Introducing a Probabilistic – structural Method for Grapheme-to-phoneme Conversion in Persian

Elham Alayiaboozar
PhD in General linguistics; Assistant Professor;
Iranian Research Institute for Information Science and Technology (IranDoc);
Corresponding Author alayi@irandoc.ac.ir

Mahmood Bijankhan
PhD in General Linguistics; Professor; University of Tehran;
mbjkhan@ut.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in SCOPUS, ISC, & LISTA
Vol. 31 | No. 4 | pp: 1121-1141
Summer 2016

Received: 2, Sep. 2015 **Accepted: 29, Nov. 2015**

Abstract: Persian writing system deviates from the ideal one due to the lack of one-to-one correspondence between graphemes and phonemes. The present study deals with this question that in spite of the absence of short vowels in Persian writing system and one-to-many and many-to-one relationships between the graphemes and phonemes, how can Persian speakers read out of vocabulary words? This study introduces a probabilistic-structural method that Persian speakers use to read out of vocabulary words in which structural information (including Persian morphology and morphophonemic rules) as well as Arabic morphological templates are considered. In order to test how the introduced method works, Persian speakers were asked to read a list out of vocabulary words. The mentioned list was used by ID3 and MLP (two methods which are used in machine learning) as input, then the outputs of the method and those of ID3 and MLP were compared with Persian speakers` pronunciations; the results proved that the introduced method functions similar to Persian speakers in reading out of vocabulary words.

Keywords: Out of Vocabulary Words, Probabilistic-Structural Method, Morphophonemic Rules, Arabic Morphological Templates

معرفی مدلی ساختاری-احتمالاتی برای تبدیل حرف به واج در متون فارسی

الهام علایی ابوذر
دکتری زبان‌شناسی همگانی؛ استادیار؛
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
پدیدآور رابط alayi@irandoc.ac.ir

محمود بی‌جن‌خان
دکتری زبان‌شناسی همگانی؛ استاد؛ دانشگاه تهران؛
mbjkhan@ut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹٤/۰٦/۱۱ **پذیرش:** ۱۳۹٤/۰۹/۰۸ مقاله برای اصلاح به مدت ۴ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۱ | شمارهٔ ٤ | ص‌ص ۱۱۲۱-۱۱٤۱
تابستان ۱۳۹۵

چکیده: در نظام‌های نوشتاری، رابطهٔ یک به یک میان واج‌ها و نگاره‌ها همواره برقرار نیست. از آنجا که در نظام نوشتاری فارسی واکه‌های کوتاه اغلب فاقد صورت نوشتاری هستند، تعداد حالت‌های ممکن خواندن کلمات خارج از واژگان افزایش می‌یابد و به این ترتیب، عمق خط فارسی زیاد در نظر گرفته می‌شود. اما با وجود چنین ویژگی‌هایی در خط فارسی، فارسی‌زبانان هنگام خواندن کلمات فارسی موجود در واژگان ذهنی خود و کلماتی که برای اولین بار با آن‌ها در متون گوناگون مواجه می‌شوند، قادرند رشتهٔ حروف را تبدیل به واج کنند. این پژوهش نشان می‌دهد که فارسی‌زبانان هنگام خواندن، با استفاده از روشی ساختاری-احتمالاتی، رشتهٔ حروف را به رشتهٔ واج‌ها تبدیل می‌کنند. منظور از بخش ساختاری روش، استفادهٔ فارسی‌زبانان از اطلاعات زبانی از قبیل: ساخت‌واژهٔ فارسی، قواعد واژواجی فارسی و آشنایی با صورت نوشتاری و تلفظ کلماتی است که با الگوهای ساخت‌واژی عربی مطابقت دارند. منظور از بخش احتمالاتی، در نظرگرفتن احتمال وقوع واکه‌های کوتاه با توجه به بافت نوشتاری است که این واقعیت می‌تواند فارغ از اطلاعات زبانی فارسی‌زبانان صورت پذیرد. در تحقیق حاضر مدلی ساختاری- احتمالاتی معرفی و عملکرد آن با نرم‌افزارهای تبدیل حرف به واج فارسی مقایسه شده است. به‌طور کلی، این نتیجه به دست آمد که عملکرد مدل ساختاری-احتمالاتی پژوهش برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان، در مقایسه با نرم‌افزارهای تبدیل حرف به واج فارسی بهتر و به تلفظ فارسی‌زبانان نزدیک‌تر است.

کلیدواژه‌ها: کلمات خارج از واژگان، قواعد واژواجی، الگوهای ساخت‌واژی عربی، مدل ساختاری-احتمالاتی

۱. مقدمه

در یک نظام نوشتاری آرمانی، رابطهٔ یک به یک میان نشانه‌های نوشتاری و واج‌ها وجود دارد، اما ارتباط نشانه‌های نوشتاری و واج‌ها لزوماً مستقیم نیست. با توجه به فقدان واکه‌های کوتاه در نظام نوشتاری زبان فارسی، وجود علائم زیروزبری که نگاره‌های وابسته محسوب می‌شوند، و رابطهٔ چند به یک و یک به چند میان نگاره‌ها[1] و واج‌ها (رابطهٔ یک به چند مانند: حرف < و> که با واج‌های/v/ ,/u/ ,/o/ مرتبط است و رابطهٔ چند به یک مانند: حروف <ط> و <ت> که با واج /t/ ارتباط دارند)، می‌توان گفت که نظام نوشتاری فارسی، از نظام نوشتاری آرمانی فاصله می‌گیرد و عمق خط فارسی (که ناظر بر رابطه میان صورت نوشتاری و نمود آوایی کلمات است) زیاد در نظر گرفته می‌شود (علایی و بی‌جن‌خان ۱۳۹۲). بنابراین، برای هر کلمهٔ خارج از واژگان (کلماتی که گویشورانی که توانایی خواندن و نوشتن دارند، برای اولین بار با آن‌ها در متون گوناگون مواجه می‌شوند) بیش از یک حالت ممکن تلفظی را می‌توان تصور کرد. به‌عنوان مثال، برای کلمهٔ خارج از واژگان <شناسدان> می‌توان این حالت‌های ممکن تلفظی را تصور کرد:

/ʃonɒsedɒn/، /ʃanɒsedɒn/، /ʃanɒsodɒn/، /ʃenɒsodɒn/، /ʃanɒsdɒn/، /ʃonɒsdɒn/، /ʃenɒsdɒn/،
/ʃenɒsedɒn/، /ʃonɒsodɒn/، /ʃenɒsadɒn/، /ʃonɒsadɒn/، /ʃanɒsadɒn/،

این در حالی است که فارسی‌زبانان در برخورد با کلمات خارج از واژگان ذهنی خود، چنین حالت‌های ممکن تلفظی، با این تعداد زیاد را برای هر کلمه در نظر نمی‌گیرند و معمولاً یک تلفظ از آن کلمه ارائه می‌دهند؛ زیرا از اطلاعات زبانی موجود در زبان فارسی برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان استفاده می‌کنند. یکی از این اطلاعات، اطلاعات ساخت‌واژی است. فارسی‌زبانان در واژگان ذهنی خود تلفظ تک‌واژهای آزاد و وابسته دارند. بنابراین، در برخورد با کلمات خارج از واژگان ذهنی خود ابتدا سعی می‌کنند تک‌واژهای موجود در کلمه را تفکیک کنند. به‌عنوان مثال، کلمهٔ <شناسدان>* را به دو تک‌واژ <شناس> و < دان> تفکیک می‌کنند و

---

۱. grapheme نگاره، کوچک‌ترین واحد نوشتاری است که میان یک جفت کمینهٔ نوشتاری تمایز ایجاد می‌کند. با طرح مفهوم نگاره می‌توان گفت علائم زیروزیری نیز نگاره محسوب می‌شود. به‌عنوان مثال، جفت‌های کمینهٔ نوشتاری مانند <تار> و < نار> تنها در تعداد نقطه‌ها از هم متمایز می‌شوند. اگر یک نقطه از قطعهٔ نوشتاری <تار> کم شود، تبدیل به قطعهٔ نوشتاری < نار> می‌شود. علائم زیروزبری مانند «تنوین»، «مد» و «تشدید» نیز نگاره محسوب می‌شوند. به‌عنوان مثال، در جفت‌های کمینهٔ نوشتاری < بنا> /bɑnɒ/ و < بنّا> /bɑnnɒ/، < اجر> /ʔɑdʒr/ و < آجر> /ʔɒdʒor/، <حکماً> /hokmɑn/ و < حکما>/hokɑmɒ/ ، به‌ترتیب، حضور یا فقدان نگاره‌های تشدید، مد و تنوین چنین جفت‌های کمینه‌ای را ایجاد کرده است. (علایی و بی‌جن‌خان ۱۳۹۲).

از آنجا که در واژگان ذهنی خود تلفظ این دو تک‌واژ را دارند، هیچ‌گاه تلفظ‌هایی مانند /ʃanpsedɒn/ و یا /ʃenpsodɒn/ * را که صورت‌های ممکن تلفظی این کلمه هستند، ارائه نمی‌دهند. بنابراین، به کمک تلفظ تک‌واژهای موجود در واژگان ذهنی‌شان، سعی در کم‌کردن عمق خط فارسی دارند. آشنایی با صورت نوشتاری وندهای تصریفی و اشتقاقی نیز باعث کم‌شدن عمق خط فارسی می‌شود. به‌عنوان مثال، فارسی‌زبانان صورت نوشتاری وند تصریفی <می-> را می‌شناسند. بنابراین، چون تلفظ مربوط به وندهای تصریفی و اشتقاقی را نیز در واژگان ذهنی خود دارند، هیچ‌گاه کلمه‌ای مانند <می‌پرداختمش> را به‌صورت /mejpardɒxtamaʃ/ * و یا /mojpardɒxtamaʃ/ * تلفظ نمی‌کنند. از دیگر اطلاعات زبانی که فارسی‌زبانان در جهت کم‌کردن عمق خط فارسی از آن استفاده می‌کنند، دانستن قواعد حرف‌نویسی و قواعد واژواجی زبان فارسی است. فارسی‌زبانان می‌دانند که هنگام اضافه‌شدن اکثر وندها به ستاک، تغییری در تلفظ ستاک و یا وند صورت نمی‌گیرد و بنابراین، این مسئلهٔ قابل پیش‌بینی نیز باعث کاهش عمق خط فارسی می‌شود. به‌عنوان مثال، هنگام اضافه‌شدن پسوند <بان> به تک‌واژ <هنر> تغییری در تلفظ ستاک یا وند صورت نمی‌گیرد و به‌صورت /honarbɒn/ تلفظ می‌شود، اما در برخی موارد در مرز تک‌واژها تغییرات واجی صورت می‌گیرد. به‌عنوان مثال، اگر پیشوند <ب> /be/ به بن مضارعی که با <ا> شروع می‌شود، اضافه شود، در تلفظ، /be/ به [bi] تبدیل می‌شود، زیرا همخوان میانجی[j] میان /i/ و /ɒ/ درج می‌شود و تلفظ [bijɒ] به‌دست می‌آید. در خط، <ی> میان <ب> و <ا> ظاهر می‌شود (<بیا>). دانستن چنین قواعد حرف‌نویسی نیز باعث کم‌شدن عمق خط فارسی می‌شود. آشنایی با صورت نوشتاری و تلفظ کلماتی که منشأ عربی دارند نیز از اطلاعات زبانی است که فارسی‌زبانان در جهت کم‌کردن عمق خط فارسی از آن استفاده می‌نمایند (بی‌جن‌خان و علایی ۱۳۹۲)؛ فارسی‌زبانان با کلماتی که از الگوهای ساخت‌واژی عربی مانند <مفعول>، <فاعل> و <فعیل> و ... تبعیت می‌کنند، آشنا هستند و در واژگان ذهنی خود کلماتی از قبیل <مقصود> بر اساس الگوی ساخت‌واژی عربی <مفعول>، <شاغل> بر اساس الگوی ساخت‌واژی عربی <فاعل> یا <فهیم> بر اساس الگوی ساخت‌واژی <فعیل> را دارند. بنابراین، در برخورد با کلماتی که خارج از واژگان ذهنی‌شان است، اما صورت نوشتاری مشابه صورت نوشتاری این الگوهای ساخت‌واژی عربی دارند، می‌توانند تلفظی شبیه تلفظ الگوهای ساخت‌واژی عربی ارائه دهند. به‌عنوان مثال، در برخورد با کلمهٔ <شاسل>* آن را با الگوی ساخت‌واژی عربی «فاعل» مطابقت می‌دهند. بنابراین، آن را /ʃɒsol/ *و یا /ʃɒsal/ * تلفظ نمی‌کنند. نگارنده در تحقیق حاضر از اطلاعات بخش ساختار زبان فارسی (اطلاعات مربوط به قواعد واژواجی زبان فارسی و اطلاعات مربوط به الگوهای ساخت‌واژی عربی قرضی در زبان فارسی) برای ارائهٔ برون‌داد واجی کلمات

خارج از واژگان، که تک‌واژهای آن‌ها در واژگان موجود است یا صورت نوشتاری آن‌ها با صورت نوشتاری الگوهای ساخت‌واژی عربی مطابقت دارد، استفاده کرده است.

«ونزکی» معتقد است که خوانندگان متون، آگاهی بالایی از فراوانی‌های نسبی وقوع حروف الفبا دارند و قادرند بدون آموزش دیدن، اطلاعات مربوط به فراوانی وقوع و ترتیب نشانه‌ها را استخراج کنند و میان آن‌ها تفاوت قائل شوند و ترکیبات با فراوانی وقوع بالا را سریع‌تر تشخیص دهند (Venezky 2004). «کسلر و تری‌من» نیز بر این باورند که توزیع واج‌ها درون هجاها از الگوهای احتمالاتی نیز تبعیت می‌کند و تشخیص و تولید یک کلمه می‌تواند تحت تأثیر تعداد کلمات دیگری قرار بگیرد که تلفظ یا املاء کم‌وبیش یکسانی دارند (Kessler & Treiman 1997). نگارنده معتقد است، فارسی‌زبانان در برخورد با کلمات خارج از واژگانی که دارای بخش/بخش‌هایی هستند و با الگوهای ساخت‌واژی عربی مطابقت نمی‌کنند، از اطلاعات مربوط به فراوانی وقوع واکه‌های کوتاه در بافت نوشتاری برای ارائهٔ برون‌داد واجی این کلمات استفاده می‌کنند. از آنجا که تحقیق حاضر رویکردی رایانه‌ای نسبت به تبدیل حرف به واج دارد، علاوه بر بررسی‌های ساختاری، لازم است مدلی برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان ارائه شود که در آن، هم از اطلاعات زبانی، مانند تأثیر الگوهای ساخت‌واژی عربی در زبان فارسی و قواعد واژواجی که نوعاً در اتصال پیشوندها و پسوندها به ستاک عمل می‌کنند و باعث تغییرات واجی درون تک‌واژها یا مرز اتصال دو تک‌واژ می‌شوند، استفاده شود و هم از احتمالات. این مدل، ساختاری- احتمالاتی نامیده شده است. در این مقاله ابتدا پیشینه‌ای ارائه می‌شود از تحقیقاتی که در زمینهٔ خط عربی، سامانه‌های تبدیل حرف به واج، خط فارسی، و مسائل مربوط به پردازش خط فارسی شده است. بخش ۳، روش تحقیق است که در آن قسمتی از برون‌دادهای واجی کلمات خارج از واژگان که توسط فارسی‌زبانان ارائه شده است، به‌عنوان نمونه، آورده شده است تا بتوان آن‌ها را با برون‌دادهای واجی مدل ساختاری- احتمالاتی تحقیق و نیز برون‌دادهای واجی حاصل از به‌کاربردن الگوریتم تبدیل حرف به واج برخی از نرم‌افزارهایی که در فارسی استفاده شده، مقایسه کرد. در بخش ۴، مدل ساختاری- احتمالاتی تحقیق و مباحث مربوط به آن ارائه می‌گردد. در بخش ۵، عملکرد نرم‌افزارهای تبدیل حرف به واج فارسی با عملکرد فارسی‌زبانان و مدل ساختاری- احتمالاتی تحقیق، به‌طور مختصر، مقایسه می‌شود و در نهایت، در بخش ۶، نتیجه‌گیری ارائه می‌گردد.

۲. پیشینهٔ تحقیق

«رابرت ترمل» به بررسی نحوهٔ خواندن کلمات ناآشنای چندهجایی می‌پردازد. منظور از

کلمات ناآشنا در این مقاله، کلماتی است که خوانندگان برای اولین بار در متون مختلف با آن‌ها مواجه می‌شوند. نگاره‌هایی که در یک کلمهٔ چندهجایی ظاهر می‌شوند به لحاظ واجی به چند طریق تفسیر می‌شوند که به تقطیع هجایی، ساخت‌واژی یا نوشتاری و به طبقهٔ نحوی و الگوی تکیه وابسته است (Trammel 1990).

«فان دن بوش» و همکاران مفهومی کمّی از عمق خط را ارائه داده‌اند. تعیین عمق خط مربوط به هر زبان از طریق اندازه‌گیری میزان پیچیدگی‌های موجود در انطباق نگاره به واج در آن زبان صورت می‌پذیرد. آن‌ها معتقدند که با استفاده از روش‌های رایانه‌ای داده‌بنیاد می‌توان اطلاعاتی دربارهٔ عمق نظام نوشتاری به‌دست آورد. برای استخراج بازنمایی واجی از صورت نوشتاری دو کار باید انجام شود: ۱) صورت نوشتاری به نگاره‌های تشکیل‌دهنده تجزیه شود. ۲) هر نگاره به واج متناظرش مربوط شود. در این مقاله سه الگوریتم یادگیری معرفی می‌شود: الف) استخراج تطابق نگاره و واج[1]، ب) درخت تصمیم[2] و ج) استنتاج تشابه‌بنیاد[3]. هدف این مقاله این است که نشان دهد آیا به‌کاربردن سه الگوریتم یادگیری داده‌بنیاد دربارهٔ سه نظام نوشتاری انگلیسی، فرانسه و هلندی، تفاوتی در عمق نوشتاری این سه زبان نمایان می‌سازد یا نه؟ برای این منظور، یکی از الگوریتم‌های یادگیری در حوزهٔ تجزیهٔ نگاره‌ای آموزش دیده است و دو الگوریتم دیگر در حوزهٔ نگاشت نگاره به واج آموزش دیده‌اند (Van den Bosch et al. 1994).

«مگردومیان» با استفاده از نظامی که در آن از ابزار زیراکس[4] مبتنی بر نظریهٔ حالت‌های محدود[5] استفاده شده، دو سطح تجزیه و تحلیل ساخت‌واژی برای زبان فارسی را توضیح می‌دهد. وی معتقد است که زبان فارسی از نقطه‌نظر تحلیل‌های رایانه‌ای دارای پیچیدگی‌هایی است. در این مقاله به تفصیل به چنین پیچیدگی‌هایی پرداخته شده است و راه حلی در چارچوب نظام ساخت‌واژه مبتنی بر نظریهٔ حالت‌های محدود پیشنهاد شده است. به نظر «مگردومیان» مزیت و اهمیت یک نظام مبتنی بر نظریهٔ حالت‌های محدود در این است که توانایی پردازش قطعات چندکلمه‌ای را در تحلیل‌گر خود دارد و نویسهٔ نیم‌فاصله را به‌عنوان یک فاصله در قواعد تقطیع کلمات به تک‌واژها در نظر می‌گیرد. با استفاده از این روش می‌توان تک‌واژهای غیرچسبان را در فارسی، قسمتی از قطعات چندکلمه‌ای در واحد واژگانی lex grammar در نظر گرفت و این کار این امکان را فراهم می‌کند که با هر دو شکل (تک‌واژهای چسبان و غیرچسبان)، در قسمت تحلیل‌گر ساخت‌واژی، به‌طور یکسان برخورد شود و دیگر نیازی به واحد مستقلی به نام

1. grapheme-phoneme correspondences extraction
2. decision tree
3. Similarity- based reasoning (SBR)
4. xerox
5. finite state theory

پیش‌پردازشگر نباشد. همچنین، نیاز به تأخیرانداختن تحلیل‌های تک‌واژهای غیرچسبان به سطح نحو نیز برطرف خواهد شد (Megerdoomian 2004).

«باک‌والتر» مسائل مربوط به خط عربی را که در تحلیل ساخت‌واژی و برچسب‌دهی[1] تعداد ۵۴۲۵۴۳ کلمهٔ عربی استخراج‌شده از سه پیکره طی سال‌های ۲۰۰۴-۲۰۰۲ با آن مواجه شدند، بررسی کرده است. این تحلیل و پردازش ساخت‌واژی با استفاده از پردازشگر ساخت‌واژی عربی «باک‌والتر»[2] انجام شده است. مهم‌ترین مسئله‌ای که وی به آن می‌پردازد، تنوعاتی است که در نظام نوشتاری عربی معاصر دیده می‌شود و این مسئله مستلزم اِعمال تغییراتی در الگوریتم پردازشگر ساخت‌واژی عربی «باک‌والتر» است (Buckwalter 2004).

«ونزکی» بر این باور است که عوامل زبان‌شناختی، روان‌شناختی، و فرهنگی، همه نقش تعیین‌کننده‌ای در نظام نوشتاری دارند؛ به این صورت که هر چه یک نظام سعی در برقراری تطابق یک‌به‌یک بین نگاره و واج داشته باشد، بیشتر به گویشوران زبان‌های دیگر، که متن مربوط به آن زبان را می‌خوانند، در خواندن کمک می‌کند. در حالی که هر چه یک نظام نوشتاری بیشتر روی ریشه‌شناسی و ساخت‌واژه تکیه کند، متون آن تنها مورد استفادهٔ افرادی است که با آن نظام نوشتاری آشنا باشند. نویسنده در این مقاله بیشتر به خط رومی در زبان‌های هند و اروپایی می‌پردازد و عواملی را که باعث می‌شوند تناظر یک‌به‌یک میان واج و نگاره وجود نداشته باشد، ذکر می‌کند. در نهایت، نویسنده چارچوبی را برای مطالعهٔ موضوعات مربوط به نظام نوشتاری ارائه می‌دهد و نکته‌ای که به آن اشاره می‌کند این است که خوانندگان متون، آگاهی بالایی از فراوانی‌های نسبی وقوع حروف الفبا دارند. فرضیهٔ غالب این است که خوانندگان قادرند بدون آموزش دیدن، اطلاعات مربوط به فراوانی وقوع و ترتیب نشانه‌ها را استخراج کنند و میان آن‌ها تفاوت قائل شوند. علاوه بر این، ترکیبات با فراوانی وقوع بالا سریع‌تر تشخیص داده می‌شوند (Venezky 2004).

«محمدمهدی عرب» و «علی عظیمی‌زاده» دو روش کلی تبدیل حرف به آوا را معرفی می‌کنند: ۱- تبدیل حرف به آوا بر اساس استفاده از قواعد واج‌شناختی. به‌عنوان مثال، در سامانهٔ سنتز گفتار فستیوال[3] که توسط «بلک»[4] و همکارانش (۱۹۹۹) تهیه شده، از قواعد واج‌شناختی استفاده شده است. یک شکل ابتدایی از قواعد واج‌شناختی مورد استفاده در این سامانه به شکل زیر است:

(LEFTCONTEXT [ITEM] RIGHTCONTEXT NEWITEMS)

1. annotation
2. Buckwalter Arabic Morphological Analyzer
3. Festival speech synthesis
4. Black

این قاعده نشان می‌دهد که چنانچه ITEM در بافت مشخص راست و چپ ظاهر شود، رشتهٔ برون‌داد باید حاوی NEWITEMS باشد. بر اساس این قاعده می‌توان مثال (#[ch] C = k) را در نظر گرفت؛ # دلالت بر مرز کلمه دارد و نشانهٔ C مجموعهٔ همهٔ همخوان‌هاست. این قاعده نشان می‌دهد که یک< ch> در ابتدای کلمه که بعد از آن یک همخوان باشد، باید به‌صورت واج /k/ تلفظ شود. نوشتن چنین قواعدی که باید به‌صورت دستی وارد شوند، بسیار دشوار و وقت‌گیر است. ۲- تبدیل حرف به آوا از طریق استفاده از واژگان زبان. این روش بر اساس مدل محاسباتی تلفظ است که در آن از داده‌های آموزشی و روش آماری استفاده می‌شود. روش آماری در این رویکرد، دسته‌بندی و درخت رگرسیون[1] است. این روش آماری با داشتن مجموعه‌ای از ویژگی‌ها، قادر است داده‌ها را پیش‌بینی کند و این کار را از طریق استفاده از سؤالات بله/ خیر دربارهٔ ویژگی‌ها انجام می‌دهد. «عرب و عظیمی‌زاده» از روش دوم در تهیهٔ یک سامانهٔ تبدیل حرف به آوا در فارسی استفاده کرده‌اند. داده‌های آموزشی مورد استفاده در این سامانه، واژگانی به حجم ۳۲۰۰۰ کلمه همراه با تلفظ مربوط به هر کلمه است. این داده‌ها از پیکره‌های متنی مختلف تهیه شده است و بنابراین، توزیع بهتر آواها را در مقایسه با دیگر سامانه‌های تبدیل حرف به آوا در فارسی که در آن‌ها داده‌ها از منابع عمومی مانند روزنامهٔ همشهری تهیه شده بود، نشان می‌دهد. صحت این سامانه ۹۳/۶۱ درصد است که نشانگر توانایی بالا در پیش‌بینی تلفظ کلمات فارسی توسط این سامانه است. همین سامانه در زبان انگلیسی هم استفاده شده و صحت آن ۹۴/۶ درصد است (Arab & Azimizadeh 2009).

«مجید نم‌نبات» و «محمدمهدی همایون‌پور» (۱۳۸۶) ساختار سامانه‌های تبدیل حرف به گفتار با معماری سه لایه‌ای را معرفی کرده‌اند. آن‌ها لایهٔ اول این سامانه را لایهٔ قانون‌گرا نامیده‌اند. لایهٔ دوم، متشکل از پنج شبکهٔ عصبی پرسپترون چندلایه‌ای و یک بخش کنترل‌کننده برای تعیین دنبالهٔ واجی متناظر با حروف است. برای تعیین دنبالهٔ واجی متناظر با حروف، از شبکه‌های عصبی استفاده شده است. بخش کنترل‌کننده نیز بخش خروجی شبکه‌ها را کنترل می‌کند تا دنبالهٔ واجی نهایی متناظر با کلمات با ساختار هجابندی فارسی مطابقت داشته باشد. در لایهٔ سوم نیز یک شبکهٔ عصبی

1. Classification And Regression Tree (CART)
زمانی که تابع هدف به‌صورت پیوسته باشد مسئلهٔ یادگیری، یک مسئلهٔ رگرسیون است مانند یادگیری رابطهٔ قیمت و مساحت خانه. وقتی ویژگی‌های برون‌داد طوری باشد که بتواند تعداد محدودی مقدار گسسته بگیرد، مسئلهٔ یادگیری یک مسئلهٔ دسته‌بندی خواهد بود؛ مانند اینکه در خرید خانه این سؤال مطرح است که آیا خانهٔ مورد نظر یک آپارتمان است؟

برای تعیین حروف مشدد، با استفاده از نتایج مراحل قبل وجود دارد. اجزاء مختلف این سامانه به‌گونه‌ای طراحی شده‌اند که در نهایت، برای هر کلمه، یک دنبالهٔ واجی منطقی تولید شود. منظور از دنبالهٔ واجی منطقی، دنبالهٔ واجی است که در آن اصول بدیهی واج‌نگاری و ساختار هجابندی زبان فارسی رعایت شده باشد. در تحقیق آن‌ها، میزان درستی به‌دست آمده برای حروف ۸۸ درصد و برای کلمات ۶۱ درصد است.

۳. روش تحقیق و تجزیه و تحلیل داده‌ها

در این بخش، دو فهرست از کلمات خارج از واژگان در اختیار فارسی‌زبانان قرار گرفت: فهرست اول شامل کلمات خارج از واژگان زبان فارسی است که بر اساس الگوهای ساخت‌واژی عربی ساخته شده است و فهرست دوم شامل کلماتی است که خود به دو دسته تقسیم می‌شود: کلماتی که همه یا بخشی از تک‌واژهایش در واژگان موجود است، اما خود کلمات خارج از واژگان هستند، و کلماتی که هیچ بخشی از آن در واژگان موجود نیست. این دو فهرست برای تلفظ در اختیار ۳۰ فارسی‌زبان با سطح تحصیلات دانشگاهی قرار گرفت تا بتوان نتایج حاصل را معیاری برای سنجش برون‌داد مدل پیشنهادی طرح قرار داد. فهرست اول شامل کلمات: <شامع>، <مقذول>، <تجسیل>، <شسیم>، <اسکام>، <تشارگ>، <لمول>، <مخاگر>، <مزادیک>، <تپدق>، <انتپاک>، <اکتعام>، <استشدان>، <مواعمه>، <متمقل>، <صاحد>، <مگشون>، <تمچین>، <پچین>، <اکوال>، <تباعم>، <شسول>، <مپادم>، <مکاعیش>، <تکشن>، <انتکال>، <ابتشاج>، <استپچان>، <ملاتبه>، <متزقل>، <ثاکل>، <مصبوق>، <تشچیف>، <دزین>، <اکداز>، <تپانق>، <کدون>، <مباظن>، <مچافیس>، <تدمع>، <اندشال>، <اجتقاذ>، <استنگال>، <مگاشده>، <متعدم>، <مانن>، <میجوف>، <تلنیم>، <لجیک>، <انقاو>، <تذاوج>، <تشون>، <مفاشغ>، <مثالیو>، <تعرگ>، <انزگار>، <التکان>، <استمکال>، <مپاینه>، <متشغم>، <عارگ>، <مدکول>، <تصکیل>، <شعیف>، <ابساز>، <تصابف>، <تدوس>، <مناعظ>، < معاتین>، <تسعل>، <انمعال>، <اعتفاو>، <استجمار>، <معابقه>، <متلرگ> و فهرست دوم شامل کلمات: <قواعدگانه>، <مرکب‌پژوه>، <خال‌و-چال>، <مصنوع‌گاه>، <بی‌پیرایشی>، <مادون‌خواه>، <شناسدان>، <افقیون>، <نانویس>، <اینترنت‌خوانی>، <موبایلی>، <گیرخانه>، <فنریات>، <گیاه‌مغزی>، <تلفنیدن>، <می-اذانیدم>، <نمی‌کاغذاند>، <ناخن‌باز>، <رفتگی>، <عمویان>، <کاهوان>، <ایمان‌گونگی>، <لایکید>، <چلم‌خوان>، <پیرایگمن>، <غذاشیم>، <مانمیس>، <شاسکلم>، <لوبکان>، <پست‌لقه>، <کچال‌کاه>، <عمیشگی>، <منلکوب>، <بیاتوم>، <تأسیم>، <میاهوان>،

<ادماکن>،<چمیان>، <آلویش>، <نماپش>، <آگوشانه>، <خیامت>، <لیوگی>، <اصلیشا>، <پشسیمان>، <ملامتظور>، <بکادوش>، <می‌صلید>، <اشتلقوس>، <اخطشم>، <ممچوپک>، <استرامذ>، <نفالود>، <حتمک>، <ماژلو>، <خسجل>، <کیامت>، و <کلاکوم> است.

بخشی از فهرست اول و برون‌داد واجی ارائه‌شده توسط آزمودنی‌ها در جدول ۱ آمده است.

جدول ۱. بررسی آماری میزان استفادهٔ فارسی‌زبانان از اطلاعات زبانی مربوط به الگوهای ساخت‌واژی عربی در زبان فارسی

درصد برون‌دادهای واجی مغایر با برون‌داد واجی الگوهای ساخت‌واژی عربی	برون‌دادهای واجی غالب مغایر با برون‌داد واجی الگوهای ساخت‌واژی عربی	درصد برون‌دادهای واجی مطابق با برون‌داد واجی الگوهای ساخت‌واژی عربی	برون‌داد واجی کلمات خارج از واژگان زبان فارسی منطبق با برون‌داد واجی الگوهای ساخت‌واژی عربی	الگوهای ساخت‌واژی عربی قرضی	کلمات
۰		۱۰۰	/ʃɒmeʔ/	فاعل	۱- شامع
۰		۱۰۰	/maGzul/	مفعول	۲- مقذول
۳/۳۳	/tedʒsil/	۹۶/۶۷	/tadʒsil/	تفعیل	۳- تجسیل
۱۶/۶۷	/ʃesim/	۸۳/۳۳	/ʃasim/	فعیل	۴- شسیم
۱۶/۶۷	/ʔoskɒm/	۸۳/۳۳	/ʔaskɒm/ /ʔeskɒm/	افعال	۵- اسکام
۷۳/۳۳	/taʃɒrag/	۲۶/۶۷	/taʃɒrog/	تفاعل	۶- تشارگ
۸۶/۶۷	/lamul/	۱۳/۳۳	/lomul/	فعول	۷- لمول
۰		۱۰۰	/ʔebteʃɒdʒ/	افتعال	۲۷- ابتشاج
۲۰	/motzaGol/ /matzaGal/	۸۰	/motazaGGel/ /motezaGGel/	متفعل	۳۰- متزقل
۹۳/۳۳	/maxɒgar/	۶/۶۷	/maxɒger/	مفاعل	۸- مخاگر
۲/۳۳	/mezɒdik/	۷۶/۶۷	/mazɒdik/	مفاعیل	۹- مزادیک
۹۰	/tapdaG/	۱۰	/tapaddoG/	تفعل	۱۰- تپدق
۶/۶۷	/ʔestaptʃɒn/	۹۳/۳۳	/ʔesteptʃɒn/	استفعال	۲۸- استپچان

بررسی آماری تک‌تک کلمات نشان می‌دهد که فارسی‌زبانان در برخورد با کلمات خارج از واژگانی که الگوی نوشتاری‌شان با الگوهای ساخت‌واژی عربی <فاعل>، <مفعول>، <تفعیل>،

<افعال>، <مفاعیل>، <افتعال>، <استفعال> و <متفعل> مطابقت دارد، تلفظی که از این کلمات ارائه می‌دهند، نزدیک به ۷۰ درصد یا حتی بیشتر، مطابق با ساخت واجی الگوهای ساخت‌واژی عربی است (بی‌جن‌خان و علایی ۱۳۹۲). بنابراین، در مدل پیشنهادی تحقیق می‌توان جایگاهی برای بررسی مطابقت صورت نوشتاری کلمهٔ خارج از واژگان با صورت نوشتاری الگوهای ساخت‌واژی غالب واردشده در فارسی در نظر گرفت.

بخشی از فهرست دوم و برون‌داد واجی ارائه‌شده توسط همان آزمودنی‌ها در جدول ۲ آمده است:

جدول ۲. بررسی آماری میزان استفادهٔ فارسی‌زبانان از اطلاعات ساخت‌واژی فارسی

برون‌دادهای واجی غالب مغایر با برون‌داد واجی مربوط به استفاده از اطلاعات زبانی	درصد برون‌دادهای واجی مغایر با برون‌داد واجی مربوط به استفاده از اطلاعات زبانی	درصد برون‌دادهای واجی مطابق با برون‌داد واجی مربوط به استفاده از اطلاعات زبانی	برون‌داد واجی بر اساس استفاده از اطلاعات زبانی مربوط به تک‌واژها	کلمات
	۰ درصد	۱۰۰ درصد	/Gavɒʔed gɒne/	قواعدگانه
	۰ درصد	۱۰۰ درصد	/morakkab paʒuh/, /markab paʒuh/	مرکب‌پژوه
	۰ درصد	۱۰۰ درصد	/xɒl o ʧɒl/, / xɒl va ʧɒl/	خال‌وچال
	۰ درصد	۱۰۰ درصد	/ʃenɒs dɒn/	شناسدان
/ʔafGijun/	۶۱/۵۴ درصد	۳۸/۴۶ درصد	/ʔofoGijun/	افقیون
/nɒnvis/	۷/۷۰ درصد	۹۲/۳۰ درصد	/nɒ nevis/	نانویس
	۰ درصد	۹۲/.۳/mɒnmis/ درصد /mɒnemis/ ۷/۷۰ درصد	/mɒn{a,e,o,Φ }mis/ /mɒn{a,e,o}m{a,e,o}j{a,e, o,Φ }s/	مانمیس
/lubkɒn/	٪۵۳/۸۴	/lobkɒn/ ۳۰/۷۸ درصد /lobakɒn/ ۱۵/۳۸	/lob{a,e,o,Φ }kɒn/	لوبکان

بررسی آماری برون‌دادهای ارائه‌شده توسط فارسی‌زبانان در این بخش نشان می‌دهد که همان‌طور که انتظار می‌رفت، فارسی‌زبانان از اطلاعات زبانی خود برای ارائهٔ تلفظ این کلمات استفاده کرده‌اند؛ زیرا در اکثر موارد درصد برون‌دادهای واجی مطابق استفاده از اطلاعات زبانی بالاتر از درصد برون‌دادهای واجی مغایر است.

٤. مدل ساختاری- احتمالاتی برای برون‌داد واجی کلمات خارج از واژگان

پیکرهٔ مورد استفاده در تحقیق حاضر، واژگان زایا است، اما مدل ساختاری- احتمالاتی پیشنهادی برای برون‌داد واجی کلمات خارج از واژگان، مدل کلی است که می‌تواند در مورد هر پیکرهٔ فارسی که حاوی تک‌واژهای فارسی است، استفاده شود. قبل از معرفی مدل ساختاری- احتمالاتی برای برون‌داد واجی کلمات خارج از واژگان لازم است انواع کلمات خارج از واژگان بررسی شوند. برای کلمات خارج از واژگان ابتدا دو حالت را می‌توان در نظر گرفت: کلمه یا بسیط است یا غیربسیط که متشکل از بیش از یک تک‌واژ است. منظور از کلمهٔ بسیط در این بخش کلمه‌ای است که با مطابقت‌دادن با تک‌واژهای درون واژگان، هیچ تک‌واژی در آن شناسایی نشود. اگر کلمه بسیط باشد، یا الگوی نوشتاری آن با الگوهای ساخت‌واژی عربی مطابقت می‌کند یا مطابقت نمی‌کند. و اگر کلمه غیربسیط باشد، دو حالت را می‌توان برای آن در نظر گرفت که به شرح زیر است:

۱. همهٔ تک‌واژهای کلمه در واژگان موجود باشد، ولی خود کلمه در واژگان نباشد، مانند کلمهٔ <اینترنت‌خوانی>*؛ تک‌واژهای <اینترنت> با تلفظ /ʔinternet/، <خوان> با تلفظ / xɒn / و <ی> با تلفظ /i/، در واژگان موجود هستند، اما کلمهٔ <اینترنت‌خوانی> در واژگان موجود نیست.
۲. اگر فرض کنیم که کلمهٔ خارج از واژگان، غیربسیط است، این کلمه nتا تک‌واژ دارد که حداقل یکی از تک‌واژهایش در واژگان موجود است و حداکثر n–۱ تک‌واژ آن در واژگان موجود نیست. به‌عنوان مثال، کلمهٔ <پانبین> *: بخش <پان> * در واژگان موجود نیست اما تک‌واژ <بین> با تلفظ /bin/ در واژگان موجود است.

برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان، در هر یک از حالت‌های مذکور، به روش زیر عمل می‌کنیم:

در حالت ۱ که همهٔ تک‌واژهای کلمهٔ غیربسیط در واژگان موجود است، در مدل مربوطه، کلمه وارد بخش قواعد واژواجی می‌شود. در این بخش ابتدا نوع کلمه مشخص می‌گردد. به این صورت که کلمهٔ مورد نظر یا مجموعه‌ای است از ستاک+وند (اشتقاقی یا تصریفی) مانند کلمهٔ <سایگی>*، یا یک کلمهٔ مرکب است، مانند کلمهٔ <سرکفش>* و یا کلمهٔ مشتق مرکب مانند کلمهٔ <رایانه‌خوانی>* است. سپس قواعد واژواجی، تغییرات واجی را که باید در مرز تک‌واژها یا درون تک‌واژها هنگام اضافه‌شدن پیشوندها و پسوندها به ستاک صورت گیرند، مشخص می‌نمایند. بنابراین، با دردست‌داشتن برون‌داد واجی تک‌واژهای شناخته‌شده (که همراه تک‌واژها در واژگان موجود است) و قواعد واژواجی، می‌توان برون‌داد واجی چنین کلمات خارج از

واژگان را که تک‌واژهای آن در واژگان موجود است، تعیین کرد. به‌عنوان مثال، برای ارائهٔ برون‌داد واجی کلمهٔ <سایگی>* روند زیر دنبال می‌شود:

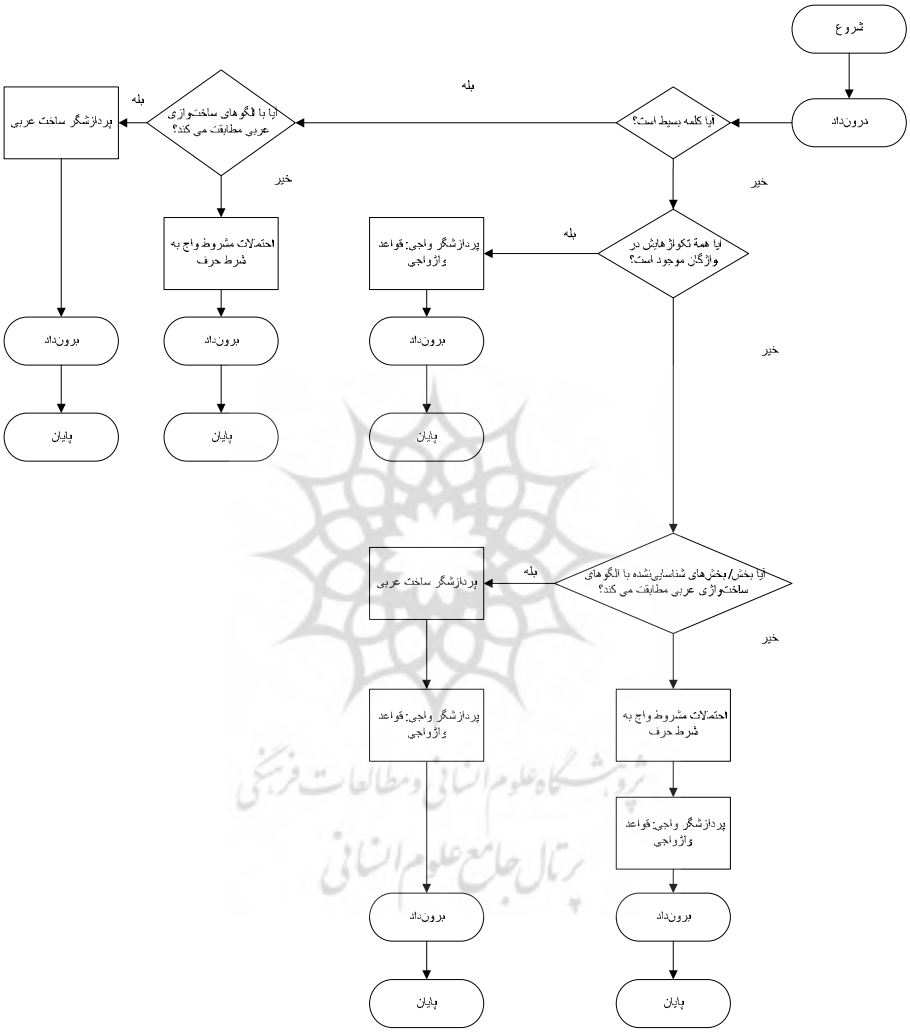
<سایگی> ⟵ پردازش واژواج ⟵ { <سایگی> → <سایه> + <ی> ; <های غیرملفوظ> → Ø / <ی> + morph [سایـ-] ; Ø → <گ> / <ی> + morph [سایـ] ; Ø → [ʃ] / sɒje]morph + /i/ } ⟵ /sɒjeʃi/

با توجه به قاعدهٔ ساخت‌واژی و واژواجی فوق می‌توان گفت، با دردست‌داشتن برون‌دادهای واجی ستاک و وند در واژگان و قاعدهٔ واژواجی، برون‌داد واجی کلمهٔ خارج از واژگان <سایگی>* به‌صورت /sɒjeʃ/ / می‌شود.

در حالت ۲ که قسمتی از کلمهٔ غیربسیط، تک‌واژ/ تک‌واژهای موجود در واژگان است و قسمت دیگر در واژگان نیست، برون‌داد واجی تک‌واژ/ تک‌واژهای شناسایی‌شده در دسترس است و قسمت خارج از واژگان ابتدا با الگوهای ساخت‌واژی عربی مطابقت داده می‌شود و چنانچه صورت نوشتاری آن بر اساس الگوهای نوشتاری بنیان‌های عربی قرضی در زبان فارسی باشد، برون‌داد واجی آن مطابق الگوی ساخت‌واژی مد نظر می‌شود. به‌عنوان مثال، بخش <مشنوس>* از کلمهٔ خارج از واژگان <مشنوسات>*، صورت نوشتاری مطابق الگوی ساخت‌واژی عربی <مفعول> دارد. بنابراین، برون‌داد واجی این بخش به‌صورت /maʃnus/ می‌شود و با دردست‌داشتن برون‌داد واجی بخش دوم (تک‌واژ تصریفی جمع‌ساز<ات> /ɒt/) در واژگان، برون‌داد واجی کل کلمه به‌صورت/ maʃnusɒt/ می‌شود. اگر صورت نوشتاری بخش خارج از واژگان با الگوهای ساخت‌واژی عربی مطابقت نداشته باشد، بخش مورد نظر وارد بخش احتمالات می‌شود که در این بخش با استفاده از توزیع احتمال واج به شرط حرف/ حروف، برون‌داد واجی بخش مورد نظر ارائه می‌شود. در واقع، در بخش احتمالات، احتمال حضور واکه‌های کوتاه با توجه به بافت نوشتاری محاسبه می‌شود.

دربارهٔ کلمات بسیط که هیچ تک‌واژی در آن‌ها شناسایی نمی‌شود، کل کلمه ابتدا با الگوهای ساخت‌واژی عربی مطابقت داده می‌شود و چنانچه صورت نوشتاری آن بر اساس الگوهای نوشتاری بنیان‌های عربیِ واردشده در زبان فارسی باشد، برون‌داد واجی آن مطابق الگوی ساخت‌واژی مد نظر می‌شود. در غیر این صورت، کلمهٔ مورد نظر وارد بخش احتمالات می‌شود و با استفاده از توزیع احتمال واج به شرط حرف/ حروف، برون‌داد واجی آن ارائه می‌گردد. با در

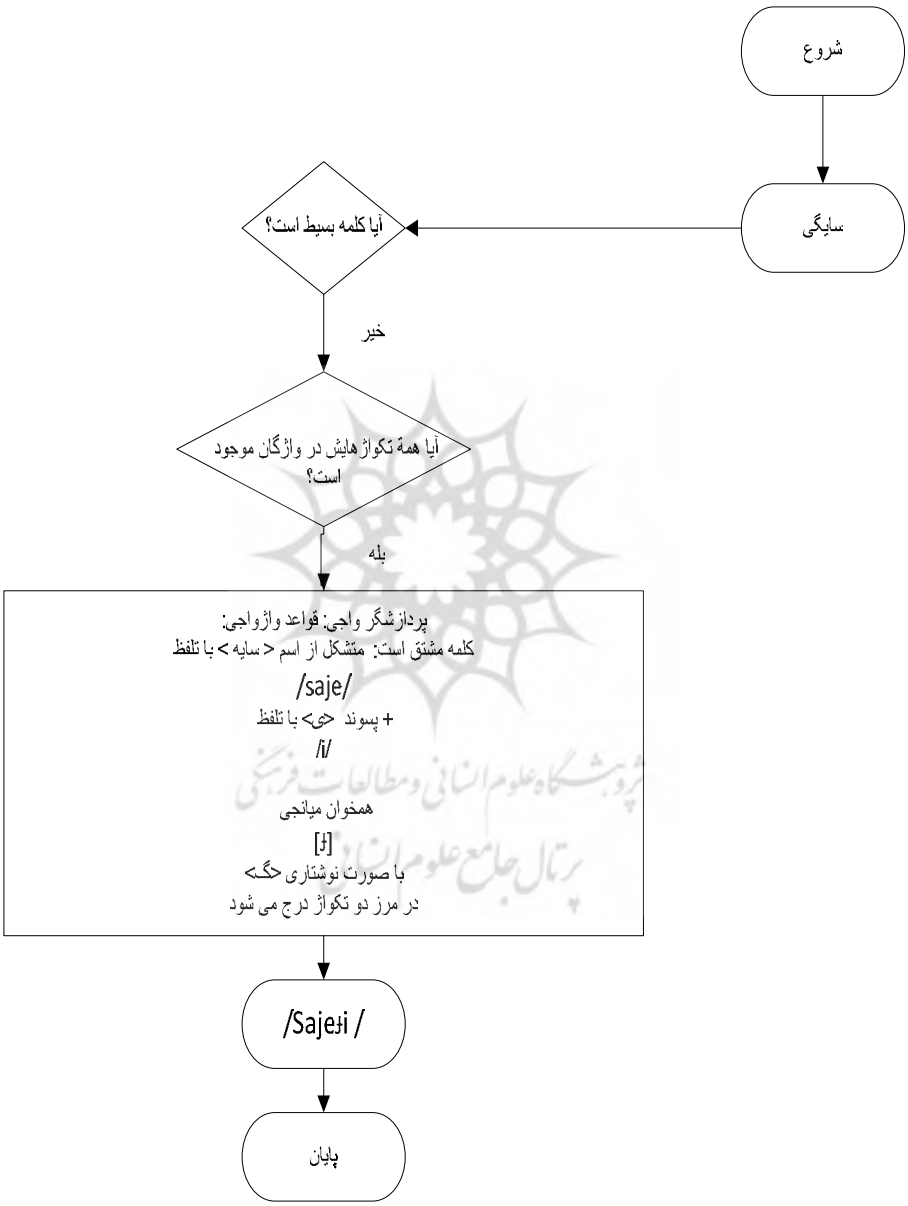
نظرگرفتن حالت‌های ذکرشده، می‌توان مدل ساختاری- احتمالاتی زیر را برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان ارائه داد.



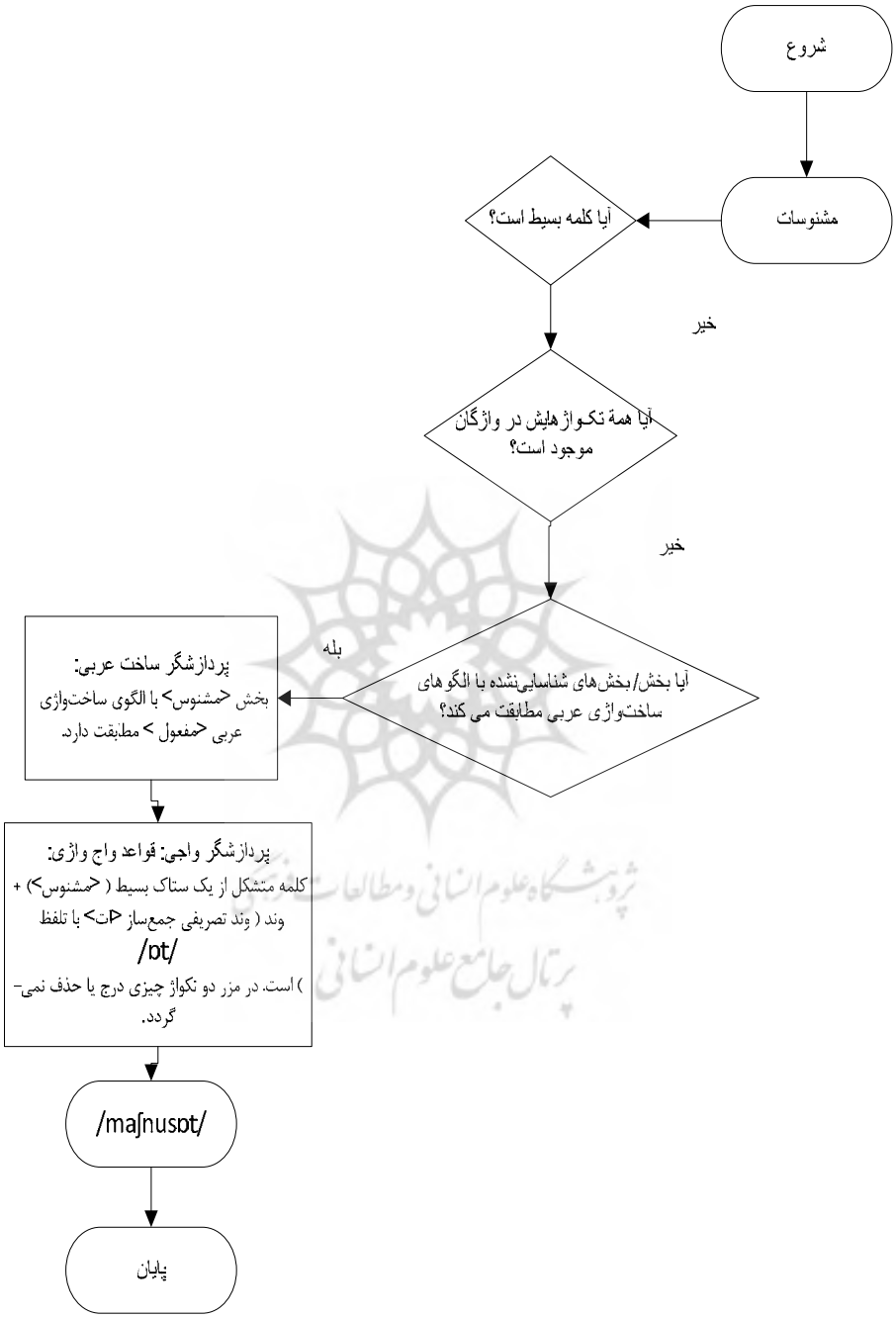
شکل ۱. مدل ساختاری-احتمالاتی برای برون‌داد واجی کلمات خارج از واژگان

به‌عنوان مثال، طبق شکل ۱، ارائهٔ برون‌داد واجی کلمات خارج از واژگان <سایگی>* (همهٔ تک‌واژهای موجود در این کلمه در واژگان موجود است)، <مشنوسات>* (یکی از تک‌واژهای این کلمه (<-ات>) در واژگان موجود است و بخش دیگر (<مشنوس>) با الگوی ساخت‌واژی

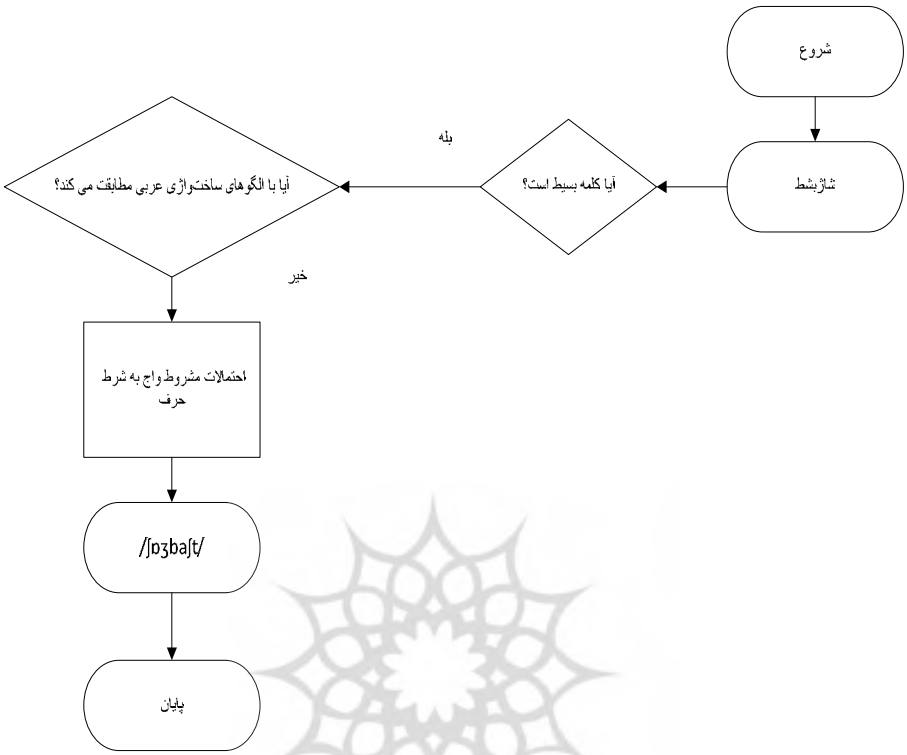
عربی<مفعول> مطابقت دارد، <شاژبشط>* و <بزاغیل>* (که کلمات بسیط هستند)، به‌ترتیب، به‌صورت زیر است:



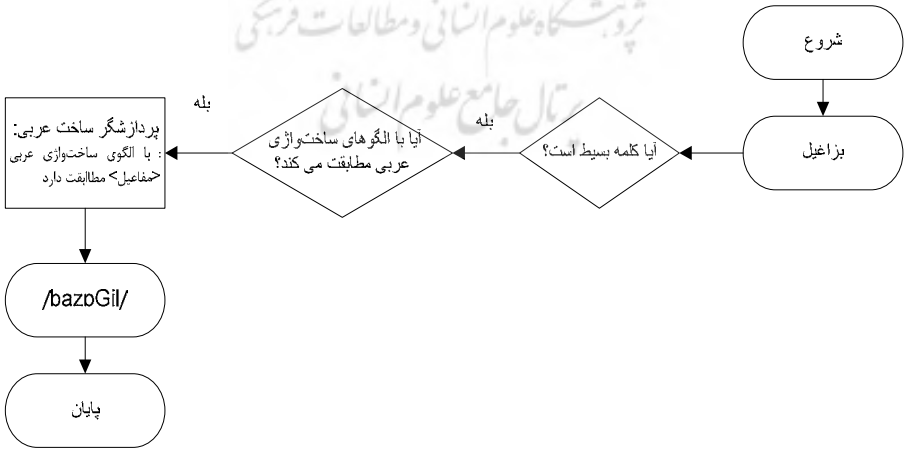
شکل ۲. ارائهٔ برون‌داد واجی کلمهٔ <سایگی> * بر اساس مدل ساختاری- احتمالاتی ارائه شده در شکل -۱



شکل ۳. ارائهٔ برون‌داد واجی کلمهٔ <مشنوسات> * بر اساس مدل ساختاری- احتمالاتی ارائه‌شده در شکل ۱



شکل ٤. ارائهٔ برون‌داد واجی کلمهٔ <شاژبشط> * بر اساس مدل ساختاری- احتمالاتی ارائه‌شده در شکل ۱



شکل ٥. ارائهٔ برون‌داد واجی کلمهٔ <بزاغیل> * بر اساس مدل ساختاری- احتمالاتی ارائه‌شده در شکل ۱

۵. مقایسهٔ عملکرد یکی از نرم‌افزارهای تبدیل حرف به واج فارسی با عملکرد فارسی‌زبانان و مدل ساختاری- احتمالاتی تحقیق

همان‌گونه که ذکر شد، به‌منظور تهیهٔ برنامهٔ رایانه‌ای تبدیل حرف به واج فارسی از *واژگان زایا* استفاده شد. *واژگان زایای* زبان فارسی حدود ۵۵ هزار مدخل واژگانی دارد و این تعداد واحد واژگانی در چارچوب قواعد تصریف کلمه می‌توانند صورت‌های تصریفی متفاوتی داشته باشند. برای عملیاتی کردن *واژگان زایا* برنامهٔ رایانه‌ای تهیه شده است که این برنامه رایانه‌ای با ارجاع به واژگان و نیز قواعد تصریف کلمه در زبان فارسی می‌تواند واحدهای زبانی نوشتار یا گفتار را به لحاظ صرفی پردازش کند و خوانش صحیح صرفی به زنجیرهٔ ورودی برنامه اختصاص دهد. *واژگان زایا* حاوی چند نوع اطلاع زبانی و غیرزبانی برای هر مدخل است؛ مانند: صورت املایی، صورت واجی، مقولهٔ واژگانی، الگوی تکیه، بسامد مدخل در یک پیکرهٔ زبانی و غیره. البته طراحی واژگان به‌گونه‌ای صورت گرفته است که امکان تغییر و یا افزایش اطلاعات دیگر نیز وجود دارد (اسلامی و همکاران ۱۳۸۳). در تحقیق حاضر از دو نوع اطلاعات موجود در *واژگان زایا* استفاده شد: صورت نوشتاری و صورت واجی کلمات. بنابراین، در مرحلهٔ آموزش نرم‌افزار نگارنده دو ستون (ستون مربوط به‌صورت نوشتاری کلمات و ستون مربوط به‌صورت واجی کلمات) را از *واژگان زایا* جدا کرده و بین دو ستون با کلید tab فاصله انداخت. سپس، با استفاده از نرم‌افزار ترازبندی واجی[1] که برای تعیین برون‌داد واجی هر حرف موجود در کلمات در آزمایشگاه پردازش هوشمند داده‌های چندرسانه‌ای دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر توسعه یافته است، کلمات در ستون صورت نوشتاری به‌صورت ستونی به حروف تشکیل‌دهنده و در ستون صورت واجی به‌صورت ستونی به واج‌های تشکیل‌دهندهٔ متناظر با حروف تقطیع شدند. با توجه به پیچیدگی ترازبندی زبان فارسی خصوصاً به‌دلیل حضور گستردهٔ کلمات عربی در این زبان، اقدامات متعددی در راستای بهبود این ترازبندی در نرم‌افزار آزمایشگاه مذکور انجام گردید. خروجی این نرم‌افزار طی چند مرحله بررسی و مشکلات ترازبندی شناسایی و برطرف گردید. فایل خروجی نهایی حاصل از ترازبندی، کنترل دستی شد و خطاهای موجود رفع گردید. پس از رفع خطاهای ترازبندی فایل داده‌های ترازبندی‌شده برای آموزش روش‌های یادگیری ماشینی شبکهٔ عصبی پرسپترون و درخت ID3 مورد استفاده قرار گرفت تا این روش‌ها پس طی کردن فاز آموزش برای تبدیل کلمات به برون‌داد واجی متناظر آن‌ها استفاده شوند.

اگر عملکرد فارسی‌زبانان در ارائهٔ برون‌داد واجی کلمات خارج از واژگان، ملاک ارزیابی

1. Phoneme alignment

عملکرد مدل تحقیق و نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون قرار گیرد، می‌توان قسمتی از عملکرد این نرم‌افزارها را با عملکرد فارسی‌زبانان در ارائهٔ برون‌داد واجی کلمات خارج از واژگان در جدول ۳ و ۴ مقایسه کرد. در این جدول‌ها برون‌دادهای مطابق برون‌دادهای ارائه شده توسط فارسی‌زبانان، با علامت (✓) و برون‌دادهای مغایر با علامت (×) مشخص شده است. لازم به ذکر است که در این قسمت، برون‌داد ارائه‌شدهٔ صحیح توسط فارسی‌زبانان، ملاک ارزیابی و مقایسه با برون‌دادهای ارائه‌شده توسط نرم‌افزارها نبود، بلکه به‌طور کلی، هر نوع تلفظی که فارسی‌زبانان برای کلمات خارج از *واژگان زایا* ارائه داده بودند، به‌عنوان تلفظِ پذیرفته‌شده برای آن کلمات در نظر گرفته شد تا ملاک ارزیابی صرفاً برون‌دادهای ارائه‌شده توسط فارسی‌زبانان باشد.

جدول ۳. مقایسهٔ برون‌دادهای واجی ارائه‌شده توسط نرم‌افزار ID3 با برون‌دادهای ارائه‌شده توسط فارسی‌زبانان

برون‌داد واجی مطابق یا مغایر با برون‌دادهای واجی فارسی‌زبانان	برون‌دادهای واجی ارائه‌شده توسط نرم‌افزار	کلمات	برون‌داد واجی مطابق یا مغایر با برون‌دادهای واجی فارسی‌زبانان	برون‌دادهای واجی ارائه‌شده توسط نرم‌افزار	کلمات
×	/maGzwl/	مقذول	✓	/taʤsil/	تجسیل
×	/taʃɒrg/	تشارگ	×	/ʃesim/	شسیم
×	/lemul/	لمول	×	/ʔeskeɒm/	اسکام
✓	/mɒnmis/	مانمیس	×	/Gazɒʃijom/	غذاشیم
✓	/ʔɒlviʃ/	آلویش	✓	/ʧamjɒn/	چمیان
✓	/ʔɒguʃɒne/	آگوشانه	×	/namɒpoʃ/	نماپش
×	/motalerg/	متلرگ	×	/moʔʔɒbeGe/	معابقه
×	/maɬaʃun/	مگشون	✓	/tamʧin/	تمچین
×	/masnuʔe gɒh/	مصنوع‌گاه	✓	/xɒl va ʧɒl/	خال‌وچال
×	/keʧɒlkeɒh/	کچال‌کاه	✓	/postalGe/	پست‌لقه

جدول ٤. مقایسهٔ برون‌دادهای واجی ارائه‌شده توسط نرم‌افزار شبکهٔ عصبی پرسپترون با برون‌دادهای ارائه‌شده توسط فارسی‌زبانان

برون‌داد واجی مطابق یا مغایر با برون‌دادهای واجی فارسی‌زبانان	برون‌داد واجی	کلمات	برون‌داد واجی مطابق یا مغایر با برون‌دادهای واجی فارسی‌زبانان	برون‌داد واجی	کلمات
×	/moGzul/	مقذول	✓	/taʤsil/	تجسیل
×	/mopoʤuf/	مپجوف	×	/tlnim/	تلنیم
×	/!ijɒmat/	کیامت	✓	/xijɒmat/	خیامت
×	/ʔaintrnt xvɒni/	اینترنت‌خوانی	×	/ʔamvijɒn/	عمویان
×	/motazGal/	متزقل	✓	/mosbuG/	مصبوق
×	/rftaʒi/	رفتگی	×	/nɒxnbɒz/	ناخن‌باز
×	/mamtʃupe!/	ممچوپک	×	/ʔextʃm/	اخطشم
×	/nafɒleud/	نفالود	✓	/ʔesterɒmez/	استرامذ
×	/ʔe!dɒz/	اکداز	×	/tapɒnG/	تپانق
×	/mnl!ub/	منلکوب	✓	/ʔamiʃʃi/	عمیشگی

به‌طور کلی می‌توان گفت که به نظر می‌رسد مدل ساختاری- احتمالاتی تحقیق حاضر در ارائهٔ برون‌داد واجی کلمات خارج از واژگان، در مقایسه با عملکرد نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون بهتر عمل کند و برون‌دادهای واجی آن به تلفظ فارسی‌زبانان نزدیک‌تر باشد؛ زیرا نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون به‌صورت احتمالاتی، احتمالات مشروط واج به‌شرط حرف را از پیکره محاسبه کرده و در ارائهٔ برون‌داد واجی کلمات خارج از واژگان تنها از احتمالات استفاده کرده‌اند. این مسئله قابل قیاس با این حالت است که فردی غیرفارسی‌زبان تنها از طریق فهرستی از کلمات فارسی که تلفظ آن‌ها در کنار آن کلمات موجود باشد، به‌طور احتمالاتی بیاموزد که چند درصد بعد از همخوان‌ها، هر یک از واکه‌های کوتاه /a/، /e/ و /o/ ظاهر شده است و یا چند درصد حرف <و> به‌صورت /v/،/o/ و /u/ تلفظ شده است. بنابراین، در این حالت از هیچ‌کدام از اطلاعات زبانی که فارسی‌زبانان در خواندن کلمات استفاده می‌کنند، برخوردار نیست. برون‌دادهای واجی نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون به‌دلیل عدم دسترسی نرم‌افزارها به اطلاعات زبانی (مانند: اطلاعات مربوط به ساخت‌واژهٔ فارسی و الگوهای ساخت‌واژی عربی) و اکتفاکردن به احتمالات، در اکثر موارد بسیار دورتر از برون‌دادهای واجی فارسی‌زبانان است.

٦. نتیجه‌گیری

در تحقیق حاضر این سؤال مطرح شد که با توجه به اینکه در خط فارسی واکه‌های کوتاه معمولاً نمایش داده نمی‌شوند و رابطهٔ حروف و واج‌ها به‌صورت چند به یک و یک به چند می‌تواند باشد (رابطهٔ یک به چند مانند: حرف <و> که با واج‌های /v/ ,/u/ ,/o/ مرتبط است و رابطهٔ چند به یک مانند: حروف <ط> و <ت> که با واج /t/ ارتباط دارند)، می‌توان گفت در خط فارسی نگاشت یک به یک حرف به واج همواره برقرار نیست. بنابراین، علی‌رغم وجود چنین ویژگی‌هایی در خط فارسی، فارسی‌زبانان هنگام خواندن کلمات فارسی موجود در واژگان ذهنی خود و کلماتی که برای اولین بار با آن‌ها در متون گوناگون مواجه می‌شوند، چگونه رشتهٔ حروف را تبدیل به واج می‌کنند؟ فارسی‌زبانان با استفاده از اطلاعات زبانی مانند اطلاعات مربوط به تکواژها، آشنایی با صورت نوشتاری وندهای تصریفی و اشتقاقی، دانستن قواعد حرف‌نویسی مربوط به اضافه‌شدن وندها به ستاک که در مرز تک‌واژها عمل می‌کنند و آشنایی با صورت نوشتاری و تلفظ کلماتی که منشأ عربی دارند، سعی در کم‌کردن عمق خط فارسی دارند تا به این ترتیب بتوانند کلماتی را که اولین بار با آن‌ها مواجه می‌شوند، به‌درستی بخوانند. تحقیق حاضر مدلی ساختاری- احتمالاتی برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان معرفی می‌کند که به‌کمک آن بتوان کلمات خارج از واژگان را تلفظ کرد. عملکرد نرم‌افزارهای تبدیل حرف به واج فارسی با عملکرد فارسی‌زبانان و مدل ساختاری- احتمالاتی پژوهش مقایسه شد و به‌طور کلی، این نتیجه به‌دست آمد که به نظر می‌رسد عملکرد مدل ساختاری-احتمالاتی تحقیق حاضر برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان، در مقایسه با نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون بهتر و برون‌دادهای واجی به تلفظ فارسی‌زبانان نزدیک‌تر باشد، زیرا نرم‌افزارهای ID3 و شبکهٔ عصبی پرسپترون به‌صورت احتمالاتی، احتمالات مشروط واج به‌شرط حرف را از پیکره محاسبه کرده‌اند و در ارائهٔ برون‌داد واجی کلمات خارج از واژگان تنها از احتمالات استفاده کرده‌اند، حال آنکه مدل ساختاری- احتمالاتی پیشنهادی، برای ارائهٔ برون‌داد واجی کلمات خارج از واژگان، مجهز به اطلاعات ساختاری زبان فارسی و اطلاعات مربوط به الگوهای ساخت‌واژی عربی است و بنابراین، عملکرد آن در مقایسه با عملکرد نرم‌افزارهایی که تنها به قوانین احتمالات در ارائهٔ برون‌داد واجی کلمات خارج از واژگان اکتفا می‌کنند، می‌تواند بهتر باشد.

فهرست منابع

اسلامی، محرم، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری، و طاهره زندی ۱۳۸۳. *واژگان زایای زبان فارسی*. اولین کارگاه پژوهشی زبان فارسی و رایانه. دانشکدهٔ ادبیات و علوم انسانی دانشگاه تهران.

بی‌جن‌خان، محمود و الهام علایی. ۱۳۹۲. بررسی الگوهای ساخت‌واژی عربی واردشده در زبان فارسی. *پژوهش‌های زبان‌شناسی تطبیقی*. ۳ (۵): ۱-۲۲.

علایی، الهام و محمود بی‌جن‌خان. ۱۳۹۲. عمق خط فارسی. *پژوهش‌های زبانی (مجله سابق دانشکدهٔ ادبیات و علوم انسانی دانشگاه تهران*. ۴ (۱): ۱-۲۰.

نم‌نبات، مجید و محمدمهدی همایون‌پور. ۱۳۸۶. تبدیل حرف به صدا در زبان فارسی به‌کمک شبکه‌های عصبی پرسپترون چندلایه‌ای. *نشریهٔ مهندسی برق و مهندسی کامپیوتر ایران*، ۵ (۳): ۱۴۷-۱۵۴.

Arab. M., and A. Azimizadeh. 2009. *Construction of a Persian letter-to-sound conversion system based on classification and regression tree in Festival*. Proceedings of the workshop on computational approach for Arabic script based languages =CAASL-3. Stanford university.

Buckwalter, T. 2004. *Issues in Arabic orthography and morphology analysis*. Proceedings of the workshop on computational approaches to Arabic script-based languages in conjunction with COLING. Switzerland.

Kessler, B., and R. Treiman. 1997. Syllable structure and the distribution of phonemes in English syllables. *Journal of memory and language. Academic press* 37: 295-311.

Megerdoomian, K. 2004. *Finite state morphological analysis of Persian.* Workshop on computational approaches to Arabic script-based languages. Switzerland.

Trammel, R. L. 1990. Variant grapheme-phoneme correspondences in unfamiliar polysyllabic words. *Language and speech* 33 (4): 293-323.

Van den Bosch, A., A. Content, W. Daelemans, and B. De Gelder, 1994. *Analyzing orthographic depth of different languages using data-oriented algorithms.* 2nd international conference on quantitative linguistics, Moscow.

Venezky, R. L. 2004. *In search of the perfect orthography. Written language and literacy*: 139-163. Amsterdam: John Benjamins publishing company.

الهام علایی ابوذر

متولد سال ۱۳۵۹، دارای مدرک تحصیلی دکتری تخصصی در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون استادیار پژوهشکدهٔ مدیریت دانش پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

زبان‌شناسی رایانه‌ای (تبدیل متن به گفتار، برچسب‌دهی خودکار به اجزاء کلام) و زبان‌شناسی پیکره‌ای از جمله علایق پژوهشی وی است.

محمود بی‌جن‌خان

متولد سال ۱۳۳۷، دارای مدرک تحصیلی دکتری تخصصی در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون استاد و مدیر گروه زبان‌شناسی دانشگاه تهران است.

آواشناسی و واج‌شناسی، زبان‌شناسی رایانه‌ای و زبان‌شناسی پیکره‌ای از جمله علایق پژوهشی وی است.