویکردی پیکره‌ای به شناسایی هویت نویسنده در زبان فارسی. *جستارهای زبانی*، ۵۱(۱۰)، ۲۹۳-۳۱۷.
- گلشیری، هوشنگ. (۱۳۷۰). *در ولایت هوا*. استکهلم: انتشارات عصر جدید.
- گلشیری، هوشنگ. (۱۳۵۰). *کریستین و کید*. تهران: کتاب زمان.
- گلشیری، هوشنگ. (۱۴۰۰). *شازده احتجاب* (۱۸). تهران: نیلوفر.
- محمود، احمد. (۱۳۵۳). *همسایه‌ها*. تهران: انتشارات امیرکبیر.
- محمود، احمد. (۱۳۸۱). *غریبه‌ها و پسرک بومی*. تهران: معین.
- میرزایی، آزاده. و صفری، پگاه. (۱۳۹۴). ساخت واژه-متن‌های تخصصی و عمومی زبان فارسی بر اساس بسامدگیری واژه‌های نقشی و محتوایی. *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*. ۱۷۵-۱۹۱.
- میرزایی، آزاده. (۱۳۹۷ الف). بازتعریف مفاهیم بند پایه و بند پیرو براساس رویکرد نقش‌گرا. *زبان و زبان‌شناسی*، ۱۳(۲۶). ۱۱۷-۱۳۲.
- (۱۳۹۷ ب). *آشنایی با زبان‌شناسی پیکره‌ای*. تهران: انتشارات دانشگاه علامه طباطبائی.
- (۱۴۰۰). رابطه قطبیت و وجهیت بندی در زبان فارسی. *مطالعات زبان‌ها و گویش‌های غرب ایران*، ۹(۱)، ۱۱۳-۱۳۵.

## References

- Al-e-Ahmad, J. (1967). *The cursing of the land*. Tehran: Ferdous. (In Persian)
- Al-e-Ahmad, J. (1971). *Five stories*. Tehran: Ferdous. (In Persian)
- Alavi, B., (2007). *Scrap Papers from Prison*. Tehran: Negah. (In Persian)
- Alavi, B., (2020). *Gilemard*. Tehran: Negah. (In Persian)
- Arefi, S., Basiri, M., & Roozmand, O. (2021). Feature selection for author identification of Persian online short texts. *Journal of Information and Communication Technology*, 47, 35-57. (In Persian)

- Argamon, S., & Koppel, M. (2013). A systemic functional approach to automated authorship analysis. *Brooklyn Journal of Law & Policy*, 21, 299.
- Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6), 802-822. <https://doi.org/10.1002/asi.20553>.
- Assi, S. M. (1997). Farsi linguistic database (FLDB). *International journal of Lexicography*, 10(3), 5.
- Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language resources and evaluation*, 45, 143-164. <https://doi.org/10.1007/s10579-010-9132-x>.
- Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4), 431-447. <https://doi.org/10.1093/applin/25.4.431>.
- Dabagh, R. M. (2007). Authorship attribution and statistical text analysis. *Advances in Methodology and Statistics*, 4(2), 149-163. <https://doi.org/10.51936/uvjx7198>.
- Daroonch, A. H., & Shariati, A. (2014). Metrics for evaluation of the author's writing styles: Who is the best? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(3). <https://doi.org/10.1063/1.4895468>.
- Dowlatabadi, M. (2022). *Kelidar*. Tehran: Farhange Moaser. (In Persian)
- Dowlatabadi, M. (2016). *The Elderly People's Elapsed Time*. Tehran: Cheshmeh. (In Persian)
- Ebrahimi, N. (1995). *A quiet Romance*. Tehran: Rouzbahan. (In Persian)
- Ebrahimi, N. (2020). *On the blue and red paths*. Tehran: Rouzbahan. (In Persian)
- Farahmandpoor, Z., Nikmehr, H., Mansoorizade, M., & Tabibzadeh Ghamsary, O. (2013). A Novel Intelligent Persian Authorship System based on Writing Style. *Soft Computing Journal*, 1(2), 26-35. (In Persian)
- Golshaie, R. (2019). Function Words as Idiolect Markers: A Corpus-based Approach to Authorship Attribution in Farsi. *Language Related Research*. 10 (3) :293-317. (In Persian)
- Golshiri, H. (1991). *Dar Velavat-e Hava*. Stockholm: Asr-e Jadid. (In Persian)
- Golshiri, H. (1971). *Christine and Kid*. Tehran: Ketab-e Zaman. (In Persian)
- Golshiri, H. (2021). *Prince Ehtejab*. (18). Tehran: Niloufar. (In Persian)
- Halliday, M. A. K., & Matthiessen, C. M. M. (2014). *Halliday's introduction to functional Grammar*. Forth Ed. Oxon: Routledge.
- Hussein Hama, H., Ali-akbari & Karimi, Y. (2022). Mood and Modality in Sorani Kurdish: A Functional Analysis. *Researches in Western Iranian Languages and Dialects*, 9(4), 1-23. <https://doi.org/10.22126/JLW.2021.6008.1504>. (In Persian)
- Jafari, A. (2008). An analysis of adjuncts in Persian: A syntacto-discoursal approach. *Dastoor*, 5(1), 128-155. (In Persian)
- Mahmoud, A. (2002). *The Strangers and The Little Native Boy*. Tehran: Moin. (In Persian)
- Mahmoud, A. (1974). *The Neighbors*. Tehran: Amirkabir. (In Persian)
- Mandenthal, T. C. (1887). The characteristics curves of composition science. *Science*, 9(214s), 237-246.
- Martinez-Galicia, J. A., Embarcadero-Ruiz, D., Ríos-Orduña, A., & Gómez-Adorno, H. (2022). Graph-Based Siamese Network for Authorship Verification. *CLEF 2022 Labs and Workshops, Notebook Papers, Bologna, Italy*.
- Mirzaei, A. (2018). *An Introduction to Corpus Linguistics*. Tehran: Allameh Tabataba'i University Press. (In Persian)

- Mirzaei, A. (2017). Redefining the Concepts of Dependent and Independent Clauses According to a Functional Approach. *Language and Linguistics*, 13(26), 117-132. (In Persian)
- Mirzaei, A. (2021). The Relationship between Polarity and Clausal Modality in Persian. *Journal of Western Iranian Languages and Dialects*, 9(1), 113-135. <https://doi.org/10.22126/jlw.2020.5372.144>. (In Persian)
- Mirzaei, A., & Safari, P. (2018). Persian discourse treebank and coreference corpus. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*, Miyazaki, Japan.
- Mirzaei, A. & Safari, P. (2014). Building specialized and general documents in Persian based on the frequency of function and content words. In *proceeding of 1st National Conference on Corpus Linguistics (175-192)*, Tehran: NevisehParsi. (In Persian)
- Najafi, M., & Tavan, E. (2022). Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis. *CLEF 2022 Labs and Workshops, Notebook Papers, Bologna, Italy*.
- Matthiessen, C. M., & Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. London and New York: Frances Pinter Publishers.
- Plecháč, P. (2021). Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns. *Digital Scholarship in the Humanities*, 36(2), 430-438. <https://doi.org/10.1093/lc/fqaa032>.
- Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi and B. Webber (2008). The Penn discourse treebank 2.0, In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.
- Radford, A. (2004). *Minimalist syntax: Exploring the structure of English*: Cambridge University Press.
- Refaeilzadeh, P., Tang, L., & Liu, H. (2009). *Cross-Validation*. Encyclopedia of database systems, 5, 532-538 . Boston: Springer. [https://doi.org/10.1007/978-0-387-39940-9\\_565](https://doi.org/10.1007/978-0-387-39940-9_565).
- Sa'edi, Gh. (2018). *Stranger in the Town*. Tehran: Negah. (In Persian)
- Sa'edi, Gh. (1998). *Ashoftehalan-e Bidarbakht*. Tehran: Negah. (In Persian)
- Segarra, S., Eisen, M., & Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, 63(20), 5464-5478. <https://doi.org/10.1109/TSP.2015.2451111>.
- Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.
- Shamsfard, M., Hesabi, A., Fadaei, H., Mansoori, N., Famian, A., Bagherbeigi, S., . Assi, S. M. (2010). Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th global WordNet conference, Mumbai, India* (Vol. 29).

- Shamsfard, M. & Bijankhan, M., (2022). *Text and Speech Processing for the Persian Language: The State of the Art and A Brief Review of the Theoretical Foundations*. Tehran: Samt.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), 538-556. <https://doi.org/10.1002/asi.21001>
- Teich, E. (1995). A proposal for dependency in systemic functional grammar-metaseiosis in computational systemic functional linguistics. *Unpublished doctoral dissertation, University of the Saarland and GMD/IPSI, Darmstadt, Germany*.
- Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 8384-8395.
- Weerasinghe, J., Singh, R., & Greenstadt, R. (2021). Feature Vector Difference based Authorship Verification for Open-World Settings. In *Proceedings of working of the evaluation Forum, Bucharest, Romania*, 2201-2207.