بررسی رابطهٔ خویشاوندی زبـان‌های ایـرانی با رویکرد زیست‌شناسی تکاملی*

روزبه تویسرکانی (پژوهشکدهٔ 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. مطالعهٔ تاریخچهٔ 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 مطالعهٔ تاریخ زبان‌ها و کشف روابط خویشاوندی و زمان تقریبی انشعاب آنها از زبان‌های

* این پژوهش را مدیون راهنمایی‌های ارزشمند جناب آقای دکتر مهدی صادقی هستیم. همچنین، لازم است از گروه زبان‌ها وگویش‌های ایرانی‌فرهنگستان 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 مطالعهٔ بسیاری از زبان‌ها، به‌ویژه زبان‌های هندواروپایی، استفاده شده است. در این مقاله ما به مطالعهٔ خویشاوندی و انشعاب تعدادی از گویش‌های ایرانی پرداخته و با استفاده از مدل‌های زیست‌شناسی تکاملی، درخت تکاملی (شجره‌نامهٔ) ٤٩ گویش ایرانی را رسم نموده‌ایم. امیدواریم با توسعهٔ 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 ۱۷٦۷میلادی، پزشکی انگلیسی به‌نام جیمز پارسونز[1] کتابی با عنوان *بازمانده‌های یافث: بررسی‌های تاریخی دربارهٔ خویشاوندی و منشأ زبان‌های اروپایی*[2] منتشر کرد. او در این کتاب، با بررسی و مقایسهٔ اعداد پایه در زبان‌های آلمانی، اسپانیایی، انگلیسی کهن، ایتالیایی، ایرلندی، بنگالی، دانمارکی، روسی، فارسی، فرانسوی، لاتینی، لهستانی، هلندی، ولش، یونانی و مقابلهٔ آنها با برابرهای 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 یافِث، پسر نوح، بوده است. آنچه امروزه تحت عنوان خانوادهٔ زبان‌های هندواروپایی شناخته می‌شود مجموعهٔ بازماندگان زبانی واحد به‌نام هندواروپایی آغازین[3] است که سخنگویان آن حدوداً در فاصلهٔ سال‌های ٤٥۰۰–۳٦۰۰ ق‌م می‌زیستند. خانوادهٔ 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 شاخهٔ اصلی را دربرمی‌گیرد که عبارت‌اند از آریایی، آلبانیایی، آناتولیایی، ارمنی، ایتالیک، بالتی‌ـاسلاوی، تُخاری، ژرمنی، سِلتی و یونانی (رضائی باغ بیدی ۲۰۰۹: ۱–٤).

مطالعات بسیاری دربارهٔ 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 کمّی و روشمند که نیازمند مدل‌سازی‌های ریاضی و شبیه‌سازی‌های رایانه‌ای هستند، کمتر استفاده شده است.

زیست‌شناسی تکاملی با تاریخ تکامل موجودات زنده سروکار دارد. از نیمهٔ دوم قرن نوزدهم به بعد، به‌تدریج پذیرفته شد که موجودات زنده طی فرایند انتخاب طبیعی، از نیاکان خود منشعب شده و گونه‌های جدیدی پدید آمده‌اند. درخت‌های فیلوژنتیک طبق نظریهٔ تکاملی داروین، رابطهٔ 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، نتیجهٔ 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 مجلهٔ *نیچر*[1] 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 دربارهٔ تاریخ انسان فراهم می‌آورند، با استفاده از روش‌های محاسباتی استخراج‌شده از زیست‌شناسی تکاملی، به تحلیل داده‌های زبان‌شناسی پرداخته‌اند. تحلیل آنها بر اساس ماتریسی از ۸۷ زبان دنیا و با ۲٤٤۹ 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 رسم‌شده در این مقاله، نشان داده شده است که زبان‌های پرتغالی و اسپانیایی رابطهٔ خویشاوندی نزدیک‌تری دارند تا زبان انگلیسی (زبان انگلیسی رابطهٔ 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 آگوست سال ۲۰۱۲، رامکو بوکائرت و همکارانش مقاله‌ای (بوکائرت و همکاران ۲۰۱۲) جنجالی با عنوان «ترسیم منشأ و توسعهٔ خانوادهٔ زبان هندواروپایی» در مجلهٔ معتبر *ساینس*[2] 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 رّد نتیجهٔ این مقاله پرداخته شد. آنها در این مقاله، با نگرش تکاملی تبارشناسی جغرافیاییِ بیزی[3] و با استفاده از ۲۰۰ واژه از ۱۰۳ زبان باستانی و معاصر هندواروپایی، توسعهٔ زبان‌ها در مناطق مختلف زمین را الگو قرار دادند. آنها دو فرضیهٔ مهم و مقابل هم دربارهٔ 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 ٦۰۰۰ سال پیش و در منطقهٔ استپ روسیه، در شمال دریای خزر باز می‌گردد. در مقابل این فرضیه، فرضیهٔ **آناتولیا** قرار دارد که پروفسور کالین رنفرو[4] برای اولین‌بار در اواخر دههٔ ۱۹۸۰ میلادی مطرح کرد و خاستگاه «نیا-زبان» این زبان‌ها را منطقهٔ آناتولی ترکیه و تاریخ آن

1) *Nature* 2) *Science* 3) bayesian phylogeography 4) Colin Renfrew

را ۸۰۰۰ تا ۹۵۰۰ سال پیش می‌داند. بوکائرت و همکارانش شواهدی یافته‌اند که فرضیهٔ 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– ۹٦۰).

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 ادامهٔ مقاله به معرفی گویش‌های مورد مطالعه و روش رسم درخت خواهیم پرداخت.

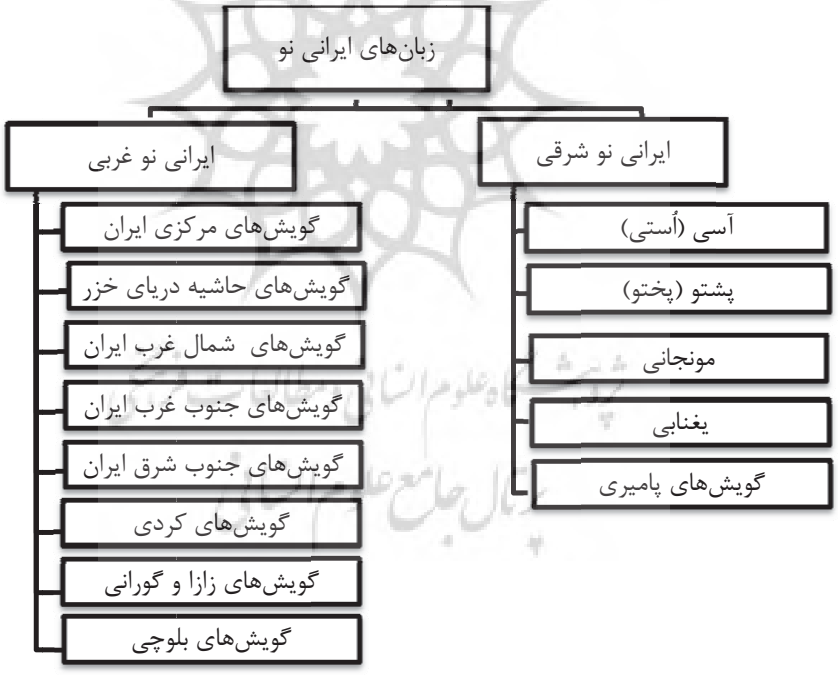
زبان‌های ایرانی

زبان‌ها و لهجه‌های ایرانی خانوادهٔ واحدی را تشکیل می‌دهند که از یک اصل مشترک منشعب شده‌اند؛ به عبارتی دیگر زبان‌ها و گویش‌ها از یک منشأ یا یک ریشهٔ زبانی واحد مشتق شده‌اند (نک: ارانسکی ۱۳۷۸: ۲٦–۲۸). زبان‌های ایرانی یکی از شاخه‌های زبان‌های هندوایرانی از خانوادهٔ 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 ریشهٔ 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 نحوهٔ استفاده از عبارات و ترکیب‌ها، امروزه خود به عنوان یک زبان شناخته می‌شوند (ویندفور ۲۰۰۹: ٤۱۸).

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 درسال ۱۳۸٤ 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 (٤۱–٦۰ 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 عمدهٔ **ایرانی نو شرقی** و **ایرانی نو غربی** جای می‌گیرند. هر کدام از این گروه‌ها، به چندین زیرگروه تقسیم می‌شوند که هریک شامل چندین گویش‌اند. این تقسیم‌بندی به‌صورت شماتیک در شکل (۱) نشان داده شده است. به عنوان مثال، گویش‌های مرکزی ایران اغلب در منطقهٔ میان اصفهان، تهران، همدان و یزد رواج دارند و به شش گروه (شمال غربی، شمال شرقی، جنوب غربی، جنوب شرقی، منطقهٔ تفرش، و دشت کویر) تقسیم می‌شود. گویش‌های گروه شمال شرقی در منطقهٔ 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 باغ بیدی 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[1] و ماتریس‌های داده‌های ریخت‌شناسی[2]، به بررسی ارتباط تکاملی گونه‌های مختلف جاندران می‌پردازد. روند تکاملی را می‌توان توسط یک درخت فیلوژنی مجسم کرد. معمول‌ترین روش‌ها برای استنباط فیلوژنی شامل صرفه‌جویی[3] (یا کمترین فرضیات)، درست‌نمایی حداکثری[4]، و استنتاج بِیزی[5] 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 اسیدآمینهٔ ترازشده[6] هستند. اطلاعات (داده‌های ورودی) و الگوریتمی که برای محاسبهٔ درخت انتخاب می‌شود، دو معیار اصلی برای ساخت درخت‌های فیلوژنی محسوب می‌شوند (بلومبرگ و همکاران ۲۰۰۳: ۷۱۹– ۷۲٤).

روش کار

سخن‌گویان هرزبانی، علاوه بر مجموعهٔ 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 هربخش دیگرِ زبان تغییر می‌کند. ما، در این نوشتار، ۲۰۰ واژه (معادل مجموعهٔ داده‌های سوادِش[7]، که مبنای کار رامکو بوکائرت و همکارانش بود و در مقدمه توضیح داده شد) را در ٤۹ گویش که به همت گروه زبان‌ها و گـویـش‌های ایـرانیِ فرهنـگستان 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) Swadesh

مجموعه‌کتاب‌هایی با عنوان **گنجینهٔ گویش‌های ایرانی** به چاپ رسیده، بررسی کردیم. این گویش‌ها عبارت‌اند از ۱- التپه ۲- رستم‌کلا ۳- مهدیرجه ٤- یخکش ٥- دیباج (چارده‌کلاته) ٦- پوروا ۷- نوکنده ۸- بنفشه‌تپه ۹- جَفاکنده ۱۰- طَرقی ۱۱- طاری ۱۲- کِشه‌ای ۱۳- طامه‌ای ١٤- نطنزی ١٥- تکیه‌ای ١٦- پاوه‌ای ۱۷- نودشه ۱۸- شرکان ۱۹- دزلی ۲۰- بزلانه ۲۱- کَلهُری ۲۲- گورانی ۲۳- سَنجابی ٢٤- کولیایی ٢٥- زَنگنه‌ای ٢٦- جَلالوَندی ۲۷- زوله‌ای ۲۸- کاکاوَندی ۲۹- هوزمانوَندی ۳۰- پاچهلک ۳۱- لاسگردی ۳۲- دوانی ۳۳- دهله‌ای ٣٤- عبدویی ۳٥- کازرونی ٣٦- کلانی ۳۷- کنده‌ای ۳۸- کوزرگی ۳۹- ممسنی ٤۰- ماسرمی (دهسروی) ٤۱- بلیانی ٤۲- بیروکانی ٤۳- حیاتی (دولت‌آبادی) ٤٤- دادنجانی ٤٥- لردارنگانی ٤٦- درونکی (مهبودی) ٤۷- دژگاهی (گوری) ٤۸- کرشی ٤۹- کرونی.

در انتخاب مجموعهٔ داده‌ها، با محدودیت‌های بسیاری مواجه بودیم. کتاب‌ها و تحقیقات متعددی به گردآوری مجموعهٔ 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 مجموعهٔ 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 مجموعهٔ بزرگ‌تری از گویش‌ها پیاده‌سازی کنیم.

گنجینهٔ گویش‌های ایرانی، که پیش‌تر به آن اشاره کردیم، شامل حدود ۲٥۰۰ واژه و ۱۰۰ جمله از گویش‌های اصیل ایرانی است که ما با توجه به نیازمان، ۲۰۰ واژه (معادل مجموعهٔ واژه‌های سوادِش) را انتخاب کرده‌ایم. مجموعهٔ واژه‌های سوادِش، دو فهرست ۱۰۰ و ۲۰۰ واژه‌ای است[1] که زبان‌شناس آمریکایی، موریس سوادش در طول سال‌های

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

۱۹٤۰ تا ۱۹۵۰، آنها را با هدف بررسی ارتباط زبان‌ها با روش آماروواژگانی[1]، تهیه کرده است، فهرست ۲۰۰ واژه‌ای شامل واژه‌هایی است که مختص زبان خاصی نیستند و در تمام زبان‌ها وجود دارند. سوادِش واژه‌ها را به صورت شهودی انتخاب کرده است.[2] 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، خویشاوندی تاریخی و...) در تفکیک آن دو تأثیر دارند؛ بنابراین تشخیص زبان از گویش دشوار است (مدرسی ۱۳۸٤: ۱۳٤). به همین‌دلیل و برای سادگی در بحث، در ادامه، ما همواره از واژهٔ **زبان** استفاده خواهیم کرد و به بررسی ٤۹ زبان خواهیم پرداخت.

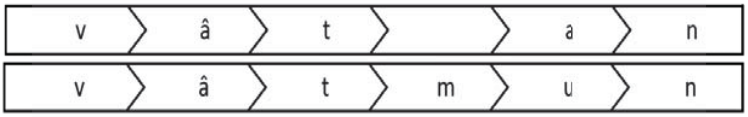
برای محاسبهٔ فاصلهٔ میان زبان‌ها از فاصلهٔ لِوِنشتاین[3] استفاده کرده‌ایم. در این روش، فاصلهٔ دو زبان از طریق جمع جبری فاصلهٔ میان واژه‌های معادل در آن دو زبان به دست می‌آید. اگر فاصلهٔ میان دو زبان (گویش) $l1$ و $l2$ را با $distance\ (l1,l2)$ و فاصله دو واژهٔ w_1^k و w_2^k (که واژه‌های معادل k- امین واژهٔ مجموعهٔ سوادِش 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 محاسبهٔ فاصلهٔ دو واژه در روش لِوِنشتاین، ابتدا باید دو واژه را با یکدیگر تراز کرد و سپس تعداد اختلاف‌های

1) lexicostatistics
2) http://en.wikipedia.org/wiki/Swadesh_list
3) levenshtein distance

آنها را به دست آورد. به عنوان مثال، دو واژهٔ vâtan و vâtmun (به معنای «گفتن»، در کشه‌ای و طامه‌ای) به صورت زیر تراز می‌شوند:



در صورتی که به حذف و تفاوت امتیاز ۱ بدهیم، فاصلهٔ این دو واژه برابر با ۲ خواهد شد؛ چون دارای یک حذف و یک تفاوت است. البته نحوهٔ 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 محاسبهٔ ماتریس فاصله‌ها (ماتریس متقارن ٤٩×٤٩، که عنصر m_{ij} بیانگر فاصلهٔ زبان **i** ام و زبان **j** ام است) با الگوریتم BioNJ (بهبودیافتهٔ الگوریتم نزدیک‌ترین مجاور[1])، درخت فیلوژنی مربوط به زبان‌ها بازسازی شد؛ سپس با استفاده از نرم‌افزار رسم درختِ (FigTree)، درخت به‌دست‌آمده 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:

ـــ در برخی از گویش‌ها، برای مفهومی خاص، چندین واژه به کار برده می‌شود؛ به عنوان مثال، در گویش بزلانه برای مفهوم «سنگین» واژه‌های sañǧīn | sañǧīnaqurs | qursa کاربرد دارند. برای اندازه‌گیری فاصلهٔ میان گویش‌ها، لازم بود یکی از واژه‌ها را انتخاب کنیم و محاسبهٔ فاصلهٔ 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 بقیهٔ 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

1) neighbor-joining

«شپش» در زبان‌های دوانی، دهله‌ای، عبدویی، کازرونی، کلانی، کنده‌ای در جدول (۱) نمایش داده شده است. با توجه به ٤٨ 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 واژهٔ šeš انتخاب و واژهٔ gendâr کنار گذاشته شد.

جدول۱. واژه‌های معادل «شپش» در شش زبان استان فارس

دوانی	دهله‌ای	عبدویی	کازرونی	کلانی	کنده‌ای
šiš	siš	seš	siš	šeš	šeš, gendâr

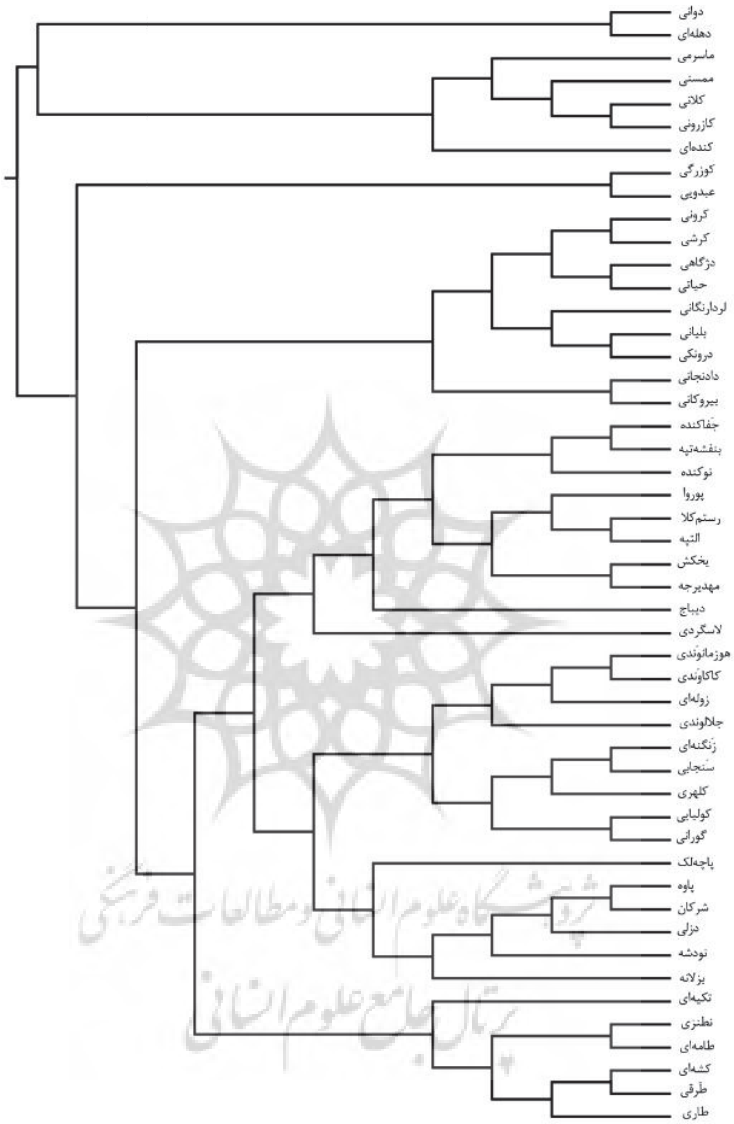
ــ چالش دیگر نداشتن برخی از واژه‌ها بود. واژه‌های «چه» و «فکر کردن» را به دلیل موجود نبودن در مجموعهٔ داده‌ها، از انتخاب خود کنار گذاشتیم و به‌جای ۲۰۰ واژه، ۱۹۸واژه را به کار بردیم.

ــ 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.

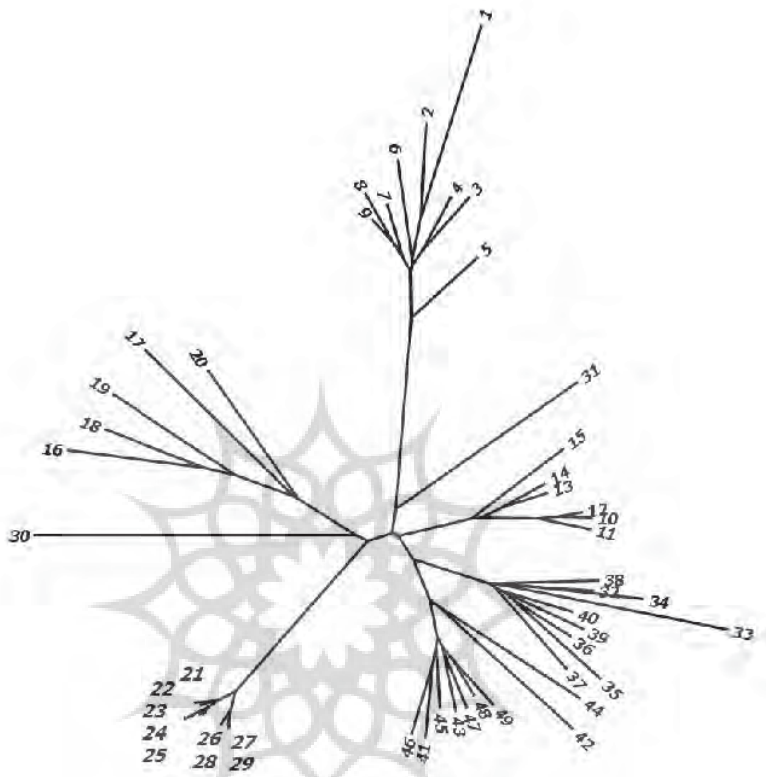
نتیجه‌گیری

درخت به‌دست‌آمده مطابقت معناداری با جغرافیای زبان‌ها دارد. در شکل (٤) دسته‌بندی زبان‌ها با شمارهٔ آنها بر روی نقشهٔ 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 (شمارهٔ ۳۱)، که در استان سمنان رایج است، از این زبان‌ها فاصلهٔ کمتری دارد.

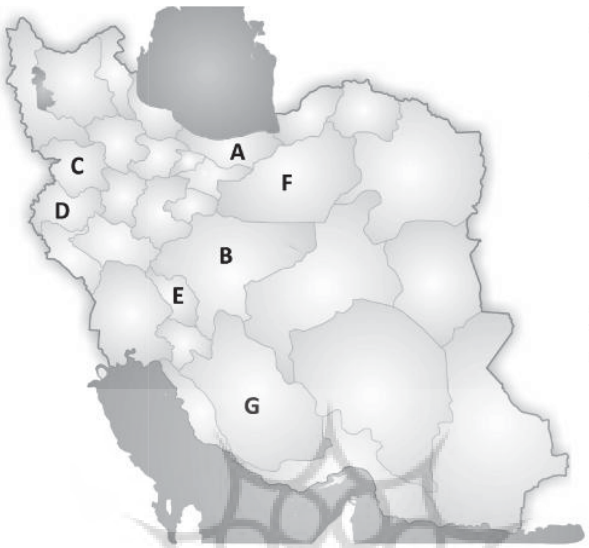
همان‌طور که در شکل (۳) می‌توان مشاهده کرد، خوشهٔ مربوط به زبان‌های شمارهٔ ۲۱ تا ۲۹ تراکم بیشتری دارد و زبان‌های این خوشه فاصلهٔ بسیار کمی از یکدیگر دارند. درواقع، می‌توان این‌گونه استنباط کرد که زبان‌های شمارهٔ ۲۱ تا ۲۹، نسبت به دیگر زبان‌های مورد مطالعه، کمتر دستخوش تغییرات شده‌اند.



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



شکل ۳. درخت خویشاوندی رسم‌شده برای ٤٩ زبان ایرانی با روش BioNJ (طرح‌بندی درخت شعاعی). شمارهٔ رأس‌ها مطابق با شمارهٔ ارائه‌شده برای زبان‌ها در ابتدای بخش «روش کار» است.



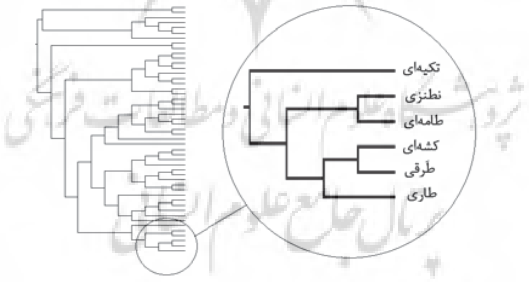
A	التپه، رستم‌کلا، مهدیرجه، یخکش، دیباج، پوروا، نوکنده، بنفشه‌تپه، جفاکنده
B	طرقی، طاری، کشه‌ای، طامه‌ای، نطنزی، تکیه‌ای
C	پاوه، نودشه، شرکان، دزلی، یزلانه
D	کلهری، گورانی، سنجابی، کولیایی، زنگنه‌ای، جلالوندی، زوله‌ای، کاکاوندی، هوزمانوندی
E	پاچه‌لک
F	لاسگردی
G	دوانی، دهله‌ای، عبدوبی، کازرونی، کلاتی، کنده‌ای، کوزرگی، ممسنی، ماسرمی، بلیانی، بیروکانی، حیاتی، دادنجانی، لردارنگانی، درونکی، دژگاهی، کرشی، کرونی

شکل ٤. نقشهٔ 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 طَرقی، 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 (٦) نمایش داده شده است، زبان‌های طرقی 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 حدّفاصل 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 نتیجهٔ 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 فاصلهٔ بیشتری دارد که کاملاً با فاصلهٔ روی نقشه و مسیر ارتباطی آنها مطابقت دارد.

شکل ۵. نقشهٔ بخشی از استان اصفهان، مسیر میان شهرهای نطنز، طامه، طرق، کشه و طار با استفاده از نقشهٔ گوگل. مسیر میان این شهرها، با شکل درخت خویشاوندی زبان‌های این شهرها، همخوانی معناداری دارد.



شکل ٦. بخشی از درخت خویشاوندی زبان‌های تکیه‌ای، نطنزی، طامه‌ای، کشه‌ای، طرقی، طاری (درخت خویشاوندی تمامی ٤٩ 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 فاصلهٔ میان شهرها بیانگر همین نکته است. اما نکتهٔ جالب توجه زبان‌های عبدویی، کلانی و ممسنی است. با اینکه عبدویی و کلانی نسبت به نورآباد (ممسنی) فاصلهٔ 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 رابطهٔ خویشاوندی نزدیکی 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 کمّی این درخت‌ها می‌تواند در بررسی خویشاوندی دور یا نزدیک زبان‌ها و چگونگی انشعاب‌ها و در صورت مطالعهٔ دقیق‌تر، تخمینی از زمان جدایی زبان‌ها، به کار آید و اطلاعات باارزشی دربارهٔ 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 ارائهٔ این نوشتار تنها معرفی روش کار و ارائهٔ نتیجهٔ حاصل از پیاده‌سازی الگوریتم‌های ساخت درخت فیلوژنی بود. آنچه مسلّم 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 رابطهٔ خویشاوندی نزدیک‌تری دارند و گویش دوانی در شاخهٔ دورتری جای گرفته است.

این ناهمخوانی‌ها را می‌توان به دو دسته تقسیم کرد. دستهٔ نخستِ 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 مقولهٔ زبان‌شناسی نگریسته شود و محققان زبان‌شناس با دانش بیشتر و کار تخصصی‌تر به بررسی این موارد بپردازند. دستهٔ دوم خطاهای مربوط به داده‌های ورودی برنامهٔ رایانه‌ای ما است ــ که باتوجه به حجم بالای داده‌ها، امری محتمل است ــ و همچنین خطاهایی است که به دلیل نبود دانش کافی در مدل‌سازی دقیق به وجود آمده است، که امیدواریم با مطالعهٔ بیشتر و در تحقیقات آتی، برطرف شوند.

منابع و کتابنامه

آذرلی، غلامرضا، ۱۳۸۷، *فرهنگ واژگان گویش‌های ایرانی*، تهران.
ارانسکی، یوسیف، ۱۳۷۸، *زبان‌های ایرانی*، ترجمهٔ 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، *واژه‌نامهٔ شصت‌وهفت گویش ایرانی*، تهران.
محمدی املشی، نصرالله‌پور و غلامی، حسین، ۱۳۹۱، *فرهنگ واژگان تاتی شال*، تهران.
مدرسی، یحیی، ۱۳۸٤، *درآمدی بر جامعه‌شناسی زبان*، تهران.

Blomberg, 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.

Bouckaert, R. et al., 2012, "Mapping the Origins and Expansion of the Indo-European Language Family", *Science*, vol. 337, pp. 957-960.

Gray, R. and Atkinson, Q., 2003, "Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin", *Nature*, vol. 426, pp. 435-439.

Windfuhr, G. and Perry, J. R., 2009, "Persian and Tajik", *The Iranian Languages*, ed. G. Windfuhr, Landon. pp. 416-544.

پیوست:

۲۰۰ واژهٔ مجموعهٔ سوادِش

فارسی	انگلیسی	فارسی	انگلیسی
همه	all	نفس کشیدن	to breath
و (واو عطف)	and	سوختن	to burn (intransitive)
حیوان	animal	بچه	child (young)
خاکستر اجاق	ashes	ابر	cloud
پشت	back	سرد	cold (weather)
بد	bad	آمدن	to come
پوست	bark (of a tree)	شمردن	to count
چون	because	بریدن	to cut
شکم	belly	روز	day (not night)
بزرگ	big	مردن	to die
پرنده	bird	کندنِ زمین	to dig
گاز گرفتن	to bite	کثیف	dirty
سیاه	black	سگ	dog
خون	blood	نوشیدن	to drink
وزیدن	to blow (wind)	خشک	dry (substance)
استخوان	bone	کند	dull (knife)
خاک	dust	چهار	four
گوش	ear	یخ زدن	to freeze
زمین	earth (soil)	میوه	fruit
خوردن	to eat	دادن	to give
تخم مرغ	egg	خوب	good

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

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

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

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

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

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

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

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