

## بررسی رابطه خویشاوندی زبان‌های ایرانی با رویکرد زیست‌شناسی تکاملی\*

روزبه تویسرکانی (پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی)  
صفر وفادار دولق (مرکز تحقیقات بیوشیمی و بیوفیزیک، دانشگاه تهران)

**چکیده:** زبان‌های رایج امروزی، محصول صدها و هزاران سال تکامل در جوامع بشری هستند. زبان‌ها یکباره خلق نشده‌اند بلکه در فرایندی تدریجی، از نیاکان خود فاصله گرفته و انشعاب یافته‌اند تا اینکه به شکل زبان‌ها و گویش‌های کنونی درآمده‌اند. مطالعه تاریخیچه این فرایند تکامل و انشعاب، موضوع تحقیقات بسیاری در سالیان گذشته بوده است. با توجه به شباهت‌های زیاد فرایند ایجاد گوناگونی و تکامل زبان‌ها و موجودات زنده و با الهام از مدل‌های توسعه‌یافته در زیست‌شناسی تکاملی، و همچنین ابزارهای محاسباتی حاصل از پیشرفت در علوم رایانه، این امکان فراهم شده است که بتوان از این مدل‌ها در مطالعه تاریخ زبان‌ها و کشف روابط خویشاوندی و زمان تقریبی انشعاب آنها از زبان‌های

---

\* این پژوهش را مدیران راهنمایی‌های ارزشمند جناب آقای دکتر مهدی صادقی هستیم. همچنین، لازم است از گروه زبان‌ها و گویش‌های ایرانی فرهنگستان زبان و ادب فارسی برای در اختیار گذاشتن مجموعه واژه‌های گویش‌های مطالعه‌شده تشکر نماییم. با سپاس فراوان از جناب آقای دکتر رضائی باغبیدی که در تمامی مراحل انجام این پژوهش از راهنمایی‌های بسیار ارزشمندشان بهره بردیم. همچنین از سرکار خانم حسن‌زاده معاون اجرایی گروه به دلیل کمک‌ها و راهنمایی‌های ارزنده‌شان تشکر می‌نماییم.

اجدادی‌شان استفاده کرد. از این روش‌ها برای مطالعه بسیاری از زبان‌ها، به‌ویژه زبان‌های هندواروپایی، استفاده شده است. در این مقاله ما به مطالعه خویشاوندی و انشعاب تعدادی از گویش‌های ایرانی پرداخته و با استفاده از مدل‌های زیست‌شناسی تکاملی، درخت تکاملی (شجره‌نامه) ۴۹ گویش ایرانی را رسم نموده‌ایم. امیدواریم با توسعه روش پیشنهادی و همچنین گردآوری و مستندسازی گویش‌های بیشتر بتوان این مطالعه را به تمامی گویش‌های ایرانی تعمیم داد.

**کلیدواژه‌ها:** زبان‌شناسی تطبیقی، زبان‌های ایرانی، تبارزایش (فیلوژنتیک)، درخت تکامل، گویش‌شناسی

#### مقدمه

در سال ۱۷۶۷ میلادی، پزشکی انگلیسی به‌نام جیمز پارسونز<sup>۱</sup> کتابی با عنوان *بازمانده‌های یافث: بررسی‌های تاریخی درباره خویشاوندی و منشأ زبان‌های اروپایی*<sup>۲</sup> منتشر کرد. او در این کتاب، با بررسی و مقایسه اعداد پایه در زبان‌های آلمانی، اسپانیایی، انگلیسی کهن، ایتالیایی، ایرلندی، بنگالی، دانمارکی، روسی، فارسی، فرانسوی، لاتینی، لهستانی، هلندی، ولش، یونانی و مقابله آنها با برابره‌های کاملاً متفاوتشان در زبان‌های ترکی، چینی، عبری و مالایی (مالزیایی)، به این نتیجه رسید که زبان‌های اروپایی، ایرانی و هندی بازماندگان زبانی واحدند که زبان یافث، پسر نوح، بوده است. آنچه امروزه تحت عنوان خانواده زبان‌های هندواروپایی شناخته می‌شود مجموعه بازماندگان زبانی واحد به‌نام هندواروپایی آغازین<sup>۳</sup> است که سخنگویان آن حدوداً در فاصله سال‌های ۴۵۰۰-۳۶۰۰ ق م می‌زیستند. خانواده زبان‌های هندواروپایی

---

1) James Parsons

2) *The Remains of Japhet: Being Historical Enquiries into the Affinity and Origin of the European Languages*

3) proto-Indo-European

ده شاخه اصلی را دربرمی‌گیرد که عبارت‌اند از آریایی، آلبانیایی، آناتولیایی، ارمنی، ایتالیک، بالتی-اسلاوی، تُخاری، ژرمنی، سِلتی و یونانی (رضائی باغ بیدی ۲۰۰۹: ۱-۴).  
مطالعات بسیاری درباره منشأ زبان‌ها، خویشاوندی آنها، تاریخ جدایی آنها از زبان‌های اجدادی‌شان و همچنین خاستگاه آنها، به همّت زبان‌شناسان انجام شده است. اغلب این بررسی‌ها بر اساس دانش زبان‌شناسی و با روش‌های کیفی و مبتنی بر نظر محقق خبره انجام شده‌اند و از روش‌های کمی و روشمند که نیازمند مدل‌سازی‌های ریاضی و شبیه‌سازی‌های رایانه‌ای هستند، کمتر استفاده شده است.

زیست‌شناسی تکاملی با تاریخ تکامل موجودات زنده سروکار دارد. از نیمه دوم قرن نوزدهم به بعد، به تدریج پذیرفته شد که موجودات زنده طی فرایند انتخاب طبیعی، از نیاکان خود منشعب شده و گونه‌های جدیدی پدید آمده‌اند. درخت‌های فیلوژنتیک طبق نظریه تکاملی داروین، رابطه خویشاوندی و نقاط انشعاب گونه‌های مختلف را نشان می‌دهند. این درخت‌ها بر اساس شباهت‌ها و تفاوت‌های مربوط به ویژگی‌های موجودات رسم می‌شوند. در حال حاضر، مبنای رسم درخت خویشاوندی گونه‌ها اطلاعات توالی‌های DNA و پروتئین‌های موجودات زنده است. در واقع، مولکول‌های DNA (که در اصطلاح از چهار حرف A، C، G و T تشکیل شده‌اند) الفبای حیات هستند و ترکیب متفاوت این الفبای واحد در گونه‌های مختلف منجر به گوناگونی موجودات شده است، همان‌گونه که آواهای زبان‌ها با ترکیب‌های مختلف منجر به پیدایش زبان‌های متفاوت گشته است.

در دهه‌های گذشته، الگوریتم‌های مختلفی برای رسم درخت فیلوژنتیک موجودات با استفاده از الفبای DNA، بر مبنای مدل تکامل داروینی، توسعه یافته‌اند. با این فرض که زبان‌ها نیز، همانند موجودات زنده، نتیجه تکامل و انشعاب از زبان‌های اجدادی هستند، این الگوریتم‌ها برای رسم درخت خویشاوندی زبان‌ها به کار رفته‌اند. در واقع، می‌توان هر زبان (یا گویش) را به صورت یک موجود زنده در نظر گرفت که کلمات آن زبان نقش ژن، و آواها نقش مولکول‌های آن را ایفا می‌کنند.

در سال ۲۰۰۳ مقاله‌ای در مجله نیچر<sup>۱</sup> توسط راسل گری و کوئنتین آتکینسون به چاپ رسید که در آن درخت خویشاوندی زبان‌های هندواروپایی نشان داده شده است (گری و آتکینسون ۲۰۰۳). آنها با این عقیده که زبان‌ها همانند ژن‌ها اطلاعاتی درباره تاریخ انسان فراهم می‌آورند، با استفاده از روش‌های محاسباتی استخراج شده از زیست‌شناسی تکاملی، به تحلیل داده‌های زبان‌شناسی پرداخته‌اند. تحلیل آنها بر اساس ماتریسی از ۸۷ زبان دنیا و با ۲۴۴۹ کلمه انجام شده است. گری و آتکینسون در این مقاله، علاوه بر ترسیم درخت خویشاوندی، زمان جدایی زبان‌ها از یکدیگر را تخمین زده‌اند؛ به‌عنوان مثال در درخت خویشاوندی زبان‌های هندواروپایی رسم شده در این مقاله، نشان داده شده است که زبان‌های پرتغالی و اسپانیایی رابطه خویشاوندی نزدیک‌تری دارند تا زبان انگلیسی (زبان انگلیسی رابطه خویشاوندی دورتری با زبان پرتغالی، در مقایسه با زبان اسپانیایی، دارد) و همچنین تخمین زده‌اند که زبان‌های انگلیسی، آلمانی، سوئدی و دانمارکی، در حدود ۱۷۰۰ سال پیش، از زبانی مشترک منشعب شده‌اند.

در اگوست سال ۲۰۱۲، رامکو بوکائرت و همکارانش مقاله‌ای (بوکائرت و همکاران ۲۰۱۲) با عنوان «ترسیم منشأ و توسعه خانواده زبان هندواروپایی» در مجله معتبر *ساینس*<sup>۲</sup> منتشر کردند که مورد توجه بسیاری از دانشمندان و محققان قرار گرفت و در نشریات و رسانه‌های فراوانی به بحث و جدل در دفاع یا رد نتیجه این مقاله پرداخته شد. آنها در این مقاله، با نگرش تکاملی تبارشناسی جغرافیایی بیزی<sup>۳</sup> و با استفاده از ۲۰۰ واژه از ۱۰۳ زبان باستانی و معاصر هندواروپایی، توسعه زبان‌ها در مناطق مختلف زمین را الگو قرار دادند. آنها دو فرضیه مهم و مقابل هم درباره خاستگاه و منشأ زبان‌های هندواروپایی را بررسی کردند. بر اساس عقیده‌ای رایج که تحت عنوان استپ یا گورچالی مطرح شده است، تاریخ «نیا-زبان» هندواروپایی به حدود ۶۰۰۰ سال پیش و در منطقه استپ روسیه، در شمال دریای خزر باز می‌گردد. در مقابل این فرضیه، فرضیه آناتولیا قرار دارد که پروفیسور کالین رنفرو<sup>۴</sup> برای اولین بار در اواخر دهه ۱۹۸۰ میلادی مطرح کرد و خاستگاه «نیا-زبان» این زبان‌ها را منطقه آناتولی ترکیه و تاریخ آن

1) *Nature*

2) *Science*

3) bayesian phylogeography

4) Colin n r

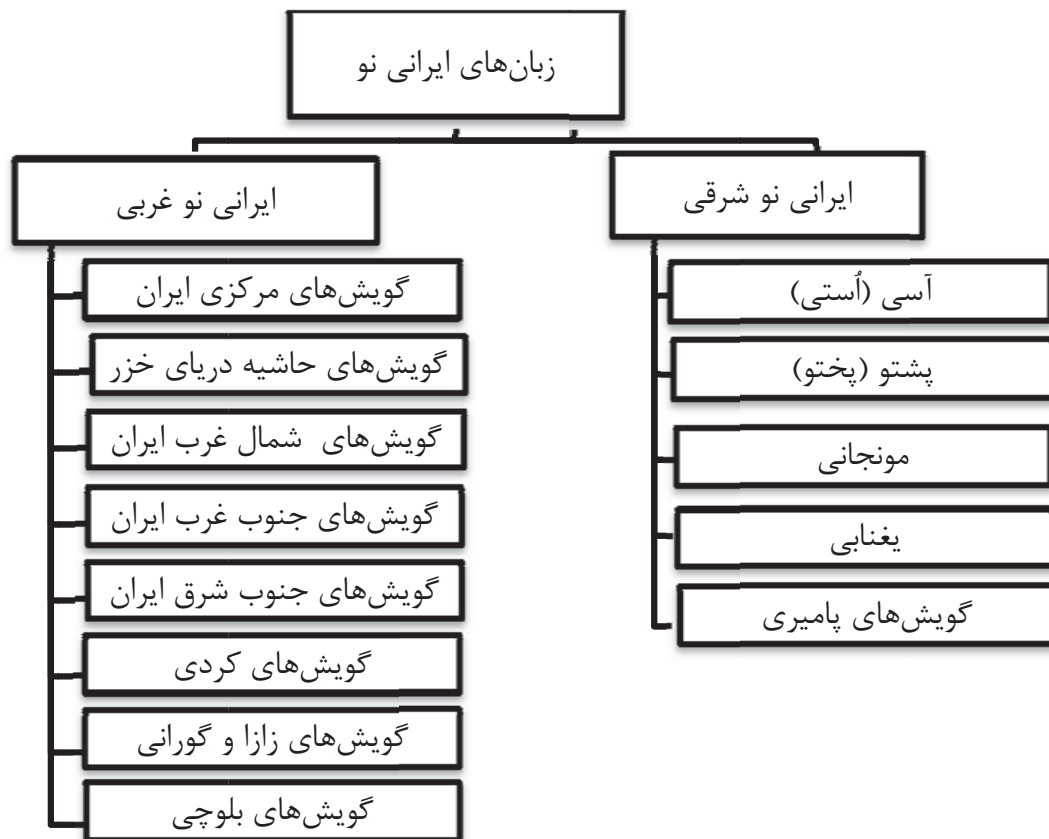
را ۸۰۰۰ تا ۹۵۰۰ سال پیش می‌داند. بوکائرت و همکارانش شواهدی یافته‌اند که فرضیه دوم را تأیید می‌کند. در واقع، این محققان از مدل‌های زیست‌شناسی و ریاضی برای درک بهتر موضوع زبان‌شناسی (که ماهیت تکاملی دارد) بهره جسته‌اند (همان: ۹۵۹-۹۶۰). در این تحقیق، هدف ما رسم درخت خویشاوندی و تکاملی گویش‌های مختلف زبان‌های ایرانی با استفاده از الگوریتم‌های رایج در رسم درخت‌های فیلوژنتیک بوده است. در ادامه مقاله به معرفی گویش‌های مورد مطالعه و روش رسم درخت خواهیم پرداخت.

### زبان‌های ایرانی

زبان‌ها و لهجه‌های ایرانی خانواده واحدی را تشکیل می‌دهند که از یک اصل مشترک منشعب شده‌اند؛ به عبارتی دیگر زبان‌ها و گویش‌ها از یک منشأ یا یک ریشه زبانی واحد مشتق شده‌اند (نک: ارانسکی ۱۳۷۸: ۲۶-۲۸). زبان‌های ایرانی یکی از شاخه‌های زبان‌های هندوایرانی از خانواده بزرگ زبان‌های هندواروپایی را تشکیل می‌دهند. منظور از اصطلاح زبان‌های ایرانی یا زبان‌های ایرانی‌تبار، گروهی از زبان‌هاست که همگی ریشه در یک زبان باستانی به نام زبان ایرانی باستان دارند و به معنی زبان‌های مرتبط با واحد سیاسی کشور امروزی ایران نیست. بنابراین، زبان‌های ایرانی یک ریشه مشترک باستانی دارند و هر زبان به گویش‌های مختلف و هر گویش به لهجه‌های مختلف تقسیم می‌شود. بعضی از گویش‌های زبان‌های ایرانی به دلیل تفاوت‌های ظاهری در نحوه استفاده از عبارات و ترکیب‌ها، امروزه خود به عنوان یک زبان شناخته می‌شوند (ویندفور ۲۰۰۹: ۴۱۸).

زبان‌های ایرانی بیشتر در ایران، افغانستان، تاجیکستان، غرب پاکستان، کردستان ترکیه، کردستان عراق و بخش‌هایی از آسیای مرکزی و قفقاز رواج دارند. برآورد می‌شود که امروزه حدود ۱۵۰ تا ۲۰۰ میلیون نفر به زبان‌های ایرانی سخن می‌گویند. طبق برآورد بنیاد زبان‌شناسی در سال ۱۳۸۴ امروزه حدوداً به ۸۷ گونه از زبان‌های ایرانی سخن گفته می‌شود، که پرکاربردترین این زبان‌ها به ترتیب شمار تقریبی سخنوران عبارت‌اند از: فارسی (۱۱۰ میلیون نفر)، پشتو (۶۰-۴۱ میلیون نفر)، کردی (۲۵ میلیون نفر)، لری (۱۰ میلیون نفر)، و بلوچی (۷ میلیون نفر) (همان‌جا).

زبان‌های ایرانی از نظر تاریخی به سه دسته تقسیم می‌شوند: (۱) زبان‌های ایرانی باستان؛ (۲) زبان‌های ایرانی میانه؛ (۳) زبان‌های ایرانی نو. زبان‌های ایرانی نو در دو گروه عمده ایرانی نو شرقی و ایرانی نو غربی جای می‌گیرند. هر کدام از این گروه‌ها، به چندین زیرگروه تقسیم می‌شوند که هر یک شامل چندین گویش‌اند. این تقسیم‌بندی به صورت شماتیک در شکل (۱) نشان داده شده است. به عنوان مثال، گویش‌های مرکزی ایران اغلب در منطقه میان اصفهان، تهران، همدان و یزد رواج دارند و به شش گروه (شمال غربی، شمال شرقی، جنوب غربی، جنوب شرقی، منطقه تفرش، و دشت کویر) تقسیم می‌شود. گویش‌های گروه شمال شرقی در منطقه کاشان و نطنز رایج هستند و شمار آنها بسیار زیاد است. از مهم‌ترین آنها می‌توان به گویش‌های کلیمیان کاشان، گویش آرانی، کشه‌ای، طاری، نطنزی، ابوزیدآبادی و بادرودی، اشاره کرد (رضائی باغ‌بیدی ۲۰۰۹: ۱۷۹-۱۸۱).



شکل ۱. دسته‌بندی زبان‌های ایرانی نو

## فیلوژنی (درخت تکاملی) در زیست‌شناسی

فیلوژنتیک شاخه‌ای در علم زیست‌شناسی است که با استفاده از داده‌های توالی‌یابی مولکولی<sup>۱</sup> و ماتریس‌های داده‌های ریخت‌شناسی<sup>۲</sup>، به بررسی ارتباط تکاملی گونه‌های مختلف جانوران می‌پردازد. روند تکاملی را می‌توان توسط یک درخت فیلوژنی مجسم کرد. معمول‌ترین روش‌ها برای استنباط فیلوژنی شامل صرفه‌جویی<sup>۳</sup> (یا کمترین فرضیات)، درست‌نمایی حداکثری<sup>۴</sup>، و استنتاج بیزی<sup>۵</sup> هستند.

همه این روش‌ها وابسته به مدل‌های ریاضی ضمنی یا صریحی هستند که تکامل شاخصه‌های مشاهده‌شده در گونه‌های مورد مطالعه را توضیح می‌دهند. این مدل‌ها معمولاً بر اساس داده‌های مولکولی ساخته می‌شوند و شاخصه‌ها در این نوع مدل‌ها نوکلئوتیدها یا رشته‌های اسیدآمینۀ ترازشده<sup>۶</sup> هستند. اطلاعات (داده‌های ورودی) و الگوریتمی که برای محاسبه درخت انتخاب می‌شود، دو معیار اصلی برای ساخت درخت‌های فیلوژنی محسوب می‌شوند (بلومبرگ و همکاران ۲۰۰۳: ۷۱۹-۷۲۴).

## روش کار

سخن‌گویان هرزبانی، علاوه بر مجموعه نظام‌های آوایی، صرفی، نحوی و معنایی زبان خود، واژگان آن زبان را نیز در ذهن دارند. مهم‌ترین بخش واژگان دانستن واژه از دیدگاه معنایی است. بخش واژگان زبان لزوماً از دگرگونی‌های فرهنگی و اجتماعی تأثیر می‌پذیرد و به همین دلیل، بیش از هر بخش دیگر زبان تغییر می‌کند. ما، در این نوشتار، ۲۰۰ واژه (معادل مجموعه داده‌های سوادش<sup>۷</sup>، که مبنای کار رامکو بوکائرت و همکارانش بود و در مقدمه توضیح داده شد) را در ۴۹ گویش که به همت گروه زبان‌ها و گویش‌های ایرانی فرهنگستان زبان و ادب فارسی جمع‌آوری شده و در

1) molecular sequencing data

2) morphological data matrices

3) parsimony

4) maximum likelihood

5) bayesian inference

6) aligned

7) a s

مجموعه کتاب‌هایی با عنوان گنجینه گویش‌های ایرانی به چاپ رسیده، بررسی کردیم. این گویش‌ها عبارت‌اند از ۱- التپه ۲- رستم‌کلا ۳- مهدیرجه ۴- یخکش ۵- دیباج (چارده‌کلاته) ۶- پوروا ۷- نوکنده ۸- بنفشه‌تپه ۹- جفاکنده ۱۰- طرّقی ۱۱- طاری ۱۲- کِشه‌ای ۱۳- طامه‌ای ۱۴- نطنزی ۱۵- تکیه‌ای ۱۶- پاه‌ای ۱۷- نودشه ۱۸- شرکان ۱۹- دزلی ۲۰- بزlane ۲۱- گلّه‌ری ۲۲- گورانی ۲۳- سَنجایی ۲۴- کولیایی ۲۵- زنگنه‌ای ۲۶- جالالوندی ۲۷- زوله‌ای ۲۸- کاکاوندی ۲۹- هوزمانوندی ۳۰- پاچه‌لک ۳۱- لاسگردی ۳۲- دوانی ۳۳- دهله‌ای ۳۴- عبدویی ۳۵- کازرونی ۳۶- کلانی ۳۷- کنده‌ای ۳۸- کوزرگی ۳۹- ممسنی ۴۰- ماسرمی (دهسروی) ۴۱- بلیانی ۴۲- بیروکانی ۴۳- حیاتی (دولت‌آبادی) ۴۴- دادنجانی ۴۵- لردارنگانی ۴۶- درونکی (مهبودی) ۴۷- دزگاهی (گوری) ۴۸- کرشی ۴۹- کرونّی.

در انتخاب مجموعه داده‌ها، با محدودیت‌های بسیاری مواجه بودیم. کتاب‌ها و تحقیقات متعددی به گردآوری مجموعه واژه‌های گویش‌های ایرانی پرداخته‌اند، که می‌توان به مواردی مانند (آذرلی ۱۳۸۷؛ اسفندیاری ۱۳۸۰؛ شالچی ۱۳۷۰؛ کیا ۱۳۹۰؛ محمدی املشی و غلامی ۱۳۹۱) اشاره کرد که به دلیل ناهمگونی با دیگر مجموعه‌ها و در موارد بسیاری، عدم مستندسازی مناسب، کنار گذاشته شدند. تنها مجموعه واژه‌هایی که گروه زبان‌ها و گویش‌های ایرانی فرهنگستان زبان و ادب فارسی آنها را گردآوری کرده‌اند ویژگی‌های لازم برای انجام کار را داشتند، که متأسفانه از لحاظ جغرافیایی محدود است. امیدواریم در آینده، با پیشرفت طرح جمع‌آوری واژگان گویش‌های ایرانی در فرهنگستان، ما نیز بتوانیم الگوریتم را بر روی مجموعه بزرگ‌تری از گویش‌ها پیاده‌سازی کنیم.

گنجینه گویش‌های ایرانی، که پیش‌تر به آن اشاره کردیم، شامل حدود ۲۵۰۰ واژه و ۱۰۰ جمله از گویش‌های اصیل ایرانی است که ما با توجه به نیازمان، ۲۰۰ واژه (معادل مجموعه واژه‌های سوادش) را انتخاب کرده‌ایم. مجموعه واژه‌های سوادش، دو فهرست ۱۰۰ و ۲۰۰ واژه‌ای است<sup>۱</sup> که زبان‌شناس آمریکایی، موریس سوادش در طول سال‌های

(۱) در سایت Wiktionary لیست دوم با ۲۰۷ معادل آورده شده است.



۱۹۴۰ تا ۱۹۵۰، آنها را با هدف بررسی ارتباط زبان‌ها با روش آمارواژگانی<sup>۱</sup>، تهیه کرده است، فهرست ۲۰۰ واژه‌ای شامل واژه‌هایی است که مختص زبان خاصی نیستند و در تمام زبان‌ها وجود دارند. سوادش واژه‌ها را به صورت شهودی انتخاب کرده است.<sup>۲</sup> به عنوان نمونه، می‌توان به واژه‌هایی مانند اعداد چهار و پنج، دریا، برف و کوه اشاره کرد. ویژگی این واژه‌ها جهان‌شمول بودن، مستقل بودن از فرهنگ و همچنین در دسترس بودن آنهاست. فهرست کامل این ۲۰۰ واژه در پیوست این مقاله آمده است.

به نظر می‌رسد معیاری قاطع برای تعیین مرز دقیق میان مفاهیم زبان و گویش که در تمام موارد قابل استفاده باشد، وجود ندارد؛ زیرا گویش شعبه‌ای از زبان است و معیارهای زبانی (فهم متقابل) و غیرزبانی (مرزهای سیاسی، نگرش افراد به زبان‌ها، ارزش و اعتبار اجتماعی آنها، وسعت و قلمرو جغرافیایی، خوشاوندی تاریخی و...) در تفکیک آن دو تأثیر دارند؛ بنابراین تشخیص زبان از گویش دشوار است (مدرسی ۱۳۸۴: ۱۳۴). به همین دلیل و برای سادگی در بحث، در ادامه، ما همواره از واژه زبان استفاده خواهیم کرد و به بررسی ۴۹ زبان خواهیم پرداخت.

برای محاسبه فاصله میان زبان‌ها از فاصله لونیشتاین<sup>۳</sup> استفاده کرده‌ایم. در این روش، فاصله دو زبان از طریق جمع جبری فاصله میان واژه‌های معادل در آن دو زبان به دست می‌آید. اگر فاصله میان دو زبان (گویش)  $l1$  و  $l2$  را با  $distance(l1, l2)$  و فاصله دو واژه  $w_1^k$  و  $w_2^k$  (که واژه‌های معادل  $k$ -امین واژه مجموعه سوادش هستند) را با  $d(w_1^k, w_2^k)$  نشان دهیم، معادله‌ای به صورت زیر خواهید داشت:

$$distance(l1, l2) = \sum_{k=1}^n d(w_1^k, w_2^k)$$

که در آن،  $n$  تعداد واژه‌های انتخابی (یعنی ۲۰۰) است. برای محاسبه فاصله دو واژه در روش لونیشتاین، ابتدا باید دو واژه را با یکدیگر تراز کرد و سپس تعداد اختلاف‌های

1) lexicostatistics

2) [http://en.wikipedia.org/wiki/Swadesh\\_list](http://en.wikipedia.org/wiki/Swadesh_list)

3) levenshtein distance

آنها را به دست آورد. به عنوان مثال، دو واژه *nāṭān* و *nāṭmūn* (به معنای «گفتن»، در کشای و طامه‌ای) به صورت زیر تراز می‌شوند:

v	â	t		a	n
v	â	t	m	u	n

در صورتی که به حذف و تفاوت امتیاز ۱ بدهیم، فاصله این دو واژه برابر با ۲ خواهد شد؛ چون دارای یک حذف و یک تفاوت است. البته نحوه امتیازدهی ممکن است کاملاً متفاوت باشد؛ یعنی برای تفاوت، امتیاز ۲ در نظر گرفته شود، زیرا هر تفاوت را می‌توان معادل یک حذف و یک جایگزینی در نظر گرفت. پس از محاسبه ماتریس فاصله‌ها (ماتریس متقارن ۴۹×۴۹، که عنصر  $m_{im}$  بیانگر فاصله زبان  $i$  ام و زبان  $m$  ام است) با الگوریتم  $\tilde{N}i\tilde{o}N$  (بهبودیافته الگوریتم نزدیک‌ترین مجاور)، درخت فیلوژنی مربوط به زبان‌ها بازسازی شد؛ سپس با استفاده از نرم‌افزار رسم درخت  $(\tilde{R}i\tilde{u}T\tilde{r}\tilde{e}\tilde{e})$ ، درخت به دست آمده رسم شد. درخت‌های رسم شده در شکل‌های (۲) و (۳) نشان داده شده‌اند.

در حین پیاده‌سازی الگوریتم با برخی چالش‌ها مواجه بودیم که در ادامه، برای روشن شدن روش کارمان، به برخی از این چالش‌ها اشاره می‌کنیم:

— در برخی از گویش‌ها، برای مفهومی خاص، چندین واژه به کار برده می‌شود؛ به عنوان مثال، در گویش بزخانه برای مفهوم «سنگین» واژه‌های  $\text{ṣāṇḥūiṇ} | \text{ṣāṇḥūiṇāāūṛṣ}$  کاربرد دارند. برای اندازه‌گیری فاصله میان گویش‌ها، لازم بود یکی از واژه‌ها را انتخاب کنیم و محاسبه فاصله گویش‌ها را بر اساس آن واژه انجام دهیم. برای افزایش دقت روش، تصمیم گرفتیم واژه‌ای را انتخاب کنیم که بیشترین شباهت را به واژه‌های معادل در دیگر زبان‌ها داشته باشد؛ به این منظور برای هر واژه، ابتدا زبان‌هایی را که برای آن مفهوم تنها یک واژه دارند در نظر گرفتیم و برای بقیه زبان‌ها، به ازای هر واژه، مجموع فاصله از زبان‌های دسته اول را محاسبه کردیم و واژه‌ای را برگزیدیم که کمترین مجموع فاصله را از زبان‌های تک‌واژه‌ای داشته باشد. به عنوان نمونه، مفهوم

۱)  $\text{ṇēiūhḥōr-ṇṇiṇiṇū}$

«شپش» در زبان‌های دوانی، دهله‌ای، عبدویی، کازرونی، کلانی، کنده‌ای در جدول (۱) نمایش داده شده است. با توجه به ۴۸ زبان دیگر (در این مفهوم، تنها یک زبان دو معادل داشت. در برخی مفاهیم، تعداد زبان‌هایی که بیش از یک واژه دارند بیشتر است)، برای معادل «شپش» در زبان کنده‌ای واژه *šêš* انتخاب و واژه *ûêṇḍâr* کنار گذاشته شد.

جدول ۱. واژه‌های معادل «شپش» در شش زبان استان فارس

دوانی	دهله‌ای	عبدویی	کازرونی	کلانی	کنده‌ای
šīš	šīš	šêš	šīš	šêš	šêš, ûêṇḍâr

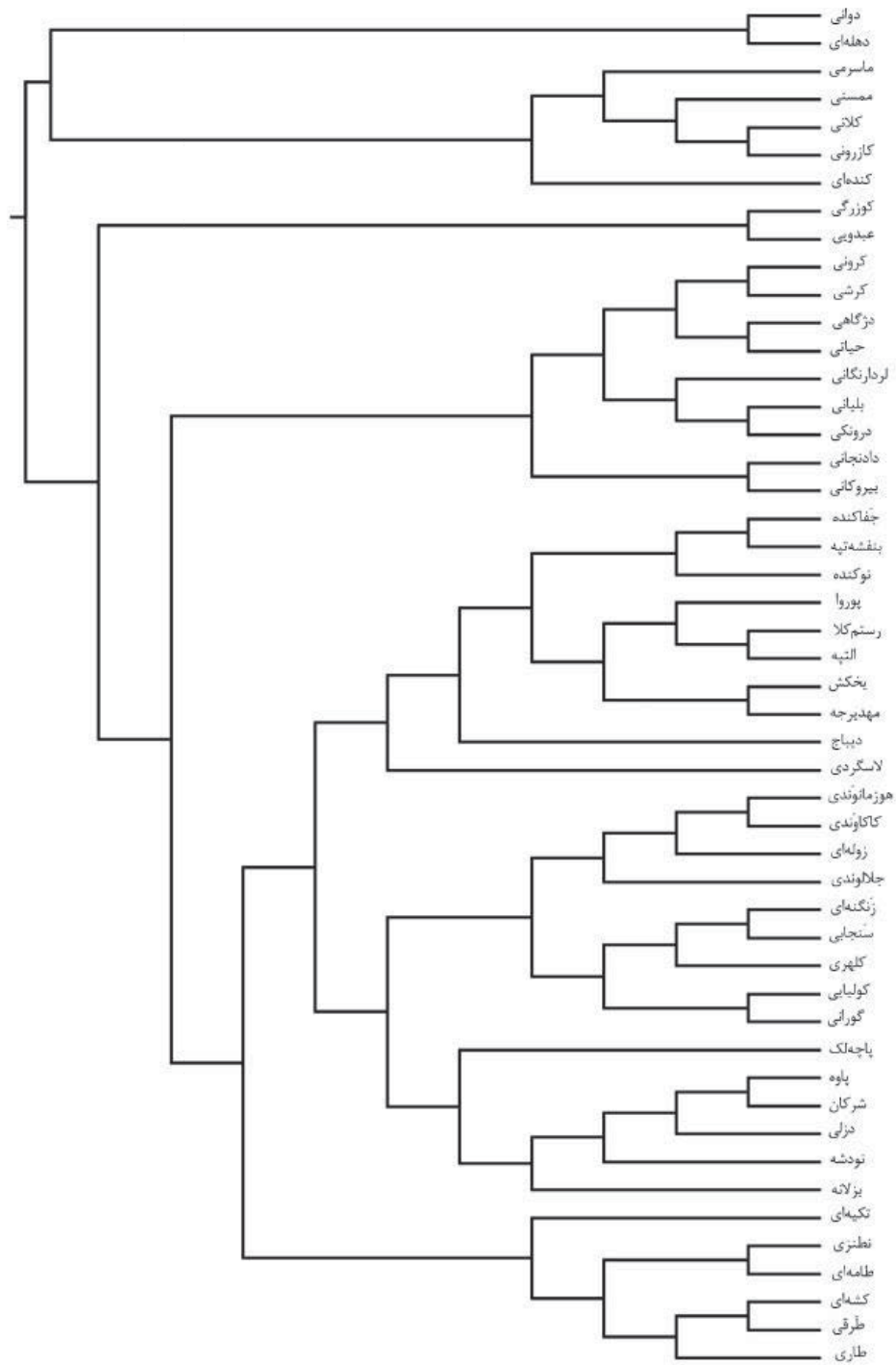
— چالش دیگر نداشتن برخی از واژه‌ها بود. واژه‌های «چه» و «فکر کردن» را به دلیل موجود نبودن در مجموعه داده‌ها، از انتخاب خود کنار گذاشتیم و به جای ۲۰۰ واژه، ۱۹۸ واژه را به کار بردیم.

— چالش دیگر در سطح آواها بود. برای صامت‌ها تفاوت‌ها را کنار گذاشتیم، به عنوان مثال، *ñ* و *n* را به *n* تبدیل کردیم. اما تفاوت در مصوت‌ها را حفظ کردیم.

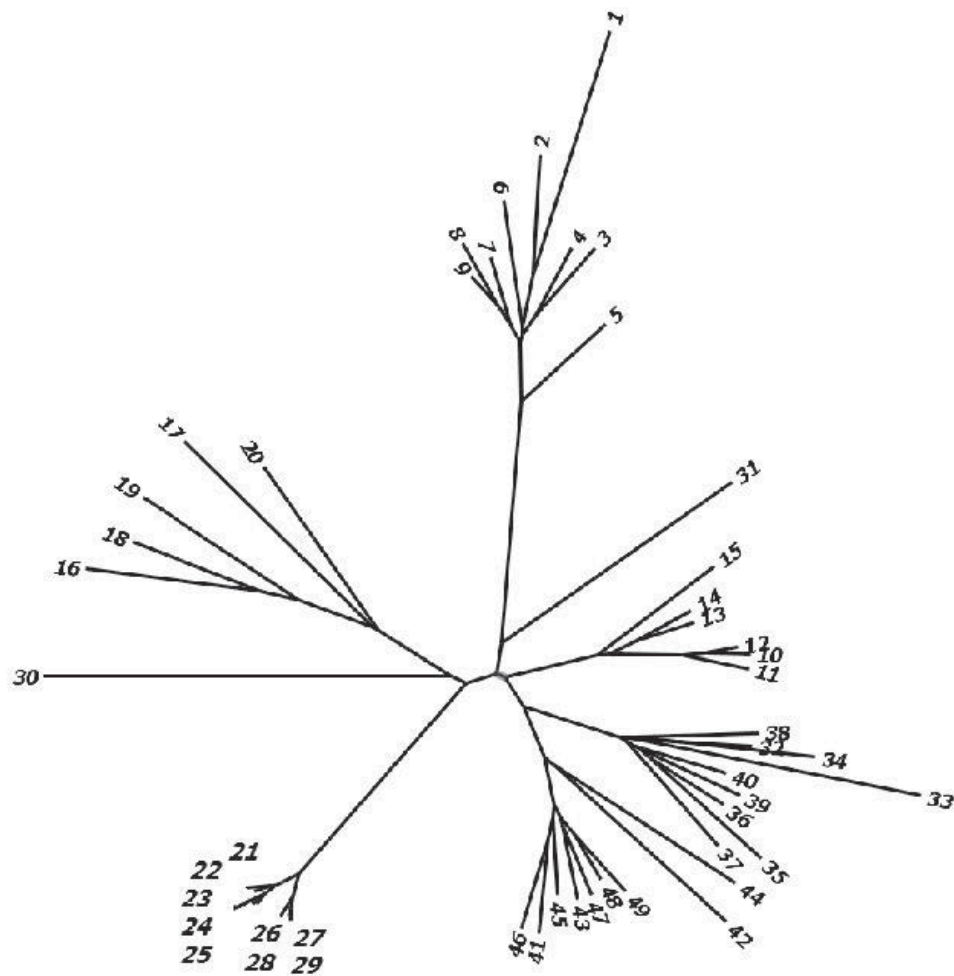
### نتیجه‌گیری

درخت به دست آمده مطابقت معناداری با جغرافیای زبان‌ها دارد. در شکل (۴) دسته‌بندی زبان‌ها با شماره آنها بر روی نقشه ایران نشان داده شده است. در شکل (۳)، در طرح‌بندی درخت شعاعی، زبان‌های ۱ تا ۹ در یک خوشه قرار گرفته‌اند که مطابق با زبان‌های استان مازندران هستند. همچنین زبان لاسگردی (شماره ۳۱)، که در استان سمنان رایج است، از این زبان‌ها فاصله کمتری دارد.

همان‌طور که در شکل (۳) می‌توان مشاهده کرد، خوشه مربوط به زبان‌های شماره ۲۱ تا ۲۹ تراکم بیشتری دارد و زبان‌های این خوشه فاصله بسیار کمی از یکدیگر دارند. در واقع، می‌توان این‌گونه استنباط کرد که زبان‌های شماره ۲۱ تا ۲۹، نسبت به دیگر زبان‌های مورد مطالعه، کمتر دستخوش تغییرات شده‌اند.



شکل ۲. درخت خویشاوندی رسم‌شده برای ۴۹ زبان ایرانی با روش BioNJ (طرح‌بندی درخت مستطیلی)



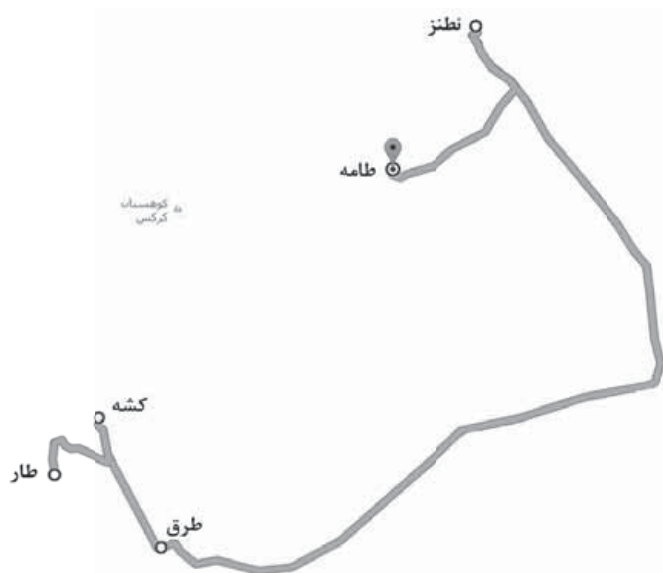
شکل ۳. درخت خویشاوندی رسم شده برای ۴۹ زبان ایرانی با روش BioNJ (طرح‌بندی درخت شعاعی). شماره رأس‌ها مطابق با شماره ارائه شده برای زبان‌ها در ابتدای بخش «روش کار» است.



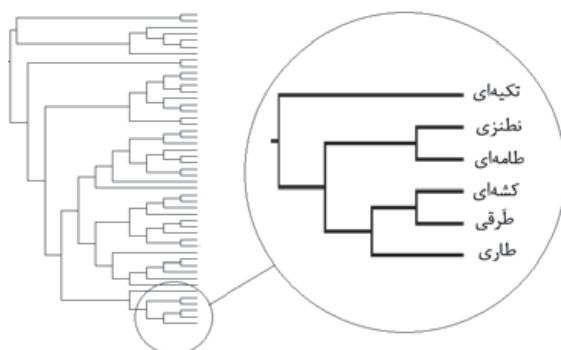
شکل ۴. نقشه ایران و دسته‌بندی زبان‌های مورد مطالعه (این زبان‌ها در استان‌های مازندران، اصفهان، کرمانشاه، کردستان، لرستان، سمنان و فارس رواج دارند).

در شاخهٔ مربوط به گروه B (زبان‌های طَرَقی، طاری، کِشه‌ای، طامه‌ای، نطنزی، تکیه‌ای) زبان‌های نطنزی و طامه‌ای رابطهٔ خویشاوندی نزدیک‌تری با یکدیگر دارند تا زبان طاری. همان‌طور که در شکل (۵)، نقشهٔ مربوط به استان اصفهان و شهر نطنز مشخص است، شهرهای نطنز و طامه، نسبت به طار، به یکدیگر نزدیک‌ترند. از طرف دیگر، بر روی نقشه، فاصلهٔ طار و کشه کمتر از فاصلهٔ طرق و کشه است ولی در درخت به‌دست‌آمده، که در شکل (۶) نمایش داده شده است، زبان‌های طَرَقی و کشه‌ای، نسبت به طاری، در شاخهٔ نزدیک به هم قرار دارند. با کمی دقت در جادهٔ بین شهرها، می‌توان پی برد که جادهٔ وصل‌کنندهٔ کشه به طار مسیری است که از نزدیکی حدفاصل طرق و طار می‌گذرد. البته این راه‌ها مربوط به سال‌های اخیر است و در قیاس با زمان شکل‌گیری و جدایی این زبان‌ها، زمان بسیار ناچیزی محسوب می‌شود اما نحوهٔ

شکل‌گیری جاده‌ها خود متأثر از عوامل دیگری مانند سادگی ایجاد مسیر (در نتیجه سادگی تردد)، الویت و حجم مراودات شهرها و... است که با معناست و نیاز به تحلیل مناسب دارد. از سوی دیگر، زبان تکیه‌ای از پنج زبان دیگر فاصله بیشتری دارد که کاملاً با فاصله روی نقشه و مسیر ارتباطی آنها مطابقت دارد.



شکل 5. نقشه بخشی از استان اصفهان، مسیر میان شهرهای نطنز، طامه، طرق، کوشه و طار با استفاده از نقشه گوگل. مسیر میان این شهرها، با شکل درخت خویشاوندی زبان‌های این شهرها، همخوانی معناداری دارد.



شکل 6. بخشی از درخت خویشاوندی زبان‌های تکیه‌ای، نطنزی، طامه‌ای، کوشه‌ای، طرقی، طاری (درخت خویشاوندی تمامی ۴۹ زبان در شکل ۲ نمایش داده شده بود، در اینجا تمرکز بر بخشی از درخت است).

در بخش زبان‌های استان فارس هم درخت به‌دست‌آمده، معنادار است؛ به‌عنوان مثال، زبان‌های دهله و دوان در شاخه‌های نزدیک‌تری قرار دارند تا زبان ماسرمی (دهسروی) و بر روی نقشه نیز فاصله میان شهرها بیانگر همین نکته است. اما نکته جالب توجه زبان‌های عبدویی، کلانی و ممسنی است. با اینکه عبدویی و کلانی نسبت به نورآباد (ممسنی) فاصله بسیار کمی دارند، در درخت به‌دست‌آمده، زبان‌های ممسنی و کلانی به یکدیگر نزدیک‌ترند تا عبدویی. علت این امر این است که در کلان، به دو زبان لری و تاجیکی تکلم می‌شود که واژه‌های گردآوری‌شده مربوط به زبان لری است و به‌طور طبیعی با زبان لری ممسنی رابطه خویشاوندی نزدیک‌تری دارد. یعنی از درخت به‌دست‌آمده، علاوه بر اطلاعات مربوط به جغرافیا، می‌توان اطلاعات تاریخی از جمله مهاجرت‌ها را نیز استخراج کرد.

درخت‌های فیلوژنی صرفاً ابزاری برای طبقه‌بندی داده‌ها نیستند؛ تحلیل کیفی و کمی این درخت‌ها می‌تواند در بررسی خویشاوندی دور یا نزدیک زبان‌ها و چگونگی انشعاب‌ها و در صورت مطالعه دقیق‌تر، تخمینی از زمان جدایی زبان‌ها، به کار آید و اطلاعات باارزشی درباره تکامل فرهنگی و اجتماعی جوامع در اختیار ما قرار دهد.

در پایان، لازم به یادآوری است که تحلیل درخت فیلوژنی مربوط به زبان‌ها نیاز به دانش‌های مختلفی از جمله زبان‌شناسی، تاریخ، جامعه‌شناسی و... دارد. هدف ما از ارائه این نوشتار تنها معرفی روش کار و ارائه نتیجه حاصل از پیاده‌سازی الگوریتم‌های ساخت درخت فیلوژنی بود. آنچه مسلم است ناهمخوانی‌هایی است که این درخت با برخی از دانسته‌های دانشمندان و محققان زبان‌شناسی ما دارد؛ به‌عنوان مثال، در کتاب *تاریخ زبان‌های ایرانی*، در بخش گویش‌های جنوب غرب ایران، گویش‌های دوانی و کازرونی در یک گروه و گویش لری ممسنی در گروه دیگری قرار دارند (رضائی باغبیدی ۲۰۰۹) اما در درخت ترسیم‌شده (شکل ۲) گویش‌های ممسنی و کازرونی رابطه خویشاوندی نزدیک‌تری دارند و گویش دوانی در شاخه دورتری جای گرفته است.

این ناهمخوانی‌ها را می‌توان به دو دسته تقسیم کرد. دسته نخست این ناهمخوانی‌ها اطلاعات جدیدی است که کمتر مورد توجه بوده و می‌تواند به‌عنوان



دستاوردی نوین در اختیار زبان‌شناسان قرار گیرد تا از منظری جدید به مقوله زبان‌شناسی نگریسته شود و محققان زبان‌شناس با دانش بیشتر و کار تخصصی‌تر به بررسی این موارد پردازند. دسته دوم خطاهای مربوط به داده‌های ورودی برنامه رایانه‌ای ما است — که باتوجه به حجم بالای داده‌ها، امری محتمل است — و همچنین خطاهایی است که به دلیل نبود دانش کافی در مدل‌سازی دقیق به وجود آمده است، که امیدواریم با مطالعه بیشتر و در تحقیقات آتی، برطرف شوند.

### منابع و کتابنامه

- آذرلی، غلامرضا، ۱۳۸۷، فرهنگ واژگان گویش‌های ایرانی، تهران.  
ارانسکی، یوسیف، ۱۳۷۸، زبان‌های ایرانی، ترجمه علی‌اشرف صادقی، تهران.  
اسفندیاری، احمد، ۱۳۸۰، گویش بروجردی، بروجرد.  
رضائی باغبیدی، حسن، ۲۰۰۹، تاریخ زبان‌های ایرانی، اوساکا.  
شالچی، امیر، ۱۳۷۰، فرهنگ گویشی خراسان بزرگ، تهران.  
کیا، صادق، ۱۳۹۰، واژه‌نامه شخصت و هفت گویش ایرانی، تهران.  
محمدی املشی، نصرالله‌پور و غلامی، حسین، ۱۳۹۱، فرهنگ واژگان تاتی شال، تهران.  
مدرسی، یحیی، ۱۳۸۴، درآمدی بر جامعه‌شناسی زبان، تهران.
- Orlitzky, S. et al., 2003, "Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits are More Labile", *Evolution: International Journal of Organic Evolution*, vol. 57, no. 4., pp. 717-745.
- Orlitzky, R. et al., 2012, "Mapping the Origins and Expansion of the Indo-European Language Family", *Science*, vol. 337, pp. 957-960.
- Orlitzky, R. and Olson, Q., 2003, "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin", *Nature*, vol. 426, pp. 435-439.
- Orlitzky, G. and Orloff, J. R., 2009, "Persian and Tajik", *The Iranian Languages*, ed. G. Orloff, London. pp. 416-544.

پیوست:

۲۰۰ واژه مجموعه سوادش

انگلیسی	فارسی
to breath	نفس کشیدن
to burn (intransitive)	سوختن
child (young)	بچه
cloud	ابر
cold (weather)	سرد
to come	آمدن
to count	شمردن
to cut	بریدن
day (not night)	روز
to die	مردن
to dig	کندن زمین
dirty	کثیف
dog	سگ
to drink	نوشیدن
dry (substance)	خشک
dull (knife)	کند
four	چهار
to freeze	یخ زدن
fruit	میوه
to give	دادن
good	خوب

انگلیسی	فارسی
all	همه
and	و (واو عطف)
animal	حیوان
ashes	خاکستر اجاق
back	پشت
bad	بد
bark (of a tree)	پوست
because	چون
belly	شکم
big	بزرگ
bird	پرند
to bite	گاز گرفتن
black	سیاه
blood	خون
to blow (wind)	وزیدن
bone	استخوان
dust	خاک
ear	گوش
earth (soil)	زمین
to eat	خوردن
egg	تخم مرغ

انگلیسی	فارسی
grass	علف
green	سبز
guts	دل و روده
hair	مو
hand	دست
head	سر
to hear	شنیدن
heart	قلب، دل
heavy	سنگین
here	اینجا
to hit	زدن
hold (in hand)	گرفتن
how	چگونه
to hunt (game)	شکار
husband	شوهر
I	من
ice	یخ
if	اگر
new	تازه
night	شب
nose	بینی
not	نه
old	پیر

انگلیسی	فارسی
eye	چشم
to fall (drop)	افتادن
far	دور
fat (substance)	چرب
father	پدر
to fear	ترسیدن
feather (large)	پر
few	کم
to fight	جنگ
fire	آتش
fish	ماهی
five	پنج
to float	سر رفتن (سر ریز)
to flow	روان
flower	گل
to fly	پرواز کردن
fog	مه
foot	پا
in	در
to kill	کُشتن
know (facts)	دانستن
lake	دریاچه
to laugh	خندیدن

انگلیسی	فارسی
one	یک
other	دیگر
person	شخص
to play	بازیگوشی
to pull	کشیدن
to push	هُل دادن
to rain	باریدن
red	قرمز
right (correct)	راست ≠ دروغ
right (hand)	راست
river	رودخانه
road	جاده
root	ریشه
rope	طناب
rotten (log)	پوسیده
rub	مالیدن
salt	نمک
sand	شن
to stand	ایستادن
to stab (or stick)	زدن
star	ستاره
stick (of wood)	چوب
stone	سنگ

انگلیسی	فارسی
leaf	برگ
left (hand)	چپ
leg	پا
to lie (on side)	خوابیدن
to live	زنده
liver	کبد (جگر)
long	دراز
louse	شپش
man (male)	مرد
many	زیاد
meat (flesh)	گوشت
mother	مادر
mountain	کوه
mouth	دهان
name	نام (اسم)
narrow	باریک
near	نزدیک
neck	گردن
to say	گفتن
scratch (itch)	خارانندن
sea (ocean)	دریا
to see	دیدن
seed	تخم (بذر)

انگلیسی	فارسی
straight	راست
to suck	مکیدن
sun	آفتاب
to swell	ورم (آماس)
to swim	شنا کردن
tail	دم
that	آن
there	آنجا
they	ایشان
thick	کلفت
thin	نازک
to think	فکر کردن
this	این
thou	تو
three	سه
to throw	انداختن
to tie	بستن
tongue	زبان
white	سفید
who	چه کسی
wide	پهن
wife	زن (همسر)
wind (breeze)	باد

انگلیسی	فارسی
to sew	دوختن
sharp (knife)	تیز
short	کوتاه
to sing	آواز
to sit	نشستن
skin (of person)	پوست
sky	آسمان
to sleep	خوابیدن
small	کوچک
to smell (perceive odor)	بویدن
smoke	دود
smooth	صاف
snake	مار
snow	برف
some	کمی
to spit	تف کردن
to split	شکافتن
to squeeze	فشار دادن
tooth (front)	دندان
tree	درخت
to turn (veer)	پیچیدن
two	دو
to vomit	قی کردن

انگلیسی	فارسی
wing	بال
wipe	نظافت
with (accompanying)	با
woman	زن
woods	جنگل
worm	کرم
you	تو / شما
year	سال
yellow	زرد

انگلیسی	فارسی
to walk	گردش کردن
warm (weather)	گرم
to wash	شستن
water	آب
we	ما
wet	خیس
what	چه
when	چه وقت
where	کجا