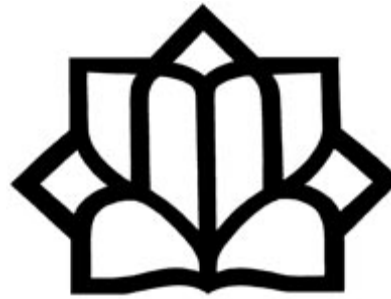


بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه کاشان

دانشکده مهندسی

گروه مهندسی کامپیوتر

سیستم هوشمند تشخیص وزن عروضی اشعار فارسی

محمد مهدی مجیری ۸۲۱۳۰۰۱۹

استاد راهنما:

جناب آقای دکتر بهروز مینایی

مرداد ماه ۱۳۸۷

ساکزاری

معمولا دانشجویان برای پر کردن صفحات و یا بر حسب عادت در این صفحه که مخصوص قدردانی می باشد، چیزی می نویسند. ولی قصد من این است که در این مجال کوتاه، واقعا از کمک بسیار بسیاری افراد، تشکری - هرچند ناچیز - بنمایم. و در همین ابتدای کار و بلکه تا انتهای کار باید اعتراف کنم که انجام این پژوهش به تنهایی امکان پذیر نبوده و ادامه آن نیز به تنهایی امکان پذیر نمی باشد.

از جناب آقای دکتر مینایی که هم مشوق بنده در انجام این پروژه بوده و هم از هیچ کمکی دریغ نورزیدند، از خانم سلیمیان که از ابتدای پروژه تا انتها از هیچ کمکی در زمینه ادبیات دریغ نکرده و با راهنمایی های خود در پیشرفت کار سهم به سزایی داشتند، از خانم دهنمکی که برای ساخت مجموعه داده، زمان بسیاری را در اختیار ما قرار دادند، که انصافا اگر کمک ایشان نبود، مطمئنا برای آزمون نهایی برنامه با مشکل برمی خوردیم. از دوست عزیز و گرامی محمد نحوی که برای آزمون برنامه به وسیله شبکه عصبی و مقایسه نتایج، کمک شایانی نمودند. از جناب آقای دکتر اکبری استاد ادبیات دانشگاه تهران که هرچند نتوانستیم از ایشان به عنوان مشاور کامل پروژه کمک بگیریم، ولی کمک و مشاوره ایشان برای شروع پروژه و ایجاد انگیزه قوی را نمی توان فراموش کرد، از آقایان حمید رضا انواری و مجتبی جمالی به خاطر کمک های فراوان برای برنامه نویسی PHP و رفع مشکلات پیاده سازی، از دوست عزیز و گرامی، مهدی نقیبی که با راهنمایی های خود راه را برای انجام پروژه هموار نمودند، و همچنین از آقایان مهدی نصیری، امید کاشفی، اکبر ثقفیان، مسعود میرسالاری، استاد حسین آهی، احمد رئیسی، محمد حسن فردوسی زاده و دیگر دوستانی که شاید نام آنها از قلم افتاده باشد. از برادرم و خواهران گرامی ام که وقت زیادی را برای انجام این پروژه در اختیار بنده قرار دادند.

باید از همه ی این عزیزان تشکر کنم که با کمک های کوچک و بزرگ خود، مرا در انجام این پروژه یاری رساندند. و البته عذرخواهی ویژه که نمی توانم در این مجال، کمک و یاری تک تک آنها را بنویسم. ان شاء الله خداوند اجر تمامی آنان را خواهد داد.

این اثر را هر چند بسیار کوچک است

تقدیم می‌نمایم به ایران اسلامی

فهرست

خلاصه	۱ -
مقدمه	۳ -
۱-۱- دلایل تحقیق بر روی عروض	۴ -
۱-۲- این سیستم چه کاری نمی تواند انجام دهد؟	۴ -
۱-۳- این سیستم چه کاری می تواند انجام دهد؟	۵ -
۱-۴- ساختار این نوشتار	۵ -
۲- عروض	۸ -
۲-۱- حرف، هجا، شعر، وزن	۸ -
۲-۱-۱- مصوت و صامت های زبان فارسی	۸ -
۲-۱-۲- هجا یا سیلاب یا بخش	۸ -
۲-۱-۳- شعر	۹ -
۲-۱-۴- وزن عروضی	۹ -
۲-۲- طریقه به دست آوردن وزن یک شعر به صورت دستی	۱۰ -
۲-۲-۱- درست خواندن	۱۰ -
۲-۲-۲- تولید رشته صامت-مصوت	۱۱ -
۲-۲-۳- تقطیع	۱۲ -
۲-۲-۴- رکن بندی	۱۳ -
۲-۲-۵- اوزان اشعار فارسی	۱۳ -
۲-۲-۶- اختیارات شاعری	۱۴ -
۲-۳- تاریخچه عروض	۱۶ -
۲-۴- در این فصل ...	۱۷ -
۳- پیش نیازهای طرح	۱۹ -
۳-۱- استاندارد خط	۱۹ -
۳-۱-۱- نخستین کارها	۱۹ -
۳-۱-۲- کد CP1256	۲۰ -

- ۲۰ - یونی کد UNICODE ۳-۱-۳
- ۲۱ - UTF-8 کد با طول متغیر ۳-۱-۴
- ۲۲ - راه حل برای مشکل استاندارد خط ۳-۱-۵
- ۲۲ - حذف نویسه‌های اضافی ۳-۲-۲
- ۲۳ - نویسه خط تیره ۳-۲-۱
- ۲۳ - فاصله مجازی و اتصال مجازی ۳-۲-۲
- ۲۳ - ساخت Data Set ۳-۳
- ۲۴ - زبان برنامه نویسی ۳-۴
- ۲۵ - در این فصل... ۳-۵
- ۲۷ - الگوریتم سیستم پیشنهادی ۴-۱
- ۲۷ - تعریف مساله ۴-۱-۱
- ۲۸ - مراحل پروژه ۴-۲
- ۲۹ - حرکت گذاری ۴-۲-۱
- ۳۰ - تبدیل مصراع حرکت گذاری شده به رشته صامت مصوت (CV) ۴-۲-۲
- ۳۱ - اصلاح رشته CV بر اساس اختیارات شاعری ۴-۲-۳
- ۳۲ - تبدیل رشته CV اصلاح شده به رشته udash ۴-۲-۴
- ۳۲ - اصلاح رشته udash بر اساس اختیارات شاعری ۴-۲-۵
- ۳۳ - انطباق یکی از اوزان معروف با رشته udash ۴-۲-۶
- ۳۳ - در این فصل... ۴-۳
- ۳۵ - انطباق ۵-۱
- ۳۵ - الگوریتم‌های انطباق رشته ۵-۱-۱
- ۳۶ - Levenshtein ۵-۱-۱
- ۳۷ - Levenshtein Distance مثالی از ۵-۱-۲
- ۳۸ - الگوریتم‌های انطباق رشته مورد استفاده ۵-۲
- ۳۸ - الگوریتم XOR ۵-۲-۱
- ۴۰ - XOR برای اعمال اختیار شاعری قلب ۵-۲-۲
- ۴۰ - XOR برای اعمال اختیار شاعری تسکین ۵-۲-۳
- ۴۱ - ترکیب نتایج توابع برازندگی ۵-۳
- ۴۱ - اولویت ۵-۳-۱
- ۴۲ - میانگین وزنی ۵-۳-۲

- ۵-۳-۳- به دست آوردن وزن هر برازندگی - ۴۲ -
- ۵-۴- در این فصل ... - ۴۳ -
- ۶- آزمون - ۴۶ -
- ۶-۱- نتیجه نهایی به تفکیک وزن - ۴۶ -
- ۶-۱-۱- تحلیل آمار به دست آمده - ۴۸ -
- ۶-۱-۲- پیشنهادات بر اساس آمار به دست آمده - ۴۸ -
- ۶-۲- مقایسه اوزان با یکدیگر - ۴۹ -
- ۶-۲-۱- اوزان نزدیک به هم - ۴۹ -
- ۶-۲-۲- مقایسه اوزان با نزدیکترین وزن - ۵۰ -
- ۶-۳- در این فصل ... - ۵۱ -
- ۷- نتایج - ۵۳ -
- ۷-۱- عروض - ۵۳ -
- ۷-۲- اهمیت و لزوم وجود این سیستم - ۵۴ -
- ۷-۳- پیشنهادات - ۵۶ -
- ۸- پیوست ها - ۵۹ -
- ۸-۱- آشنایی با طبقه بندی زبان های جهان و زبان اوستایی - ۵۹ -
- ۸-۲- جدول ارکان مشهور - ۶۲ -
- ۸-۳- جدول اوزان پر کاربرد - ۶۳ -
- ۸-۴- اوزان مثنوی های مشهور - ۶۵ -
- ۸-۵- نویسه های فارسی و هماندها - ۶۸ -
- ۸-۵-۱- صفحه کلید فارسی - ۶۸ -
- ۸-۵-۲- نویسه های فارسی و یونی کد آنها - ۶۸ -
- ۸-۵-۳- نویسه های عربی و همانند آن در فارسی - ۷۰ -
- ۸-۵-۴- فاصله مجازی - ۷۰ -
- ۸-۶- توضیحات مربوط به Data Set تولید شده - ۷۲ -
- ۹- واژه نامه - ۷۵ -
- ۱۰- منابع - ۸۳ -

خلاصه

سابقه تاریخی شعر و ادب فارسی و همچنین دست‌یازی‌های بیگانگان به سوابق تاریخی این عرصه، لزوم ورود تحقیقات کامپیوتری و بین رشته‌ای در این زمینه را طلب می‌نماید. یکی از مشکل‌ترین بخش‌های رشته‌ی ادبیات فارسی تشخیص وزن عروضی است، به صورتی که تشخیص وزن عروضی در بعضی اشعار برای اساتید ادبیات هم مشکل‌زا می‌شود. با توجه به این‌که تاکنون هیچ پژوهشی در این زمینه انجام نشده، ورود به این عرصه ضروری به نظر می‌رسد. ما در این پژوهش ابتدا روش و شیوه‌های اخذ دانش و ایجاد پایگاه دانش قوانین وزن عروضی شعر فارسی را بیان می‌نماییم. سپس الگوهای ۳۱ گانه وزن عروضی تبیین می‌گردد. برای سنجش توانایی برنامه، یک مجموعه هزار بیتی از اشعار حافظ و مولوی جمع‌آوری و آماده‌سازی شده است. در این پژوهش، پس از پنج مرحله پردازش ابتدایی مانند حرکت‌گذاری، تبدیل به رشته صامت - مصوت، اعمال اختیارات و ضرورت‌های شاعری و بر روی بیت شعر ورودی، بیت ورودی تبدیل به رشته‌ای از «U» و «-» خواهد شد که «U» در این رشته نشان‌دهنده هجای کوتاه و «-» نشان‌دهنده هجای بلند می‌باشند. سپس این رشته ورودی با استفاده از الگوریتم‌های یافتن شباهت دو رشته مانند Levenshtein Distance، شعر ورودی به شبیه‌ترین ۳۱ وزن مشهور فارسی نسبت داده می‌شود. نتایج اولیه به دست آمده صحت بیش از ۶۵٪ را نشان می‌دهد.

فصل اول

مقدمه

۱- مقدمه

ادبیات فارسی، زمینی بکر است که به جهات مختلف، کسی جز در موارد نادر حاضر به ورود به این عرصه و زدن حرف نو نیست. برخی بر این باورند که آنچه گذشتگان کرده‌اند فصل الخطاب کاری است که می‌توان کرد و برخی از آن می‌ترسند که مورد طعن دیگران قرار گیرند. دیگر همه جریان آمدن شعر نو و هجمه انتقادات به نیما را می‌دانیم. و شبیه همین ماجرا در علم عروض است. شاید اگر امثال دکتر خانلری و مسعود فرزاد هم این عقاید را داشتند، ما هنوز در حال حفظ کردن زحافات و درجا زدن در علم عروض بودیم. زمان زیادی از به وجود آمدن عروض علمی به صورت امروزی نمی‌گذرد (کمتر از ۵۰ سال) ولی هم اکنون شاهدیم که کمتر استاد ادبیاتی عروض را فقط به شیوه‌ی قدیم آموزش می‌دهد. قبل از رایج شدن عروض علمی، عروض فنی غریب بود که کمتر کسی از اهل ادب با معیارهای آن آشنایی داشت و لذا آنان که عروض می‌دانستند در هاله‌ای از اسرار به نظر می‌رسیدند [۱].

هر چه قدر حرف نو زدن در ادبیات مشکل است، ورود کامپیوتر به این عرصه از آن مشکل‌تر. با وجود پیشرفت‌های فراوان در تحقیقات کامپیوتری، (به صورتی که محققان در برخی زبان‌ها همچون انگلیسی و اسپانیایی موفق به سرایش شعر به وسیله کامپیوتر شده‌اند [۱۲-۱۵]) هنوز اساتید ادبیات و ادیبان ما حاضر به قبول توانایی‌های این ابزار نیستند. و از طرف دیگر اساتید کامپیوتر نه اطلاع چندانی با مشکلات ادبیات دارند و نه حاضرند در این زمین بکر کاری نو بکنند. اکثر کارهایی که در ارتباط با زبان فارسی و رایانه انجام شده مربوط به مباحث زبان فارسی است.^۱ معدود کارهایی که در زمینه ادبیات و رایانه انجام می‌گیرد، از کامپیوتر در حد یک ویرایشگر استفاده می‌شود. مانند برنامه‌هایی که در دیوان شعرا جستجو می‌کنند و یا فال حافظ می‌گیرند. از جسورانه‌ترین کارهایی که در این زمینه انجام شده می‌توان به [۲] اشاره کرد، که آن هم در مقایسه با کارهایی که در زبان‌های دیگر انجام شده، بسیار کوچک می‌نماید.

یکی از مشکل‌ترین بخش‌های رشته‌ی ادبیات فارسی تشخیص وزن عروضی است، به صورتی که تشخیص وزن عروضی در بعضی اشعار برای اساتید ادبیات هم مشکل‌زا می‌شود. با توجه به این‌که

^۱ برای اطلاع از این موضوع می‌توانید به مقالات اولین و دومین کارگاه پژوهشی «زبان فارسی و رایانه» مراجعه نمایید.

تاکنون هیچ پژوهشی در این زمینه انجام نشده، ورود به این عرصه ضروری به نظر می‌رسد. البته باید توجه کرد که عمر عروض علمی که ما در این پژوهش از آن استفاده کرده‌ایم به ۵۰ سال هم نمی‌رسد.

۱-۱- دلایل تحقیق بر روی عروض

عروض برای شناختن اجزاء شعر و تشخیص وزن صحیح از وزن شکسته و ناصحیح است و یادگیری آن هرگز برای به دست آوردن توانایی لازم برای سرودن اشعار نیست [۴]. وجود وزن برای شعر لازم است و آشنایی شاعران و محققان ادبی با علم عروض واجبتر. امروزه عروض جزو یکی از نظام‌های زبان شناسی محسوب می‌شود و در مطالعات مربوط به فنولوژی (واج شناسی) و سبک‌شناسی و نظام‌های دیگر ادبی کاربرد وسیع یافته است [۱]. آشنایی با عروض و به دست آوردن وزن عروضی یکی از روش‌های تحقیق بر روی اشعار شعرا، مقتضیات زمان شاعر، روحیات شاعر و ... می‌باشد.

به دلیل آنکه تشخیص وزن و امور مربوط به آن برای بعضی اساتید هم مشکل می‌نماید و همچنین نیاز مبرمی که به این علم وجود دارد، بر آن شدیم سیستمی طراحی کنیم که راه را بر دانش‌پژوهان آسان‌تر نماید و همچنین سرعت تحقیقات را بالاتر ببرد. زیرا هرچند استادی در این زمینه بسیار قوی باشد، نمی‌تواند به سرعت یک کامپیوتر، وزن یک مجموعه بزرگ را تشخیص دهد. البته توضیح خواهیم داد که این پژوهش در ابتدای کار است و نیاز به ادامه آن توسط پژوهشگران دیگر دارد.

۱-۲- این سیستم چه کاری نمی‌تواند انجام دهد؟

سیستم طراحی شده تنها قادر به تشخیص ۳۱ وزن معروفی است که از [۱] استخراج شده‌اند. البته باید توجه داشت که [۱۶] تعداد اوزان پرکاربرد را ۲۹ عدد دانسته و [۵] آنها را ۳۳ تا دانسته که البته خود اعتراف می‌کند که این اوزان اضافی بسیار کم کاربردتر از دیگران هستند. تحقیقات نشان می‌دهد که ۹۹ درصد اشعار در همین ۲۹ وزن سروده شده‌اند [۳]. و البته به اعتقاد [۱] قریب ۹۰ درصد و بلکه بیشتر اشعار فارسی به ۳۱ وزن اول سروده شده‌اند. با این توضیحات مشخص می‌شود که به دست آوردن همین ۳۱ وزن از مجموع بیش از ۳۰۰ وزن شناخته شده فارسی، کاری شایسته باشد. البته برای ادامه کار می‌توان تمامی این ۳۰۰ وزن را به مجموعه اضافه کرد. اگر وزنی در مجموعه ۳۱ تایی نباشد، سیستم سعی می‌کند آن را به یکی از این اوزان نسبت دهد، بنابراین وزن اشتباه تشخیص داده می‌شود و هیچگونه گزارشی هم، جز احتمالاً برازندگی پایین نمی‌دهد.

همچنین این سیستم قادر به تشخیص درستی و یا نادرستی وزن نیست، یعنی اگر شاعری بیتی را با وزنی اشتباه و یا ناهماهنگ سروده باشد، اگر اختلاف جزئی باشد، سیستم آن را تشخیص نخواهد داد. که این مورد هم می‌تواند زمینه تحقیقات بعدی باشد.

در میان اشعار فارسی، به دلیل قواعد عروضی خاص رباعی، این سیستم وزن این نوع شعر را به خوبی تشخیص نمی‌دهد. همچنین این مشکل برای وزن شعر نو نیز وجود دارد.

۱-۳- این سیستم چه کاری می‌تواند انجام دهد؟

این سیستم بیت ورودی را با تبدیلات میانی به رشته‌ای از «U» و «-» تبدیل می‌کند که «U» نشانه هجای کوتاه و «-» نشانه هجای بلند است. سپس سعی می‌کند این رشته را به یکی از ۳۱ رشته موجود در پایگاه داده - که همان ۳۱ وزن پرکاربرد هستند - نسبت دهد. برای این کار از الگوریتم‌های شباهت رشته^۱ استفاده می‌کند. به این صورت که رشته به دست آمده را با تمام ۳۱ رشته موجود مقایسه نموده و رشته‌ای را به عنوان وزن بیت معرفی می‌کند که بیشترین برازندگی^۲ را داشته باشد.

۱-۴- ساختار این نوشتار

به دلیل آنکه این پژوهش یک پژوهش میان‌رشته‌ای بوده و همچنین این گزارش برای پژوهشگران رایانه آماده شده، ناگزیر به ارائه توضیحاتی در مورد عروض بودیم. بنابراین در فصل دوم به صورت مختصر در مورد تاریخچه و روش به دست آوردن وزن یک شعر به صورت دستی توضیحاتی داده‌ایم. همچنین در پیوست اول هم توضیحاتی در مورد مبدأ زبان فارسی آورده شده است.

فصل سوم به تشریح پیش‌نیازهای طرح و همچنین دشواری‌های تحقیقات کامپیوتری در رابطه با زبان فارسی می‌پردازد. در فصل چهارم به تشریح روند کارکرد برنامه و الگوریتم‌های استفاده شده می‌پردازد. در فصل پنجم که انطباق نام دارد، الگوریتم‌های شباهت رشته و نحوه استفاده از آنها را بررسی می‌کند. فصل ششم آزمون‌های انجام شده بر روی برنامه و همچنین نتایج به دست آمده را

^۱ String Search(String Matching)

^۲ Fitness

بررسی می‌نماید. فصل هفتم به جمع‌بندی و شرح پیاده‌سازی انجام شده پرداخته و راه ادامه کار را روشن‌تر می‌نماید. پس از آن در فصل هشتم پیوست‌ها قرار گرفته است که در خلال توضیحات هر فصل به آنها ارجاع داده شده است.

در آخر برای راحتی پژوهش‌گران واژه‌نامه‌ای نوشته شده که در آن لغات تخصصی ادبیات به کار برده شده در متن توضیح داده می‌شود.

فصل دوم

عروض

۲- عروض

۲-۱- حرف، هجا، شعر، وزن

۲-۱-۱- مصوت و صامت‌های زبان فارسی

در فارسی ۶ مصوت داریم:

سه مصوت کوتاه: $\text{---}^{\text{ـ}}$

سه مصوت بلند: آ ای او

بقیه اصوات را صامت می‌گویند. زبان فارسی دارای ۲۳ صامت است. برخی صامت‌های زبان فارسی در خط دارای چند علامت هستند. مانند: «ز = ظ = ذ = ض».

در این پژوهش صامت را با «C» و مصوت کوتاه با «V» و مصوت بلند با «VV» نشان داده می‌شوند.

۲-۱-۲- هجا یا سیلاب یا بخش

هجای فارسی مجموعه‌ای از یک مصوت و یک یا دو یا سه صامت است. بنابراین تعداد هجاهای هر کلمه مساوی تعداد مصوت‌های آن است یعنی در یک عبارت همان تعداد هجاها هست که مصوت هست. هر هجا در یک دم یا بازدم ادا می‌شود. در کلمه «آمدم» سه مصوت وجود دارد. پس «آمدم» دارای سه هجاست. در زبان فارسی شش نوع هجا (مجموعه صامت و مصوت) وجود دارد، اما در عروض فارسی سه نوع آن مستعمل است که به آن‌ها هجاهای عروضی گویند. یعنی کلمات زبان فارسی در شعر فقط به سه نوع هجا بخش می‌شوند.

هجاهای عروضی از نظر کمیت بر دو گونه‌اند: بلند و کوتاه. در این پژوهش هجای کوتاه را با «-» و هجای بلند را با «U» نشان می‌دهیم.

هجاهای عروضی بلند عبارتند از:

صامت + مصوت کوتاه + صامت	یعنی CVC	مانند: شن، در، بن
صامت + مصوت بلند	یعنی CVV	مانند: با، بو، بی

هجای عروضی کوتاه عبارت است از:

صامت + مصوت کوتاه	یعنی CV	مانند: نه، که، چو
-------------------	---------	-------------------

۲-۱-۳- شعر

سخنی خیال انگیز که موزون و قافیه دار است. خواجه نصیرالدین طوسی وزن را به دلیل خیال انگیز بودن از فصول ذاتی شعر دانسته [۳]. مثال زیر که در [۳] نوشته شده، تفاوت سخن معمول با شعر را به خوبی بیان می‌دارد:

خورشید طلوع کرد	تنها یک سخن معمول و خبری است
گل خورشید شکفته شد	سخنی خیال انگیز ولی بدون وزن
گل خورشید شکفت	سخنی خیال انگیز و موزون

بی وزن کردن شعر، روح شعر را از بین می‌برد:

دانه چو طفلی است در آغوش خاک	روز و شب این طفل به نشو و نماست
دانه چو طفلی در آغوش خاک است	این طفل روز و شب به نشو و نماست

ناگفته نماند که شعر نو نیز موزون است و تنها تفاوت آشکار آن با شعر سنتی، کوتاهی و بلندی مصراع‌ها می‌باشد [۱]. البته اشعاری هم بعد از شعر نو با نام‌های سپید و موج نو به وجود آمد که وزن موجود در شعر نو هم در آنان وجود نداشت.

۲-۱-۴- وزن عروضی

وزن، نظم و تناسب خاصی است در اصوات شعر (=هجاءها) این نظم و تناسب اصوات به انحای گوناگون نزد ملل مختلف مبین نوعی آهنگ و موسیقی است [۱].

۲-۱-۴-۱- انواع وزن در زبان‌های گوناگون

در شعر سنتی هر زبانی، تساوی هجاهای هر مصراع، در وزن دخیل است. علاوه بر این عامل مشترک، وزن شعر هر زبانی مبتنی بر عامل خاصی است:

وزن عددی: وزن مبتنی است بر تساوی تعداد هجاهای هر مصراع (یعنی عامل خاصی در اینگونه وزن دخیل نیست) وزن اشعار فرانسوی و ایتالیایی و اسپانیایی از این گونه است.

وزن تکیه‌ای: وزن مبتنی بر تکیه که بر هجاها واقع می‌شود. وزن اشعار انگلیسی و آلمانی چنین است.

وزن کمی: وزن مبتنی بر امتداد زمانی، یعنی کمیت (کوتاهی و بلندی) هجاهاست. وزن شعر فارسی و عربی و سنسکریت و یونانی باستان و لاتین از این دست است.

وزن نواختی: وزن بر حسب زیری و بمی اصوات (هجاها) مشخص می‌شود. وزن شعر چینی و ویتنامی از این قبیل است.

۲-۴-۱-۲- مبانی وزن شعر فارسی

چنان که گفتیم وزن شعر فارسی مبتنی بر توالی‌های خاصی از هجاهای کوتاه و بلند در یک مصراع و تکرار عین آن در مصراع‌های دیگر است. بنابراین وزن شعر فارسی مجموعه‌ای از هجاهای کوتاه و بلند است که با نظم مشخصی پی در پی قرار گرفته باشد. هر نظم خاص، وزن خاصی را به ذهن متبادر می‌کند. پس: شعر فارسی به طور کلی مبتنی بر کمیت (کوتاه و بلندی) هجاهاست [۱].

۲-۲- طریقه به دست آوردن وزن یک شعر به صورت دستی

قواعد به دست آوردن وزن عروضی یک شعر را می‌توان در چند صفحه خلاصه کرد و یا در یکی دو ساعت به صورت کامل بیان کرد. ولی این بدان معنی نیست که پس از یادگیری قواعد عروض می‌توان به راحتی وزن عروضی یک شعر را به دست آورد. این امر نیاز به تمرین و ممارست فراوان دارد و مانند مسائل ریاضی اشعاری وجود دارد که پیدا کردن وزن عروضی آنها برای دانشجویان بسیار مشکل است.

علاوه بر مسائل فوق، اختلاف نظر هم بین اساتید وجود دارد، که پس از تحقیقات فراوان ما عروض علمی را انتخاب نمودیم که قابلیت الگوریتمیک کردن آن بسیار بهتر از عروض سنتی است. هم‌چنین بین اساتیدی که از این روش استفاده کرده‌اند، ما از روش دکتر سیروس شمیسا استفاده نموده‌ایم. البته در برخی از مراحل مجبور به استفاده از روش‌های دیگر شدیم، ولی اصل روش از کتاب ایشان [۱] استخراج شده است.

۲-۲-۱- درست خواندن

باید توجه داشت برای به دست آوردن وزن عروضی، صورت ملفوظ شعر برای ما مهم است، نه صورت مکتوب. به همین دلیل است که «ز»، «ض»، «ظ» و «ذ» برای ما تفاوتی ندارد. و در مقابل «و» در سه کلمه «سود»، «تو» و «وجد» برای ما متفاوت است، زیرا به سه صورت مختلف خوانده می‌شود.

در اولی به صورت یک مصوت بلند، در دومی به صورت یک مصوت کوتاه و یک صامت و در سومی به صورت یک صامت خوانده می‌شود.

برای درست خواندن دو روش پیشنهاد شده است: ۱- حرکت‌گذاری ۲- آوانویسی (نوشتن به خط فونتیک) به این دو روش املای عروضی گویند. البته نکات مختلفی هم در نوشتن املای عروضی پیشنهاد شده^۱ که می‌توان برای اطلاع از آنان به کتاب‌های عروض مراجعه کرد.

البته به این نکته هم باید توجه داشت که لازم نیست دقیقاً بیت را درست خواند، بلکه کافی است که بدانیم که ترتیب مصوت کوتاه و بلند و صامت در بیت چیست. مثلاً «حَسَن» با «حَسین» تفاوتی ندارد، زیرا در هر دو ترتیب صامت-مصوت به صورت ذیل است: صامت- مصوت کوتاه-صامت-مصوت کوتاه-صامت. ولی در عوض «حُسَن» با کلمات فوق متفاوت است، زیرا بر روی صامت دوم حرکتی نیست.

۲-۲-۲- تولید رشته صامت-مصوت

در روش شمیسا و کامیار این مرحله وجود ندارد و در دل مرحله قبل نهفته شده. در این بخش رشته صامت و مصوت به صورت مخفف‌های V, C تولید می‌شود. به این صورت که به جای صامت از C و به جای مصوت کوتاه یک V و به جای مصوت بلند دو V قرار می‌دهیم. به این صورت در پیدا کردن هجای کوتاه و بلند در مرحله بعد مشکلی پیدا نمی‌شود.

به عنوان مثال به تبدیل کلمه «توانا» (tavAnA)^۲ به رشته CV توجه کنید:

C	«ت» یک صامت است بنابراین به جای آن یک C قرار می‌دهیم.
CV	«-» یک مصوت کوتاه است بنابراین به جای آن یک V قرار می‌دهیم.
CVC	«و» به صورت صامت خوانده می‌شود، بنابراین به جای آن یک C قرار می‌دهیم.
CVCVV	«ا» یک مصوت بلند است، بنابراین به جای آن VV قرار می‌دهیم.
CVCVVC	«ن» یک صامت است بنابراین به جای آن یک C قرار می‌دهیم.
CVCVVCVV	«ا» یک مصوت بلند است، بنابراین به جای آن VV قرار می‌دهیم.

^۱ مثلاً شمیسا برای آنکه قانون سه واکی بودن هجای بلند رعایت شود، پیشنهاد داده بر سر مصوت بلند علامت مد قرار داده شود و یا از الف کوتاه و ضمه استفاده شود.

^۲ باید توجه داشت که این روش استفاده شده برای نوشتن تلفظ‌ها در این پژوهش است و روش استاندارد فونتیک نیست.

بنابراین رشته صامت-مصوت تولیدی برای کلمه «توانا» به صورت CVCVVCVV می‌باشد. البته فراموش نشود که برای به دست آوردن وزن صحیح باید کلمات در جمله تبدیل شوند و نه به صورت کلمه کلمه، یعنی کسره اضافی آخر کلمات هم محسوب شود.^۱

توجه: رشته‌های انگلیسی از چپ به راست نوشته شده. مثلاً «ت» که ابتدا صامت (C) و سپس مصوت (V) است، به صورت CV نوشته‌ایم و نه VC

۲-۲-۳- تقطیع

تقطیع در لغت به معنای قطعه قطعه کردن است و در اصطلاح عروض به معنای قطعه قطعه کردن شعر به هجای کوتاه و بلند است. در این سیستم هجای کوتاه را با «U» و هجای بلند را با «-» نشان می‌دهیم. خروجی این مرحله رشته Udash می‌باشد که بدون اعمال اختیارات شاعری، وزن اصلی شعر نیست، بلکه وزن تقطیعی است.

برای به دست آوردن این رشته، باید از ابتدای رشته CV شروع کنیم. در صورتی که به قالب CVV یا CVC رسیدیم به جای آن «-» قرار دهیم، در صورتی که به CV یا C تنها رسیدیم به جای آن «U» قرار دهیم. باید توجه داشت که در فارسی هیچ هجایی با مصوت (V) آغاز نمی‌شود، بنابراین نباید به گونه‌ای این تقسیم بندی را انجام داد که به V برسیم. در مثال زیر اگر ابتدای شعر (توا = CVCVV) را به صورت CVC جدا نموده بودیم، در بخش بعد مجبور بودیم از V شروع کنیم. بنابراین از ابتدای رشته CV را جدا کرده و پس از آن هم به دلیل گفته شده مجبوریم CVV جدا کنیم و نه CV.

تَوَانَا بُودَ هَرِ کِه دَانَا بُودَ										
تَ	وا	نا	بُ	وَد	هَر	کِه	دا	نا	بُ	وَد
CV	CVV	CVV	CV	CVC	CV	CVV	CVV	CV	CVC	CVC
u	-	-	u	-	-	-	-	-	u	-

در این مرحله هم مانند مرحله قبل باید مجموع کلمات با یکدیگر در نظر گرفته شوند.

^۱ به دلیل مشکل بودن تشخیص کسره اضافه آخر کلمه، این مورد در سیستم ما در نظر گرفته نشده است.

۲-۲-۴- رکن بندی

بعد از تقطیع اشعار به هجای کوتاه و بلند (رشته Udash) با توجه به جدول ارکان^۱، به جای هر چند هجا، یک رکن قرار می‌دهیم. همانطور که در صرف عربی، قالب‌هایی برای کلمات اختراع کرده‌اند و مثلاً می‌گویند شاعر بر وزن فاعل، در عروض نیز برای (یا مجموعه چند هجای کوتاه و بلند) قالب‌هایی وضع کرده‌اند و مثلاً می‌گویند «می‌آیم» (---) بر وزن مفعولن. بنای این قالب بر «ف»، «ع» و «ل» است. البته می‌توان به جای آن هر معادل دیگری به کار برد. مانند «ت» و «تن» که در این صورت «مفاعیلن» می‌شود: «ت تن تن تن».

مثال قبل و رشته Udash به دست آمده را رکن بندی می‌نماییم:

تَوَانَا بُودَ هَرِ کِه دَانَا بُودَ										
تَ	وَ	نَا	بُ	وَدَ	هَرِ	کِه	دَا	نَا	بُ	وَدَ
CV	CVC	CVV	CV	CVC	CVC	CV	CVV	CVV	CV	CVC
u	-	-	u	-	-	-	-	-	u	-
فعولن			فعولن			فعولن			فعل	

بنابراین وزن عروضی «توانا بود هر که دانا بود»، «فعولن فعولن فعولن فعل^۲» می‌باشد. باید به این نکته هم توجه کنید که اگر شاعری مصراع اول را با یک وزن خاص سرود به ناچار باید مصاریع بعدی را هم با همین وزن بسراید. همچنین اوزان یک مثنوی همگی یکسان هستند^۳، مثلاً مثنوی معنوی تماماً بر وزن «مفتعلن مفتعلن فاعلن» (کد ۱۰۲۱) سروده شده است.

۲-۲-۵- اوزان اشعار فارسی

از اجتماع ارکان مختلف، اوزان گوناگونی پدید می‌آید. در هیچ وزنی کمتر از دو و بیشتر از چهار رکن وجود ندارد. به عبارت دیگر کوتاه‌ترین مصراع شعر سنتی فارسی دارای دو رکن و طویل‌ترین آن دارای چهار رکن است. در عروض سنتی برای هر یک از اوزان اسمی نهاده بودند، مثلاً اوزان مبتنی بر

^۱ این جدول در پیوست دوم آورده شده.

^۲ متقارب مضمن محذوف. کد ۱۰۲۳ در جدول اوزان پرکاربرد.

^۳ برای اطلاع از اوزان مثنوی‌های مشهور به پیوست چهار مراجعه نمایید.

«فاعلاتن» را «رمل» و اوزان مبتنی بر «مستفعلاتن» را «رجز» می‌گفتند. در عروض جدید نیازی به حفظ اسامی اوزان نیست و در مواقع لزوم می‌توان به کتب عروضی قدیم مراجعه کرد [۱].

اوزان شعر فارسی بر سه نوع است:

- الف) اوزانی که شاعران به آنها شعر سروده‌اند و تعداد آنها متجاوز از ۲۵۰ تا است.^۱
- ب) اوزانی که عروضیان خود برای آنها شعر به عنوان مثال آورده‌اند.
- ج) اوزانی که حتی برای آنها مثال هم آورده نشده. [۳]

با اختلاف نظر فراوان، تقریباً همه صاحب نظران این رشته اعتقاد دارند که بیش از ۹۰ درصد اشعار فارسی بر ۳۳ وزن سروده شده است.^۲ بنابراین برای محدود کردن دامنه کار، ما در این پژوهش فقط از ۳۱ وزن استفاده کرده‌ایم.^۳

۲-۲-۶- اختیارات شاعری

کمتر شعر است که پس از تقطیع شعر به هجای کوتاه و بلند و رکن‌بندی، وزن اصلی شعر را نشان دهد، بلکه وزنی به دست می‌آید که به آن وزن تقطیعی گویند. برای تبدیل وزن تقطیعی به وزن اصلی (یعنی وزنی که در فهرست اوزان درج شده است) توجه به دقایقی که به آنها اختیارات شاعری گویند لازم است. البته بعضی از مواردی که در زیر می‌آید در کتب عروض تحت عنوان اختیارات شاعری نیامده، بلکه بعضی تحت عنوان «قواعد تقطیع» و برخی دیگر تحت عنوان «ضرورات و استثنا» آورده شده. ضمناً موارد زیر تمامی اختیارات شاعری نیست، بلکه گزیده‌ای از آنهاست.

۱. هجای کوتاه در آخر مصراع، بلند محسوب می‌شود. بنابراین پس از یافتن رشته Udash باید آخرین حرف آن را تبدیل به «-» بنماییم.

۲. شاعر مختار است در آخر مصراع یک یا دو حرف صامت، اضافه بر فرمول بیاورد (یا نیاورد). بنابراین پس از یافتن رشته CV باید Cهای اضافه آخر رشته (C تنها) را حذف نماییم.

^۱ طبق نظر دکتر کامیار بیش از ۲۵۰، طبق نظر دکتر شمیسا حدود ۳۰۰ و طبق نظر الول سانن ۱۱۵ تا

^۲ طبق آمارری که از روی اشعار ۴۲ تن از شاعران قدیم و جدید تهیه شده، نزدیک به ۹۹ درصد اشعار این شعرا بر ۲۹ وزن بوده. شمیسا اعتقاد دارد بیش از ۹۰ درصد اشعار بر ۳۲ وزن است. مسعود فرزاد اوزان پرکاربرد را ۳۳ دانسته.

^۳ این اوزان با نام و رشته Udash مربوطه در پیوست سوم آورده شده.

۳. از آنجا که کوچکترین جزء تشکیل دهنده وزن، هجای کوتاه است، از این روی در صورت برخورد به صامت تنها (C) در وسط رشته^۱ CV، آن را همانطور که قبلاً هم گفته شد، به عنوان یک هجای کوتاه (-) می‌شناسیم.

۴. در کلماتی از قبیل خواست، گوشت، بیست، کارد (طرح هجایی CVVCC) هرگاه صامت آخر به وسیله مصوت بعد از خود جذب نشود (با حذف همزه‌ی کلمه بعد)، همواره از تقطیع ساقط خواهد بود.

۵. هرگاه بعد از نونی که بعد از مصوت بلند قرار گرفته است، سکون یا مکث باشد، از کمیت مصوت بلند کاسته می‌شود

مرنجان دلم را که این مرغ وحشی (u--u)

مَرَنجان دِلَم													
م	َ	ل	َ	د		ن	ا	ج	ن	َ	ر	َ	م
c	v	c	v	c		c	v	c	c	v	c	v	c
-				u		-			-			u	

همانطور که در بالا می‌بینید، به جای VV در جایگاه «الف» V تنها قرار داده‌ایم.

۶. شاعر مختار است به جای فعلاتن در رکن اول هر مصراع، فاعلاتن بیاورد^۲ [۱].

۷. شاعر مختار است به جای دو هجای کوتاه (UU) یک هجای بلند بیاورد. عکس این مورد صحیح نمی‌باشد^۳. این اختیار که در اصطلاح به آن «تسکین» می‌گویند، جز در آغاز مصراع (مگر به ندرت) در همه جا قابل اعمال است.

^۱ در پایان رشته، حذف می‌شود.

^۲ در عروض فارسی عکس آن صحیح نیست، یعنی نمی‌توان به فاعلاتن در رکن اول وزنی، فعلاتن گفت. ولی در عربی عکس تبدیل فوق معمول است.

^۳ در عربی عکس این مورد جایز است و این یکی از مهم‌ترین تفاوت‌های عروض فارسی و عربی است.

۸. شاعر مختار است در برخی اوزان به جای «U»، «U-» بیاورد، و یا بالعکس عمل کند. این عمل را در اصطلاح «قلب» گویند. قلب در تبدیل مفتعلن به مفاعلن و بالعکس دیده می‌شود.

۲-۳- تاریخچه عروض

در اینکه منشا عروض فارسی اوزان شعری قبل از اسلام است یا دستگاه عروض عربی، بین محققان اختلاف است.

به نظر می‌رسد که دستگاه عروض شعر فارسی در اساس مقتبس از دستگاه عروض شعر عربی باشد، اما بعدها به مناسبت اختلاف دو زبان، بین این دو عروض فرقه‌هایی پیدا شده باشد. با این همه مسلم است که برخی از اوزان و حتی قوانین عروضی منشا ایرانی دارند.

اول کسی که در ادبیات عرب به تدوین و تنظیم یا وضع این علم پرداخت خلیل ابن احمد بصری نحوی است. او به پانزده بحر قائل بود. بعد از او، ابوالحسن سعید بلخی، یک بحر بر پانزده بحر افزود. سرانجام علمای ایرانی عروض از قبیل ابوالحسن سرخسی بهرامی، بزرجمهر قاینی و منشوری سمرقندی در اواخر قرن چهارم و اوایل قرن پنجم این علم را توسعه دادند. در اواخر قرن پنجم و اوایل قرن ششم امام حسن قطان مرزی با اختراع دو شجره اُخرب و اُخرم جهت استخراج اوزان مختلف رباعی به گسترش عروض کمک کرد. سرانجام نوبت به شمس قیس رازی - سرآمد علمای ایرانی عروض - رسید. شمس قیس در عروض، کتاب بسیار مهم «المعجم فی معاییر اشعار العجم» را نوشت که تا امروز کامل‌ترین مرجع و منبع برای مطالعه در عروض سنتی است. اهمیت این کتاب به حدی بود که جمیع آثار عروضی قبل از خود را تحت الشعاع قرار داد، چنان که امروزه از عروض‌دانان قبل از شمس قیس کتابی در دست نیست. بعد از شمس قیس هم هیچ حرکت اصیلی در عروض رخ نداد و تمام کتب عروضی تحت تاثیر مستقیم المعجم نوشته شد.

تنها کتابی که می‌توان آن را همسنگ المعجم شمرد «معیار الاشعار» خواجه نصیر طوسی است که در همان زمان تالیف المعجم نوشته شده است و در برخی از زمینه‌ها از نظر ابتکار و استقلال رای بر المعجم سر است.

بدین ترتیب عروض بعد از شمس قیس و خواجه نصیر حدود هفت قرن در دکود ماند. در قرون اخیر (از قرون ۱۸ به بعد) مستشرقان که به دلیل آشنایی با عروض لاتین و یونان باستان، با تقطیع

هجایی آشنایی داشتند، اشعار فارسی و عربی را بدان اسلوب تقطیع کردند و باب مطالعات جدید را گشودند.

در دوران مشروطه که آشنائی با شعر اروپایی حاصل شد، بحث‌هایی هم در زمینه نو کردن عروض در گرفت. کسانی چون عبدالحسین آیتی و یحیی دولت آبادی و دیگران کوشش‌هایی در این زمینه کردند که کلا سطحی بود و هیچ کدام موفق به ارائه نظریه بنیادی اصیلی نگشتند.

در روزگار ما اشخاص متعددی با شیوه‌های گوناگون به بحث در عروض پرداختند و برخی از آنان موفق به ارائه نظریات تازه‌ای گشتند که مجموعاً به عروض شکل علمی و تازه داد و توانست گوهر ساده و بی پیرایه عروض را از میان غبار و خروارها نام و اصطلاح بیهوده و توجیه‌های اشتباه بیرون آورد. از میان ایشان، مسعود فرزاد، دکتر پرویز ناتل خانلری و ابوالحسن نجفی را باید نام برد [۱].

۲-۴- در این فصل...

در این فصل در مورد برخی اصطلاحات شعر و عروض و تاریخچه عروض توضیحات مختصری داده شد. همچنین طریقه به دست آوردن وزن عروضی یک شعر به صورت دستی آموزش داده شد. از خواننده انتظار نمی‌رود که پس از خواندن این فصل توانایی لازم برای به دست آوردن وزن عروضی اشعار فارسی را داشته باشد. زیرا این امر نیاز به تمرین و ممارست فراوان دارد. این آموزش فقط به منظور آشنایی با موضوع نوشته شده است.

در فصل بعد در مورد برخی پیش‌نیازهای سیستم، مشکلات موجود بر سر راه سیستم و همچنین راه‌حلی که استفاده شده است، صحبت می‌شود.

فصل سوم

پیش‌نیازهای طرح

۳- پیش نیازهای طرح

برای آنکه این پژوهش انجام شود، قبل از هر مرحله می‌بایست پیش‌نیازهایی را انجام داد. برخی این پیش‌نیازها مربوط به تمامی پژوهش‌هایی است که بر روی متون فارسی انجام می‌شود و برخی دیگر مختص این پژوهش خاص می‌باشد.

۳-۱- استاندارد خط

یکی از مهمترین دشواریهای زبان فارسی (خط فارسی) وجود نداشتن الگوی استاندارد نوشتاری برای زبان فارسی (یا رعایت نکردن قانونهای آماده شده تا کنون برای آن) است. روند فارسی‌سازی و استاندارد نمودن زبان فارسی برای رایانه فراز و نشیبهای زیادی داشته است. در اینجا گوشه‌ای از آن آورده می‌شود.

۳-۱-۱- نخستین کارها

کُد `ascii` استاندارد ۱۲۷ نویسه دارد و چون این کُد در یک بایت (۸ بیت) گذاشته می‌شود ۱۲۸ حالت دیگر از ۲۵۶ حالت آن باقی می‌ماند. در آغاز هر نویسه^۱ی فارسی به یکی از این جاهای باقی مانده نگاشته می‌شد. این روش پیش از آن برای دیگر زبانها در دیگر کشورها به کار گرفته شده بود. در این روش هر برنامه‌نویس به دلخواه این نگاشت را با برنامه‌نویسی فراهم می‌کرد. کم‌کم برنامه‌هایی چون `vegaf`، سایه، سپند، ساحر و ... نوشته شد که این نگاشت را انجام می‌دادند و برخی تابعهای کتابخانه‌ای نیز برای برنامه‌نویسی فراهم می‌کردند. برخی برنامه‌نویسان این برنامه‌ها را برای فارسی‌سازی در نرم‌افزارهای خود به کار می‌بردند. هنوز نیز بسیاری از سایتهای فارسی نگاشت ویژه‌ی خود را به کار می‌برند و فونت (`font`) ویژه‌ی خود را دارند.

^۱ Character

۳-۱-۲- کد CP1256

سپس شرکت‌های نرم‌افزاری بزرگ دنیا که علاقه‌مند به فروش نرم‌افزارهای خود در کشورهای دیگر بودند به کمک همین روش (۱۲۸ حالت آزاد) کدی را ساختند که همه‌گیر شد. یکی از بزرگترین در دسترس‌های این کد برای فارسی رعایت نکردن ترتیب چهار نویسه‌ی ویژه‌ی فارسی (پ،چ،ژ،گ) بود. زیرا که این شرکت‌ها پایه‌ی کار خود را بر زبان عربی گذاشته بودند و سپس فارسی را بدان افزوده بودند. برای هر زبان یا مجموعه‌ی زبان نزدیک به هم که با این روش کدگذاری شده بودند نام ویژه‌ی برگزیده شد. زیرا باید به گونه‌ای می‌شد فهمید که ۱۲۸ حالت آزاد را برای چه زبانی به کار برده‌اند. کد عربی با این روش cp1256 (یا windows1256) نامیده شد. این کد در ایران آنقدر همه‌گیر شد که آن را unicode می‌خوانند که نامگذاری نادرستی است. یکی از قاعده‌هایی که در این کد رعایت شد، تغییر شکل نویسه با توجه به نویسه‌های پس و پیش آن است. به این ترتیب برای هر حرف فارسی تنها یک نویسه به کار گرفته می‌شود و تنها شکل نویسه هنگام در کنار هم گذاشتن نویسه‌ها در یک واژه تغییر می‌یابد.

۳-۱-۳- یونی کد unicode

یونی کد یک کد جهانی است که چند شرکت بزرگ نرم‌افزاری دنیا به همراهی سازمان جهانی استاندارد (ISO/JTC1/SC2) آن را به وجود آورده‌اند. با چشم‌پوشی از فراز و نشیب‌های این روش به سادگی می‌توان گفت که در این روش به جای یک بایت، دو بایت به کار برده می‌شود. این کدگذاری رویه‌ی پایه‌ی چند زبانی^۱ نامیده می‌شود. این کد ۶۵۵۳۵ نویسه دارد.

یونی کد از مجموعه نویسه‌های محدود ۸ بیتی بسیار فراتر رفته و با ظرفیت بیش از یک میلیون نویسه، امکان کدگذاری همه‌ی زبانهای نوشتاری دنیا را فراهم می‌کند. افزون بر این برای برگزیدن خط و زبان نوشته، به کدهای کنترلی نیاز ندارد. یونی کد رفتار یکسانی را با نویسه‌های الفبایی، نویسه‌های اندیشه‌نگار و نمادها و نشانه‌ها دارد که امکان به کار بردن آمیخته‌ی آنها را با یکدیگر فراهم می‌کند. یونی کد افزون بر تعیین کد عددی و نام برای هر نویسه که در استانداردهای همانند، به کار گرفته شده است؛ اطلاعات یونی کدی بیشتری را نیز برای پردازش و نمایش نوشته‌ها فراهم می‌کند؛ که از آن میان می‌توان به جهت نویسه و ویژگی‌های الفبایی اشاره کرد.

^۱ Basic Multilingual Plane BMP

در این کد نیز زبان‌هایی که نزدیک به هم هستند در یک گروه گذاشته شده‌اند و باز چهار حرف «پ»، «چ»، «ژ» و «گ» ترتیب درستی ندارند. در این کد چند نویسه همانند با شماره‌های گوناگون وجود دارند؛^۱ مانند چند «ی» و دو «ک» و ... این دشواری دیگری است که به دشواری‌های پیشین افزوده شده است. با اینکه این استاندارد به روشنی زبان‌هایی را که این نویسه‌های همانند در آنها کاربرد دارند، نشان می‌دهد ولی برخی از سیستم‌ها (مانند win2000 و winxp) این قانون‌ها را رعایت نمی‌کنند و در هنگام برنامه‌نویسی سردرگمی ایجاد می‌کنند. برخی نرم‌افزاری تحت ویندوز به خوبی یونی‌کد را پشتیبانی می‌کند مانند unipad و Notepad++. سیستم عامل linux نیز این استاندارد را رعایت می‌کند و در نرم‌افزار gedit در نسخه linux mandrake 9.1 به خوبی فارسی استاندارد پشتیبانی شده است و صفحه‌ی کلید استاندارد ایران نیز رعایت شده است (بر خلاف winxp که صفحه‌ی کلید ویژه خود را داراست که بیشتر عربی است تا فارسی). همواره بهترین و به‌روزترین جا برای دیدن آخرین نسخه استاندارد یونی‌کد در www.unicode.org است.

۳-۱-۴- کد با طول متغیر UTF-8

کد `ascii` سالها استاندارد رایانه‌ها بوده است بسیاری از سخت‌افزارها، شبکه‌ی جهانی و ... بر پایه‌ی آن ساخته شده‌اند. بنابراین عوض کردن این کد (یک بایتی) اگر ناممکن نباشد بسیار پرهزینه و زمانگیر خواهد بود. چاره‌ای که اندیشیده شد به کار بردن کدی با تعداد بایت متغیر بود که با یونی‌کد نگاشت یک به یک داشته باشد و همچنین ۱۲۸ نویسه `ascii` را به همان شکل یک بایتی باقی گذارد.

برای نگاشت دیگر نویسه‌های یونی‌کد، ۱۲۸ حالت باقی مانده از نخستین بایت و احتمالاً بایت‌های پس از آن به کار گرفته می‌شود. نخستین بایت نشانه‌ای است که تعداد بایت‌های پس از خود و رسته یونی‌کد را نشان می‌دهد در حالتی که نویسه یک نویسه‌ی اسکی باشد همان یک بایت کافی خواهد بود [۹].

این سیستم روزبه‌روز در حال گسترش است. در ایران نیز این استاندارد پذیرفته شده است ولی هنوز همه‌گیر نشده است.

^۱ برای اطلاع از نویسه‌های یکسان به پیوست پنجم مراجعه کنید

۳-۱-۵- راه حل برای مشکل استاندارد خط

در این پژوهش ما از استاندارد UTF-8 استفاده کرده‌ایم، البته باید توجه داشت که نویسه‌های فارسی در این استاندارد و استاندارد یونی‌کد تفاوت چندانی ندارند. فقط کافی است در هنگام ذخیره فایل‌هایی که حاوی متن هستند، آنها را به صورت UTF-8 ذخیره کنید، هنگام ورود اطلاعات از طریق فیلدهای ورودی هم نگرانی وجود ندارد. تنها نگرانی موجود، وجود نویسه‌های مشابه است که در پیوست پنجم به تفصیل در این مورد صحبت شده است. بنابراین برای کار با خط فارسی، ۱- متون خود را به صورت UTF-8 ذخیره نموده^۱ و در برنامه‌هایی که متن فارسی را از فایل نمی‌خواند، متن ورودی را با استفاده از توابع آماده‌ای - که تقریباً تمامی نسخه‌های جدید زبان‌های برنامه‌نویسی دارا می‌باشند - به یونی‌کد یا UTF-8 تبدیل نمایید. ۲- توجه کنید که انواع «ک» و «ی» را قبل از پردازش به یک نوع تبدیل کنید و یا اینکه هر دو نوع را در نظر بگیرید.

در این پژوهش ما از توابع نوشته شده php برای تبدیلات utf-8 و کار با آن استفاده کرده‌ایم، همچنین قبل از هر پردازشی نویسه‌های مشابه در متن ورودی را تبدیل می‌نماییم.

۳-۲- حذف نویسه‌های اضافی

یکی از پیش پردازش‌هایی که در کار با متون در هر زبانی انجام می‌شود، حذف نویسه‌های اضافی^۲ است. این نویسه‌ها معمولاً نویسه‌هایی از قبیل «؟» ، «-» هستند که در قواعد نوشتاری باید نوشته شوند ولی برای پردازش متن نیازی به آنها نیست.

در متون فارسی علاوه بر نویسه‌های معمول در زبان‌های دیگر دو نوع نویسه خاص زیاد مورد استفاده قرار می‌گیرند که باید به آنها توجه شود. این دو نویسه، یکی خط تیره^۳ است و دیگری فاصله مجازی^۴ و اتصال مجازی^۵ است.

^۱ اگر حتی با ویرایشگر ساده notepad هم امکان پذیر است، به این صورت که در بخش ذخیره به عنوان (Save as) کدینگ مورد نظر خود را به UTF-8 تغییر دهید. در ویرایشگرهای پیشرفته‌تر نظیر notepad++ به راحتی کدینگ یک فایل را می‌توانید تغییر دهید.

^۲ Black List

^۳ Em Dash

^۴ ZERO-WIDTH NON-JOINER

^۵ ZERO-WIDTH JOINER

۳-۲-۱- نویسه خط تیره

این نویسه معمولاً برای تراز کردن انتهای خطوط مورد استفاده قرار می‌گیرد، معمولاً در نرم‌افزار Office Word با استفاده از Shift+J یا Ctrl+Num- قابل نوشتن است. در کلمه «مهم» اگر چندبار از این نویسه خاص استفاده شود به صورت «مهم» درخواهد آمد. در این پژوهش این نویسه به همراه بقیه نویسه‌های اضافی حذف خواهد شد.

۳-۲-۲- فاصله مجازی و اتصال مجازی

این دو نویسه -که درباره آنها و روش ورود آنها در صفحه کلیدهای مختلف در پیوست پنجم به تفصیل سخن گفته شده- معمولاً برای جدا کردن «می» از ابتدای کلمه در کلماتی مانند «می‌روم» و جدا کردن «ها» از انتهای کلماتی مانند «کتاب‌ها» مورد استفاده قرار می‌گیرند. این کار کمک می‌کند تا این کلمات با فاصله جدا نشوند (در صورتی که با فاصله جدا شوند دو کلمه مجزا در نظر گرفته می‌شوند). و ضمناً به صورت جدا نویسی نوشته شوند. (طبق نظر فرهنگستان فارسی «می‌روم» صحیح است نه «میروم»). البته موارد گفته شده برای فاصله مجازی است و نه اتصال مجازی. برای دیدن کاربرد اتصال مجازی به این دو کلمه توجه کنید: «ه» و «ه» در اولی پس از نویسه «ه» هیچ نویسه‌ای قرار ندارد و در دومی پس از آن یک اتصال مجازی قرار داده شده.

نویسه فاصله اضافی مانند نویسه‌های دیگر به راحتی قابل حذف نیست. زیرا اگر آن را به طور کلی حذف کنیم، ممکن است خواندن کلمه کمی فرق کند، اگر به جای آن فاصله قرار دهیم، کلمه تبدیل به دو کلمه مجزا خواهد شد. بنابراین باید بسته به پردازش مورد نظر با این نویسه برخورد کرد.

۳-۳- ساخت Data Set

یکی از مشکلات موجود بر سر تحقیقات فارسی نبود یک مجموعه داده^۱ برای آزمایش برنامه‌های تولیدی می‌باشد. در صورتی که انواع مجموعه داده‌ها به زبان انگلیسی بر روی فضای سایبر وجود دارد.

^۱ Data Set

در مورد این پروژه خاص نیز این مشکل وجود داشت. به همین دلیل بر آن شدیم که یک مجموعه داده مناسب برای این پروژه طراحی بنماییم. برای این کار از اشعار دیوان غزلیات خواجه حافظ شیرازی [۶] و دیوان شمس مولانا جلال الدین [۷] استفاده کردیم. البته متن تایپ شده این اشعار بر روی اینترنت موجود بود، که با وجود غلط‌های تایپی و همچنین عدم رعایت تصحیح‌های معتبر مجبور به استفاده از همین نسخ شدیم.

برای یافتن اوزان اشعار در مورد دیوان حافظ، در [۶] وزن هر یک از غزل‌ها نوشته شده بود، که با کدگذاری و تبدیل به فرمت دلخواه از آن استفاده کردیم. و در مورد دیوان شمس از خانم ده‌نمکی یکی از اساتید ادبیات کمک گرفتیم، که جا دارد در این جا باز هم از زحمات ایشان قدردانی نمایم. ایشان وزن عروضی بیش از ۴۵۰ غزل از این مجموعه را استخراج نمودند.

با حذف موارد اختلافی و استفاده از یک انتخاب تصادفی بیش از ۱۰۰۰ بیت از اشعار حافظ و مولانا همراه با کد وزن آنها به صورت قالب استاندارد تعریفی^۱ درآوردیم. یکی از مواردی که در ساخت این مجموعه داده مورد توجه قرار گرفته تعداد هر کد وزن است. تعداد هر کد وزن بسته به کاربرد هر وزن تعیین شده، مثلاً برای کد وزن ۱۰۰۲ که پرکاربردترین وزن فارسی است بیشترین تعداد بیت در نظر گرفته شده است. برای بالا بردن دقت اوزان انتخابی تمامی ابیات انتخابی دوبار بررسی گردیده است.

۳-۴- زبان برنامه نویسی

برای پیاده سازی این پژوهش زبان‌های مختلفی مورد بررسی قرار گرفت، به دلیل آنکه قصد داشتیم این پژوهش را برای علاقه‌مندان در اینترنت به اشتراک بگذاریم، به دنبال برنامه‌ای تحت وب بودیم که به راحتی بتواند پردازش متن را انجام دهد.

زبان برنامه نویسی Python و PHP دو گزینه خوب برای این کار بودند، ولی با توجه به اینکه زبان PHP توسط اکثر سرورها پشتیبانی شده و همچنین نیاز به هیچ واسط دیگری ندارد، این زبان برای کار انتخاب شد.

^۱ برای اطلاع از قالب تعریفی به پیوست ششم مراجعه نمایید.

یکی از مشکلات PHP که برنامه‌نویسان قول داده‌اند در نسخه ۶ این زبان حل شود، عدم پشتیبانی کامل UTF-8 است. برای حل این مشکل از یکی از بسته‌های آماده^۱ که توسط برنامه‌نویسان این زبان نوشته شده بود استفاده کردیم. این کلاس آماده تمامی توابع کار با رشته PHP را هماهنگ با استاندارد UTF-8 می‌نماید.

۳-۵- در این فصل...

در فصلی که گذشت، پیش‌نیازهای طرح و همچنین روش رفع نیازها بررسی گردید. در این فصل استاندارد خط و استاندارد دی که ما مورد استفاده قرار داده‌ایم، نویسه‌های اضافی و حذف آنها، ساخت مجموعه داده و زبان برنامه‌نویسی برای پیاده‌سازی مورد بحث قرار گرفت.

در فصل بعد شرح خواهیم داد که چگونه پس از رفع نیازها، تمامی صحبت‌هایی که در فصل دوم در مورد به دست آوردن عروض گفته‌ایم، تبدیل به یک روش سیستماتیک خواهد شد. در آن جا به شرح الگوریتمی که برای این کار ابداع شده است پرداخته می‌شود.

^۱ package utf8, author: Chris Smith, chris@jalakai.co.uk, www.phpclasses.com

فصل چهارم

الگوریتم سیستم

پیشنهادی

۴- الگوریتم سیستم پیشنهادی

برای حل این مساله خاص (یافتن وزن عروضی اشعار فارسی)، می‌توان از روش‌های مختلفی استفاده کرد، بر اساس بررسی‌های انجام شده، ما از عروض علمی و همچنین شیوه دکتر شمیسا با تغییراتی استفاده کرده‌ایم. در این روش ابداعی که نخستین الگوریتم پیشنهادی برای این کار است، مساله یافتن وزن به شش مرحله تقسیم می‌شود که جلوتر توضیح خواهیم داد.

۴-۱- تعریف مساله

برای تقریب ذهن و همچنین خارج کردن مساله از حالت تخصصی ادبیات مساله تشخیص وزن عروضی اشعار فارسی را به صورت زیر تعریف می‌نماییم:

« پس از تغییرهایی بر روی یک رشته ورودی و تبدیل به آن رشته خروجی، سعی می‌شود رشته خروجی را در یکی از دسته‌های موجود جای داد»

رشته ورودی در این‌جا، یک بیت شعر فارسی است.

رشته خروجی، رشته‌ای از «U» و «-» که در آن «U» به معنای هجای کوتاه و «-» به معنای هجای بلند است.

تغییرات به معنای تغییر حروف به رشته صامت مصوت و سپس تبدیل این رشته به U- و همچنین اعمال اختیارات و ضرورات شاعری است. تغییرات به دو دسته کلی تقسیم‌بندی می‌شوند:

۱- تغییرات قطعی: این تغییرات در طول روند تبدیل رشته ورودی به خروجی اعمال می‌شوند. که خود به دو دسته حتمی و مشروط تقسیم می‌شوند:

الف: تغییرات قطعی حتمی: تغییراتی که حتما باید انجام شوند، مثلا تغییر رشته CV به رشته Udash که بر اساس قوانینی این تغییرات اعمال می‌شوند

ب: تغییرات قطعی مشروط: این تغییرات را می‌توان با قرار دادن شروطی حتمی دانست. بعضی اختیارات شاعری، تغییر قطعی مشروط می‌باشند. به عنوان مثال: شاعر مختار است به جای فعلاتن در

رکن اول هر مصراع، فاعلاتن بیاورد[۱]. درست است که در این اختیار قطعیت وجود ندارد، ولی با توجه به اوزان موجود، می‌توان این اختیار شاعری را به یک تغییر قطعی مشروط تبدیل کرد.

۲- تغییرات غیر قطعی : این تغییرات قابل اعمال در طول روند تبدیل ورودی به خروجی نمی‌باشند. می‌توان این تغییرات را به عنوان تابع شباهت با وزن خاص در تشخیص الگو به کار برد.

۴-۲- مراحل پروژه

برنامه به شش مرحله مختلف به صورت زیر تقسیم می‌شود:

۱- حرکت گذاری

۲- تبدیل مصراع حرکت گذاری شده به رشته CV

۳- اصلاح رشته CV بر اساس اختیارات شاعری

۴- تبدیل رشته CV اصلاح شده به رشته udash

۵- اصلاح رشته udash بر اساس اختیارات شاعری

۶- انطباق یکی از اوزان معروف با رشته udash

مراحل یک تا پنج همان روند تبدیل رشته ورودی به خروجی می‌باشد، و مرحله ششم تشخیص الگو. توجه شود در این برنامه ما فقط قادر به یافتن وزن شعر ورودی خواهیم بود، بنابراین اگر وزن شعر ورودی اشتباه باشد و یا اینکه وزن دو مصراع همخوانی نداشته و یا اشتباه از اختیار شاعری استفاده شده باشد و از این قبیل، برنامه قادر به تشخیص نبوده و خروجی قابل اعتمادی نخواهد داد.



۴-۲-۱- حرکت گذاری

توجه شود که کلمات ما تک تک حرکت گذاری خواهند شد. بنابراین قادر به تشخیص حرکت آخر کلمه نخواهیم بود و این امر دقت برنامه را کاهش داده است. دو حالت برای حرکت گذاری داریم:

۴-۲-۱-۱- با استفاده از مجموعه لغات

در این برنامه ما از یک مجموعه لغات (Lexicon) فارسی استفاده کرده ایم. مجموعه استفاده شده از یک برنامه متن باز به نام «واژگان زبانی فارسی» برداشته شده و پس از اعمال تغییراتی با فرمت جدید مورد استفاده قرار گرفته اند. این لغات شامل نوعی تلفظ به صورت فونتیک می باشند، که کار ما را در تبدیلات آتی بسیار آسان می نماید. البته در پایگاه داده این برنامه خصوصیات دیگری از کلمات هم نوشته شده، که ما فقط تلفظ آنها را مورد استفاده قرار دادیم. در این پایگاه داده بیش از ۵۰۰۰۰ کلمه فارسی جمع آوری شده است. بعضی از این کلمات در اشعار ما کاربردی نداشته (مانند آث میلان) و برخی از کلمات استفاده شده در اشعار در این پایگاه داده وجود ندارد (مانند حالات مختلف افعال). استفاده از این پایگاه داده برای حرکت گذاری کلمات نسبت به حالت بعدی، دقت بسیار بالاتری دارد.

توجه شود که ممکن است کلمه ای در مجموعه لغات نباشد ولی به راحتی و با استفاده از پیشوند و پسوند ساخته شود، مانند کلمه ای «می خواهم» که کلمه «خواه» در مجموعه کلمات موجود است و «می خواهم»، خیر. بنابراین دو مجموعه لغت پیشوند و پسوند هم ساخته شد، در صورتی که کلمه ای با این پیشوند و پسوندها ساخته شود، بتوان تلفظ آن را به دست آورد.

۴-۲-۱-۲- بدون استفاده از مجموعه لغات

در صورتی که کلمه ای در پایگاه داده بود، تلفظ آن به صورت دقیق به دست می آید، در غیر این صورت با استفاده از قوانین حرکت گذاری این کار انجام می شود. قوانین حرکت گذاری اعمال شده به صورت زیر است:

۱. اگر حرف «الف»، «ی» و یا «و» باشد، روی آن هیچ حرکتی قرار نمی دهد، مانند «ابر»
۲. اگر حرف الف نباشد ولی حرف بعدی آن الف باشد، باز هم حرکتی روی حرف قرار نمی گیرد، مانند حرف «م» در کلمه «مار»
۳. اگر حرف، آخرین حرف کلمه باشد هم حرکتی روی آن قرار نمی گیرد

۴. در غیر از سه مورد بالا، حرکت «-» روی حرف قرار داده می‌شود. توجه شود که نوع حرکت اهمیتی ندارد. زیرا مثلاً دو کلمه «حَسَن» و «حَسین» هر دو از دو هجای کوتاه و بلند تشکیل شده‌اند.

جالب است که با این روش (که تخمین بسیار ضعیفی می‌باشد) و بدون استفاده از مجموعه لغات، تا حدود ۳۰ درصد، برنامه وزن‌های ورودی را صحیح تشخیص داد.

۴-۲-۲- تبدیل مصراع حرکت‌گذاری شده به رشته صامت مصوت (CV)

اگر در مرحله قبل کلمه‌ای در مجموعه لغات یافت شد، تبدیل آن به CV کار بسیار راحتی می‌باشد، این کار به وسیله جدول زیر و چند قانون ساده به صورت تقریباً دقیق انجام خواهد شد.

تبدیل تلفظ‌های موجود در مجموعه لغات به CV

تلفظ حرف	حرف	نشانه	نوع	مثال	تلفظ مثال
A	ا	vv	مصوت بلند	ناسزا	nAsezA
i	ی	vv	مصوت بلند	حصین	hasin
u	او	vv	مصوت بلند	سوز	suz
a	َ	v	مصوت کوتاه	بهجت	bahjat
e	ِ	v	مصوت کوتاه	زرنگ	zerang
o	ُ	v	مصوت کوتاه	مناظره	monAzere
'	ء	c	صامت	آیه	'Aye
Other	بقیه حروف	c	صامت	بهشت	beheSt

در صورتی که کلمه ورودی به وسیله قوانین حرکت‌گذاری، حرکت‌گذاری شده، باید در این مرحله طبق قوانین زیر کلمه حرکت‌گذاری شده تبدیل به رشته CV شود.

- ۱- به جای هر یک از مصوت‌های کوتاه (َ، ِ، ُ) یک v قرار می‌دهد
- ۲- به جای هر یک از مصوت‌های بلند (ا، ی، و) vv قرار می‌دهد
- ۳- به جای صامت‌ها، c قرار می‌دهد

در این تبدیلات موارد زیر رعایت می‌شود:

۱- در کلماتی مانند «تَوَسَّل»، «مِیْخانه» ویا «تَوَانا» که حروف «و» و «ی» به عنوان صامت استفاده شده‌اند و نه مصوت، تابع به جای این حروف C قرار می‌دهد و نه VV. تشخیص این امر بدین صورت است: اگر حرف قبل از «ی» یا «و» یک مصوت باشد (مانند «تَوان» و یا «تَوَانا») و یا بعد از این حروف یک مصوت باشد (مانند «مروان») این حروف حتما صامت هستند، در غیر این صورت به عنوان مصوت بلند شناخته خواهند شد.

۲- اگر حرف ابتدای کلمه «آ» بود:

a. اگر حرف آخر کلمه قبل برابر صامت (C) بود باید به جای «آ» CVV قرار دهیم.
b. در غیر این صورت، باید به جای آن CVV قرار داد. زیرا کلمات فارسی هیچیک با مصوت آغاز نمی‌شوند. و در کلماتی مانند «آب» حرف اول همزه به علاوه مصوت «ا» می‌باشد.

۳- اگر حرف ابتدای کلمه «ا» بود:

a. اگر حرف آخر کلمه قبل برابر صامت (C) بود، باید به جای «ا» CV قرار دهیم.
b. در غیر این صورت، به جای آن CV قرار می‌دهیم.
۴- نون ساکن آخر کلمه که قبل از آن مصوت بلند آمده باشد. به جای نون ساکن حرف n که معادل یک صامت (C) است قرارداده شده. این حرف n برای تشخیص اختیار شاعری شماره ۴ نوشته می‌شود.

۴-۲-۳- اصلاح رشته CV بر اساس اختیارات شاعری

با توجه به اختیارات شاعری تغییرات زیر روی رشته CV انجام می‌شود.

۱- اختیار شاعری شماره ۲: شاعر مختار است در آخر مصراع یک یا دو حرف صامت، اضافه بر فرمول بیاورد (یا نیاورد). بنابراین اگر در انتهای رشته چند C تنها وجود داشت، آنها را حذف می‌نماییم.

۲- اختیار شاعری شماره ۴: در کلماتی از قبیل خواست، گوشت، بیست، کارد (طرح هجایی CVVCC) هرگاه صامت آخر به وسیله مصوت بعد از خود جذب نشود (با حذف همزه‌ی کلمه بعد) همواره از تقطیع ساقط خواهد بود. یعنی CVVCCC تبدیل به CVVCC خواهد شد.

۳- اختیار شاعری شماره ۵: هرگاه بعد از نونی که بعد از مصوت بلند قرار گرفته است، سکون یا مکث باشد، از کمیت مصوت بلند کاسته می‌شود. یعنی $vvn \dots$ تبدیل به $vc \dots$ خواهد شد.

دو تغییر اول جزء تغییرات قطعی است، ولی سومی را نمی‌توان به صورت قطع بیان کرد، ولی با تخمین خوبی می‌توان آن را به برنامه اضافه کرد.

۴-۲-۴ تبدیل رشته cv اصلاح شده به رشته udash

هر هجا باید با صامت (C) شروع شود. بنابراین، باید از C شروع کرده و جلو برویم تا هجای کوتاه و بلند مشخص شود:

۱. هجای کوتاه: این هجا را با U نشان می‌دهیم، به دو صورت دیده می‌شود: C و یا cv
۲. هجای بلند: این هجا را با - نشان می‌دهیم. به دو صورت دیده می‌شود: CVV و یا CVC

۴-۲-۵ اصلاح رشته udash بر اساس اختیارات شاعری

در این مرحله تغییرات زیر بر روی رشته Udash انجام خواهد شد:

۱- اختیار شاعری شماره ۱: هجای کوتاه در آخر مصراع بلند حساب می‌شود. بنابراین هجای آخر مصراع تبدیل به هجای بلند می‌شود. به عبارت دیگر، آخرین کاراکتر رشته به «-» تبدیل می‌شود

۲- اختیار شاعری شماره ۶: شاعر مختار است به جای فعلاتن در رکن اول هر مصراع، فاعلاتن بیاورد. با بررسی‌های صورت گرفته بر وری اوزان موجود (الگوها) هر کجا در ابتدای مصراع فعلاتن (UU--) آمده پس از آن یا دوباره فعلاتن آمده و یا مفاعلن (U-U-) بنابراین با یک تخمین خوب دو تبدیل زیر را در صورتی که ابتدای رشته باشد انجام می‌دهیم:

"U--UU" → "UU--UU"	فاعلاتن فعلاتن <----
"U--UU" → "UU—U-"	فاعلاتن مفاعلن <-----

۳- در هیچ وزنی سه هجای کوتاه پشت سرهم نمی‌آیند [۱]. بنابراین با یک تخمین خوب UUUU را تبدیل به UUU- می‌نماییم

۴-۲-۶- انطباق یکی از اوزان معروف با رشته udash

پس از به دست آوردن رشته Udash نوبت به تطبیق آن با یکی از اوزان مشهور می‌رسد. در این مرحله رشته به دست آمده را با تمامی ۳۱ وزن موجود در پایگاه داده مقایسه کرده و نزدیکترین رشته را به عنوان وزن اعلام می‌نماییم. برای این کار می‌توان از الگوریتم‌های معروف استفاده کرد ولی به دلیل حالت خاص مسئله و همچنین برای اعمال برخی اختیارات شاعری باید از الگوریتم‌های دیگری هم استفاده کرد. به دلیل اهمیت این بخش از کار، این بخش در فصلی جداگانه شرح داده شده است.

۴-۳- در این فصل...

در فصلی که مورد مطالعه قرار دادید، الگوریتم شش مرحله‌ای برای یافتن عروض پیشنهاد شد. در الگوریتم پیشنهادی ابتدا شعر حرکت‌گذاری شده، سپس تبدیل به رشته صامت-مصوت شده و پس از اصلاح این رشته، رشته مذکور تبدیل به رشته Udash خواهد شد. پس از اصلاح این رشته نوبت به تطابق این رشته با یکی از اوزان معروف می‌رسد.

در این فصل پنج مرحله‌ی اول به طور کامل توضیح داده شد، ولی به دلیل اهمیت مرحله‌ی ششم و همچنین گستردگی و تفاوت آن با مراحل قبل، در فصل بعد به تشریح این مرحله پرداخته خواهد شد. آن‌جا به صورت کامل با الگوریتم‌ها و روش‌های مورد استفاده در این مرحله آشنا خواهید شد.

فصل پنجم

انطباق

۵- انطباق

پس از یافتن رشته Udash در مراحل ۱ تا ۵ نوبت به انتخاب یک وزن از ۳۱ وزن موجود در پایگاه داده می‌رسد. در این مرحله باید رشته Udash به دست آمده را با همه ۳۱ وزن موجود در پایگاه داده مقایسه نموده تا شبیه‌ترین وزن به رشته خروجی (رشته Udash به دست آمده در مراحل قبل) را بیابیم. توجه کنید که یک وزن خاص دارای فقط یک نوع رشته Udash می‌باشد. مثلاً رشته‌ی Udash وزن «خفیف مخبون محذوف» که با «فعلاتن مفاعلن فعلن» خوانده می‌شود، به صورت زیر است: «UU-U-U--UU-» این وزن همیشه باید به این رشته ختم شود و این رشته نیز به هیچ وزن دیگری نمی‌رسد.^۱

۵-۱- الگوریتم‌های انطباق رشته

الگوریتم‌های انطباق رشته^۲ یا جستجوی رشته یکی از مباحث جذاب در رشته کامپیوتر می‌باشد. این الگوریتم‌ها معمولاً به این صورت کار می‌کنند که رشته ورودی را با یک مجموعه رشته مقایسه کرده و برازندگی^۳ رشته ورودی با هر یک از رشته‌های مجموعه را اعلام می‌نمایند. البته این حداقل کاری است که این الگوریتم‌ها انجام می‌دهند، بسته به نوع مساله نزدیکی به رشته دیگر اعلام می‌شود. این الگوریتم‌های کاربردهای فراوانی دارند. اگر در موتور جستجوی Google کلمه‌ای اشتباه وارد کنید، مثلاً fitness را وارد کنید، در این صورت گوگل به شما هشدار می‌دهد که «آیا منظورتان کلمه‌ی fitness بوده است؟» در نرم‌افزار MS Word هم اگر این کلمه را تایپ کنید، زیر کلمه خط قرمزی می‌کشد که کلیک راست کردن می‌توانید یکسری کلمات صحیح پیشنهادی را ببینید. تمامی این کارها با الگوریتم‌های شباهت رشته انجام می‌شود.

^۱ البته اوزان ذو بحرین وجود دارند که در واقع دو یا چند اسم هستند برای یک وزن و نه دو یا چند وزن جداگانه

^۲ String Searching OR String Matching

^۳ Fitness

از جمله این الگوریتم‌ها می‌توان به soundex اشاره کرد که مخصوص تلفظ کلمات می‌باشد [۱۹]. این الگوریتم برای زمانی مفید است که شما تلفظ کلمه را بدانید ولی طرز نوشتن آن را خیر.

و یا می‌توان به الگوریتم‌های شباهت Rabin-Karp، Finite state automaton، Bitap، Boyer-Moore و بسیاری دیگر اشاره کرد.

کارکرد مهم‌تر این الگوریتم‌ها جستجوی الگو^۱ می‌باشد. کاری که ما در این پژوهش به آن نیاز داریم. البته باید توجه کرد که هر الگوریتم برای موقعیت خاصی ساخته شده، و نزدیکی را به صورت خاصی مطرح می‌کند. یکی نزدیکترین از نظر تلفظ و دیگری نزدیکترین از لحاظ داشتن نویسه‌های یکسان.

۵-۱-۱- Levenshtein

levenshtein نام الگوریتمی از دسته الگوریتم‌های انطباق رشته می‌باشد. این الگوریتم توسط شخصی به همین نام در سال ۱۹۶۵ مطرح شد، البته این نام پس از مرگ وی بر این الگوریتم گذاشته شد. با آن که این الگوریتم، الگوریتم جدیدی نیست ولی به دلیل کارایی بالای آن هنوز مورد استفاده است. این الگوریتم کمترین تغییری^۲ که نیاز است تا یک رشته تبدیل به رشته دیگر شود را بیان می‌کند. البته به این الگوریتم Levenshtein Distance نیز می‌گویند. شیوه کار این الگوریتم به صورت زیر است:

گام ۱: نام رشته اول را s و نام رشته دوم را t قرار می‌دهیم.

طول s را n و طول t را m می‌نامیم.

اگر $n=0$ بود، m را برگردان و خارج شو.

اگر $m=0$ بود، n را برگردان و خارج شو.

یک ماتریس d را که شامل $0 \dots m$ سطر و $0 \dots n$ ستون می‌باشد، بساز.

گام ۲: سطر اول را از 0 تا m مقدار دهی کن.

ستون اول را از 0 تا n مقدار دهی کن.

^۱ Classification

^۲ منظور از تغییر یا عملیات، حذف یک نویسه، افزودن یک نویسه و یا جایگزینی یک نویسه می‌باشد. که آخری نیز یک تغییر محسوب می‌شود.

گام ۳: هر نویسه s ، i از ۱ تا n را با هر نویسه t ، j از ۱ تا m را به صورت زیر مقایسه کن:
اگر $s[j]=t[j]$ بود، $cost[i][j]$ را صفر قرار می‌دهیم، در غیر این صورت ۱ قرار می‌دهیم.

گام ۴: $d[i,j]$ از ماتریس ساخته شده را برابر کمینه مقادیر زیر قرار بده:

الف) $d[i-1,j]+1$

ب) $d[i,j-1]+1$

ج) $cost[i][j]+d[i-1,j-1]$

گام ۵: گام چهار را برای تمامی خانه‌های ماتریس تکرار کن، $d[n,m]$ برابر فاصله Levenshtein می‌باشد.

۵-۱-۲- مثالی از Levenshtein Distance

برای درک بهتر الگوریتم به دست آوردن فاصله Levenshtein در زیر دو رشته را با هم مقایسه می‌نماییم. رشته اول "GUMBO" و رشته دوم "GAMBOL" می‌باشد. جواب به دست آمده باید مبین این مساله باشد که با چند حرکت (افزودن نویسه، حذف نویسه، جایگزینی نویسه) می‌توان از رشته "GUMBO" به رشته "GAMBOL" رسید (و یا بالعکس).

رشته "GUMBO" را رشته t نامیده و رشته "GAMBOL" را s .

بنابراین n برابر ۶ و m برابر ۵ خواهد بود.

ماتریس‌های تولید شده به صورت زیر است، هم اکنون باید با توجه به الگوریتم آنها را کامل کنیم.

ماتریس d							ماتریس COST						
		G	U	M	B	O							
	0	1	2	3	4	5		G	U	M	B	O	
G	1							G					
A	2							A					
M	3							M					
B	4							B					
O	5							O					
L	6							L					

در ماتریس cost در خانه هاشور زده شده ($\text{cost}[4][4]$) باید عدد صفر را قرار دهیم زیرا B (سطر چهارم) و B (ستون چهارم) برابرند. ولی در خانه شطرنجی ($\text{cost}[2][2]$) عدد یک را قرار می‌دهیم، زیرا A (سطر دوم) و U (ستون دوم) برابر نیستند.

ماتریس d							ماتریس COST						
		G	U	M	B	O							
		0	1	2	3	4	5		G	U	M	B	O
G	1	0	1	2	3	4		G	0	1	1	1	1
A	2	1	1	2	3	4		A	1	1	1	1	1
M	3	2	2	1	2	3		M	1	1	0	1	1
B	4	3	3	2	1	2		B	1	1	1	0	1
O	5	4	4	3	2	1		O	1	1	1	1	0
L	6	5	5	4	3	2		L	1	1	1	1	1

البته باید توجه داشته باشید که ماتریس cost در الگوریتم ساخته نمی‌شود، بلکه در حین کار هربار یک خانه از آن محاسبه می‌شود، ولی برای فهم بهتر الگوریتم این ماتریس را آورده‌ایم.

در خانه آخر عدد دو قرار گرفته است، بنابراین برای تبدیل این دو رشته به یک دیگر به دو حرکت نیاز داریم: الف) تبدیل A به U ب) افزودن L

۵-۲- الگوریتم‌های انطباق رشته مورد استفاده

برای انطباق رشته در این برنامه از چند الگوریتم انطباق رشته استفاده شده، یکی از آنان را که الگوریتمی معروف و عمومی بود (Levenshtein)، در بالا توضیح دادیم. به جز این الگوریتم، چند الگوریتم ابداعی دیگر نیز مورد استفاده قرار گرفته است، البته این ابداع از روی اختیارات شاعری صورت گرفته.

۵-۲-۱- الگوریتم XOR

ابتدا XOR کردن دو نویسه را این‌گونه تعریف می‌نماییم: «اگر دو نویسه یکسان بودند، یک و در غیر این صورت صفر برگردان» با این تعریف در رشته Udash اگر هر دو نویسه «U» و یا «-» بودند، یک و اگر یکی «U» و دیگری «-» بود صفر برمی‌گرداند.

برای آنکه XOR را بر روی دو رشته انجام دهیم، باید XOR تعریف شده در بالا را بر روی تک تک نویسه‌های دو رشته اعمال کنیم، یعنی نویسه i ام رشته اول با نویسه i ام رشته دوم مقایسه شود. البته این مقایسه باید تا طول کوچکترین رشته انجام شود. یعنی اگر یک رشته طولی برابر ۱۵ و دیگری طولی برابر ۱۷ داشت، باید مقایسه تا نویسه پانزدهم انجام شود. اختلاف طول را به دست آورده (مثلاً اختلاف دو رشته ۱۷ و ۱۵ نویسه‌ای را برابر ۲ قرار می‌دهیم) و در هر بار مقایسه خروجی‌های مقایسه را اختلاف طول جمع می‌نماییم. در آخر مقدار به دست آمده از طول بزرگترین رشته کم کرده و بر طول بزرگترین رشته تقسیم می‌نماییم.

در فرمول بالا base برابر طول رشته بلندتر است. distance هم از جمع اختلاف طول با نتیجه XOR هر یک از نویسه‌ها به دست می‌آید. برای فهم بهتر به مثال زیر توجه نمایید:

رشته اول	-	U	-	-	-	U	U	U	-	U	-	U	U	U	U	-	16
رشته دوم	U	U	-	-	U	-	U	U	-	U	-	-	U	U			14
XOR	1	0	0	0	1	1	0	0	0	0	0	1	0	0			4

در مثال بالا رشته اول طولی برابر ۱۶ و رشته دوم طولی برابر ۱۴ دارد. بنابراین base را ۱۶ مقداردهی می‌نماییم. در ابتدا distance برابر با اختلاف طول دو رشته یعنی ۲ قرار می‌دهیم. خروجی‌های XOR هر یک از مراحل را با هم جمع نموده (۴) و سپس آنها را با distance جمع می‌نماییم ($4+2=6$). عدد به دست آمده (۶) را از base کم کرده ($16-6=10$) و بر base تقسیم می‌نماییم. ($10/16 = 0/625$) عدد به دست آمده ($0/625$) برازندگی رشته اول نسبت به رشته دوم را نشان می‌دهد. (و یا بالعکس)

توجه کنید که این الگوریتم به صورتی که گفته شد، مورد استفاده قرار نخواهند گرفت، بلکه از این الگوریتم برای اعمال برخی از اختیارات شاعری که نمی‌توان به صورت مستقیم در برنامه قرارداد استفاده می‌شود.

۵-۲-۲ XOR برای اعمال اختیار شاعری قلب

شاعر مختار است در برخی اوزان به جای «U-»، «U-» بیاورد، و یا بالعکس عمل کند. این عمل را در اصطلاح «قلب» گویند [۱]. به دلیل آنکه به راحتی نمی‌توان مشخص کرد که آیا شاعر از این اختیار استفاده کرده یا اگر استفاده کرده، در کجای شعر از این اختیار استفاده کرده، نمی‌توان این اختیار را مستقیماً در برنامه اعمال کرد. بنابراین باید از یک تابع برازندگی با وزنی مخصوص برای این کار استفاده کرد. به این منظور از الگوریتم XOR که در بالا شرح داده شد، با تعییر ذیل استفاده می‌نماییم:

برای هر نویسه اگر پاسخ برابر یک شد، بررسی که آیا نویسه‌ی بعدی هم (به صورت معکوس) یک می‌شود. یعنی آیا دو نویسه به صورت «U-» و «U-» در دو رشته قرار دارند و یا به طرز دیگری. اگر به این صورت باشد، خروجی برای هر دو نویسه صفر خواهد شد. محاسبه برازندگی هیچ تفاوتی با XOR نمی‌کند. برای فهم بهتر به مثال زیر توجه کنید:

رشته اول	-	U	-	-	-	U	U	U	-	U	-	U	U	U	U	-	16
رشته دوم	U	U	-	-	U	-	U	U	-	U	-	-	U	U			14
XOR	1	0	0	0	0	0	0	0	0	0	0	1	0	0			2

بخش هاشور خرده این مورد را توضیح می‌دهد. بنابراین با این الگوریتم عدد به دست آمده برای distance برابر با ۴ خواهد شد. و نتیجه برابر ۰/۷۵ خواهد شد.

$$\frac{16 - 4}{16} = 0.75$$

۵-۲-۳ XOR برای اعمال اختیار شاعری تسکین

شاعر مختار است به جای دو هجای کوتاه (UU) یک هجای بلند بیاورد. عکس این مورد صحیح نمی‌باشد. این اختیار که در اصطلاح به آن «تسکین» می‌گویند، جز در آغاز مصراع (مگر به ندرت) در همه جا قابل اعمال است [۱]. در این اختیار شاعری هم مانند قبل به دلیل نامعلوم بودن وقوع و مکان وقوع، مجبور به استفاده از یک تابع برازندگی هستیم. در اینجا نیز باز از تابع XOR با تغییراتی استفاده می‌نماییم.

در صورتی که خروجی برای یک نویسه ۱ شد و نویسه رشته اول برابر «U» بود، بررسی می‌کنیم که آیا نویسه بعدی در رشته اول هم برابر «U» است. اگر انگونه بود خروجی را برابر با صفر می‌نماییم. برای فهم بهتر به مثال زیر توجه نمایید:

رشته اول	-	U	-	-	-	U	U	U	-	U	-	U	U	U	U	-	16
رشته دوم	U	U	-	-	U	-	U	U	-	U	-	-	U	U			14
XOR	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0		3

قسمت هاشورخورده به خوبی این مطلب را بیان می‌نماید. بنابراین با این الگوریتم عدد به دست آمده برای distance برابر با ۵ خواهد شد. و نتیجه برابر ۰/۶۸۷۵ خواهد شد.

$$\frac{16 - 5}{16} = 0.6875$$

۵-۳- ترکیب نتایج توابع برازندگی

همانطور که در بالا شرح داده شد، توابع برازندگی توضیح داده شده هر یک برای کاری مناسب هستند، ولی به‌طور قطع نمی‌توان یکی را بر دیگری ترجیح داد. بنابراین باید به گونه‌ای برازندگی‌های تولید شده از هر یک از این توابع را با یکدیگر ترکیب کرد^۱.

۵-۳-۱- اولویت

برای ترکیب برازندگی‌ها و به دست آوردن یک برازندگی مناسب، داشتن یک مقدار اولویت^۲ ضروری به نظر می‌رسد. البته این مساله با توجه به نوع خاص مسئله و وجود اولویت در بین اوزان به راحتی قابل استخراج است.

طبق بررسی‌های انجام شده بر روی اوزان اشعار فارسی از ابتدا تا کنون، محققان هشت کد وزن پرکاربرد فارسی را به ترتیب زیر بیان نموده‌اند[۳]:

1002، 1003، 1004، 1001، 1009، 1005، 1008، 1006

^۱ Combine Result

^۲ Priority

بنابراین ما هم به این اوزان بر همین ترتیب اولویت بالاتری می‌دهیم. به اولی (۱۰۰۲) اولویت ۱۰۰ به بعدی ۹۹ و به همین ترتیب تا ۹۳، به بقیه اوزان موجود هم اولویت ۹۰ می‌دهیم. این اولویت‌ها کمک می‌کند در مواردی که دو برازندگی نزدیک داشتیم، به یک وزن پرکاربرد نزدیک‌تر شویم.

۵-۳-۲- میانگین وزنی

برای ترکیب دو یا چند برازندگی از میانگین وزنی استفاده خواهیم کرد. به این صورت که برای کنترل تاثیر هر یک از توابع برازندگی به هر یک وزنی داده و با زیاد کردن آن، تاثیر آن تابع را بیشتر و با کم کردن وزن آن، تاثیر تابع را کم می‌نماییم. برای این کار از فرمول زیر استفاده می‌نماییم:

$$\text{fitness}_j = \frac{\sum_{i=1}^n \text{fitness}_{ij} \times \text{weight}_i}{\sum_{i=1}^n \text{weight}_i} \times \text{priority}_j$$

این تابع میزان برازندگی (fitness) رشته Udash به الگوی ژام (وزن ژام) را محاسبه می‌کند. برای این کار، ابتدا به وسیله هر یک از توابع شباهت (اگر n تابع شباهت وجود داشته باشد) برازندگی رشته Udash به الگوی ژام محاسبه می‌شود (fitness_{ij}) سپس این مقدار که عددی بین صفر و یک است، در وزنی که به این تابع داده شده ضرب می‌شود. این مقادیر با یکدیگر جمع شده و سپس تقسیم بر وزن‌ها می‌شوند. این اوزان در برنامه‌ای جداگانه به دست می‌آیند. در آخر مقدار محاسبه شده، در اولویت الگو (priority_j) که بر اساس میزان تکرار این الگو در اوزان فارسی مقدار گذاری شده- ضرب می‌شود. این کار برای تمامی الگوها انجام می‌شود و الگویی که بالاترین عدد را کسب کند، به عنوان برنده اعلام می‌شود.

۵-۳-۳- به دست آوردن وزن هر برازندگی

باید توجه داشت که نمی‌توان برازندگی‌ها را با هم جمع نموده و تقسیم بر تعداد بنماییم. زیرا قدرت تفکیک بعضی از این توابع بیشتر از دیگری است و همچنین بدون وجود هر یک از این برازندگی‌ها قدرت تشخیص برنامه پایین می‌آید. بنابراین برنامه‌ای نوشته شد که با تغییر وزن هر یک از

برازندگی‌ها بهترین جواب را به دست آورد. برای به دست آوردن وزن توابع (آموزش)^۱، مجموعه داده (Data Set) را به دو بخش تقسیم کردیم. ۷۰۰ بیت را برای به دست آوردن وزن و ۳۰۰ بیت دیگر برای آزمون^۲ برنامه به کار بردیم.

برای به دست آوردن وزن‌ها در این‌گونه موارد الگوریتم ژنتیک پیشنهاد می‌شود. ولی ما در این پژوهش از راه ساده‌تری این کار را انجام دادیم. گرچه می‌توان در پژوهش‌های بعدی از الگوریتم ژنتیک هم استفاده کرد. برای این کار ما ابتدا حدود وزن مناسب هر یک از برازندگی‌ها را با آزمایش دستی به دست آوردیم، سپس با استفاده از چند حلقه (for) تودرتو ۷۰۰ بیت را با وزن مختلف برای برازندگی‌های مختلف آزمایش نمودیم و بالاترین درصد پاسخ (۶۵/۲ درصد) را انتخاب نموده و وزن‌هایی که با آنها این درصد به دست آمده به عنوان وزن برازندگی‌ها انتخاب کردیم. برای آزمایش درستی این وزن‌ها، برنامه را با وزن‌های به دست آمده برای ۳۰۰ بیت دیگر آزمایش کردیم که نتیجه نزدیک به همان نتیجه قبل شد. (۶۵/۵ درصد)

وزن‌های به دست آمده برای سه تابع برازندگی به صورت زیر می‌باشند:

وزن	تابع برازندگی
۹	Levenshtein
۴	XOR برای اعمال اختیار شاعری قلب
۳	XOR برای اعمال اختیار شاعری تسکین

۵-۴- در این فصل...

در این فصل توضیحاتی درباره‌ی الگوریتم‌های انطباق رشته و اهداف و قابلیت‌های آنها داده شد. همچنین درباره الگوریتم‌های مورد استفاده در این پژوهش، عملکرد و دلیل استفاده از آنان توضیح داده

^۱ Train

^۲ Test

شد. در آخر روش ترکیب این الگوریتم‌ها و همچنین چگونگی استفاده از آنها در این طرح را شرح دادیم.

پس از پایان یافتن مراحل کار، نوبت به آزمون برنامه می‌رسد. در فصل بعد آزمون‌هایی که بر روی برنامه انجام شده و همچنین نتایج این آزمون‌ها و به همراه تحلیل این نتایج خواهد آمد. آن جا دلیل برخی نتایج ضعیف توضیح داده خواهد شد.

فصل ششم

آزمون

۶- آزمون

پس از تکمیل مراحل کار و همچنین تعیین وزن، نوبت به امتحان نهایی برنامه رسید برای این کار ۱۰۰۰ بیتی که برای آموزش و آزمون به کار برده بودیم و همچنین تعداد دیگری بیت شعر، برای امتحان نهایی آماده کردیم. برای امتحان نهایی برنامه حدود ۱۳۰۰ بیت شعر آماده شد. از ۱۲۹۷ بیت شعر داده شده، وزن ۸۵۱ بیت صحیح تشخیص داده شده، و ۴۴۶ بیت را به غلط تشخیص داده‌ایم. یعنی برنامه ۶۵/۶۱ درصد توانسته است جواب بدهد.

۶-۱- نتیجه نهایی به تفکیک وزن

برای آنکه بتوان بهتر نتایج را تحلیل کرد و همچنین بتوان روشی برای بهبود نتایج پیدا کرد، هر وزن را به طور جداگانه بررسی کرده تا ببینیم در هر وزن چه درصدی توانسته‌ایم نتیجه بگیریم.

در مورد جدول زیر باید نکاتی را توضیح دهیم. ستون اول کد وزن است که می‌توانید وزن هر یک از این کدها را در پیوست ۳ ببینید. ستون دوم تعداد هر یک از این اوزان در مجموعه داده (Data Set) می‌باشد. ستون سوم تعداد تشخیص صحیح این وزن را مشخص می‌نماید، مثلاً در مورد کد ۱۰۰۲، از ۱۲۸ تعداد بیت در مجموعه داده تعداد ۱۱۸ بیت صحیح تشخیص داده شده، و تعداد ۱۰ بیت نیز غلط تشخیص داده شده. بنابراین ۹۲/۱۹ درصد بیت این وزن صحیح تشخیص داده شده. ستون آخر به معنای تشخیص اوزان دیگر به این کد است. مثلاً در مورد کد ۱۰۰۲، تعداد ۴۸ بیت از مجموع ۱۲۹۷ بیت به غلط کد ۱۰۰۲ تشخیص داده شده‌اند. در آخرین سطر هم به مجموع ارقام اشاره شده. البته در مورد درصد تشخیص باید گفته شود که این مقدار (۵۶/۵۸) میانگین درصد تشخیص صحیح می‌باشد. درست است که در کل ۶۵/۶۸ درصد جواب گرفته‌ایم ولی به طور میانگین تنها ۵۶/۵۸ درصد هر وزن درست تشخیص داده شده، این مورد بهتر می‌تواند اطمینان به ما دهد که اگر وزنی را این سیستم تشخیص داد، چقدر می‌توان به آن اطمینان کرد.

جدول آماری صحت تشخیص هر وزن

کد وزن	تعداد در مجموعه داده	تعداد تشخیص صحیح	تعداد تشخیص غلط این وزن	درصد تشخیص صحیح	تعداد تشخیص غلط با این وزن
۱۰۰۱	۱۰۳	۶۰	۴۳	۵۸.۲۵	۷۲
۱۰۰۲	۱۲۸	۱۱۸	۱۰	۹۲.۱۹	۴۸
۱۰۰۳	۱۳۱	۹۱	۴۰	۶۹.۴۷	۴۹
۱۰۰۴	۷۹	۷۲	۷	۹۱.۱۴	۴۶
۱۰۰۵	۹۵	۸۵	۱۰	۸۹.۴۷	۱۲
۱۰۰۶	۷۹	۶۶	۱۳	۸۳.۵۴	۱۱
۱۰۰۷	۳۰	۲۰	۱۰	۶۶.۶۷	۱۰
۱۰۰۸	۶۸	۳۱	۳۷	۴۵.۵۹	۳۴
۱۰۰۹	۲۹	۲۵	۴	۸۶.۲۱	۳۱
۱۰۱۰	۱۱۲	۷۲	۴۰	۶۴.۲۹	۹
۱۰۱۱	۱۸	۹	۹	۵۰	۱
۱۰۱۲	۴۰	۳۱	۹	۷۷.۵	۳۱
۱۰۱۳	۱۵	۱۳	۲	۸۶.۶۷	۱
۱۰۱۴	۱۶	۰	۱۶	۰	۰
۱۰۱۵	۱۵	۱۳	۲	۸۶.۶۷	۱
۱۰۱۶	۱۳	۱۱	۲	۸۴.۶۲	۵
۱۰۱۷	۱۸	۱۴	۴	۷۷.۷۸	۰
۱۰۱۸	۶۱	۲۵	۳۶	۴۰.۹۸	۰
۱۰۱۹	۸	۲	۶	۲۵	۹
۱۰۲۰	۱۳	۹	۴	۶۹.۲۳	۱۴
۱۰۲۱	۹	۳	۶	۳۳.۳۳	۲
۱۰۲۲	۱۱	۶	۵	۵۴.۴۴	۱۱
۱۰۲۳	۲۳	۱۴	۹	۶۰.۸۷	۳
۱۰۲۴	۳۲	۹	۲۳	۲۸.۱۳	۴
۱۰۲۵	۲۲	۱	۲۱	۴.۵۵	۰
۱۰۲۶	۲۷	۴	۲۳	۱۴.۱۸	۷
۱۰۲۷	۱۶	۷	۹	۴۳.۷۵	۱۸
۱۰۲۸	۲۳	۸	۱۵	۳۴.۸۷	۳
۱۰۲۹	۲۷	۲۳	۴	۸۵.۱۹	۷
۱۰۳۰	۱۸	۰	۱۸	۰	۰
۱۰۳۱	۱۸	۱۰	۸	۵۵.۵۶	۰
مجموع	۱۲۹۷	۸۵۱	۴۴۶	۵۶.۵۸	۴۴۶

۶-۱-۱- تحلیل آمار به دست آمده

اگر خوب به جدول بالا دقت کنید، می‌توانید اوزانی را ببینید که نتیجه بسیار بدی را تولید کرده‌اند. کد وزن‌های ۱۰۳۰ و ۱۰۱۴ هیچ بیت درست تشخیص داده شده ندارند و در کد ۱۰۲۵ فقط ۴ درصد جواب گرفته‌ایم. علت چیست؟

مطلبی که در وهله اول به نظر می‌رسد که علت اصلی این امر باشد، نزدیکی اوزان به یکدیگر است که جلوتر به این امر بیشتر خواهیم پرداخت. وزن ۱۰۳۰ بسیار شبیه به وزن ۱۰۰۴ است.

-U---U---U---U- 1004

-U---U---U---U-- 1030

همانطور که می‌بینید، این دو وزن فقط در یک «-» در آخر با هم تفاوت دارند، بنابراین طبیعی است که با یکدیگر اشتباه گرفته شوند. و البته به خاطر ساختار برنامه، وزن ۱۰۰۴ که بالاتر از ۱۰۳۰ است به عنوان وزن صحیح تشخیص داده می‌شود. و همانطور که در جدول بالا می‌بینید، ۴۸ بار وزن‌های دیگر به عنوان وزن ۱۰۰۴ تشخیص داده شده‌اند. و در آمار دیگری که تهیه شد، مشخص گردید که از ۱۸ بیتی که وزن آنها ۱۰۳۰ بوده، ۱۱ مورد ۱۰۰۴ تشخیص داده شده‌اند، که این به معنی تایید حدس ما می‌باشد.

و همین‌گونه است، وزن ۱۰۱۴ با وزن ۱۰۰۲. جلوتر وزن‌های شبیه به هم را بیان خواهیم کرد.

۶-۱-۲- پیشنهادات بر اساس آمار به دست آمده

باتوجه به نزدیکی بعضی اوزان، به نظر می‌رسد، قرار دادن اوزان شبیه به هم در یک دسته و اعلام یک گروه وزن، به عنوان وزن تشخیص داده شده، هم باعث بهبود نتایج شده و هم باعث اطمینان بیشتر به برنامه.

مساله دیگر که می‌توان به آن توجه کرد و در پژوهش‌های بعدی از آن برای بهبود نتایج استفاده کرد، استفاده از ساختار سلسله مراتبی می‌باشد. به این صورت که با استفاده از یک تابع تخمین، ابیات را در یک دسته وزن جا داد، سپس با استفاده از یک تابع تخمین دیگر سعی کرد بیت را در یک زیر دسته کوچکتر جا داد و بالاخره با یک تابع تخمین دقیق‌تر وزن بیت را در زیر دسته تشخیص داد. این کار مستلزم دسته‌بندی اوزان بر اساس کارکرد هر تابع تخمین و همچنین انجام یک کار آماری گسترده برای به دست آوردن اشتباهات هر تابع تخمین می‌باشد.

۶-۲- مقایسه اوزان با یکدیگر

کار آماری بسیار خوبی که می‌تواند به بهبود برنامه در پژوهش‌های آینده کمک کند، مقایسه وزن‌های موجود بر اساس توابع تخمین می‌باشد. در این قسمت ما هر بار یکی از اوزان را به توابع تخمین به عنوان ورودی داده‌ایم تا توابع تخمین، برازندگی هر وزن نسبت به اوزان دیگر را بسنجند، مسلماً باید هر وزن بیشترین برازندگی را با خود داشته باشد، که این امر محقق شده است. ولی مسئله مهم این است که هر وزن با کدام وزن‌ها برازندگی بیشتری دارد. این تشخیص می‌تواند به دسته بندی صحیح، کمک بسیاری بکند.

به عنوان مثال در بالا به صورت دستی دیدیم که وزن ۱۰۰۴ که به ۱۰۳۰ بسیار نزدیک است. و بر اساس توابع برازندگی هم می‌توان این مساله را نشان داد. برنامه وزن ۱۰۰۴ را ۹۳.۷۵ به وزن ۱۳۰ نزدیک می‌داند و برعکس.

۶-۲-۱- اوزان نزدیک به هم

در این قسمت با توجه به توابع شباهت، اوزان نزدیک به هم، همراه با درصد نزدیکی را ارائه خواهیم داد:

جدول اوزان نزدیک به هم

کد وزن ۱	کد وزن ۲	درصد شباهت	تفاوت
۱۰۰۱	۱۰۲۵	۹۳.۷۵	یک «-» اضافه در آخر
۱۰۰۲	۱۰۱۴	۹۳.۷۵	یک «-» اضافه در آخر
۱۰۰۴	۱۰۳۰	۹۳.۷۵	یک «-» اضافه در آخر
۱۰۱۵	۱۰۲۲	۹۰.۹۱	یک «-» اضافه در آخر
۱۰۲۲	۱۰۲۳	۹۲.۲۸	یک «-» اضافه در آخر

همانطور که در بالا مشاهده می‌نمایید، تفاوت در این ۸ وزن فقط داشتن یک هجای بلند در آخر است، و این می‌تواند به ساخت تابع شباهتی برای جدا کردن اوزان بر اساس هجای آخر منجر شود. البته باید توجه شود، شباهت‌های آورده شده، فقط به بالای ۹۰ درصد تعلق دارد، وگرنه بعضی دیگر از اوزان نیز تا ۸۰ درصد به هم نزدیک هستند که می‌توان علل آن را نیز بررسی کرد. مثلاً کد وزن ۱۰۰۹ و ۱۰۲۶ که ۸۴ درصد شبیه هستند، فقط در یک هجای بلند، (البته در وسط رشته) با هم تفاوت دارند.

۶-۲-۲- مقایسه اوزان با نزدیکترین وزن

در این بخش تمامی اوزان را با نزدیکترین وزن و همچنین با دیگر اوزان مقایسه می‌شود. در هر وزن نزدیکترین وزن به آن و درصد نزدیکی آمده است. همچنین میانگین درصد شباهت با وزن‌های دیگر را محاسبه کرده‌ایم. بدیهی است که هرچه میانگین شباهت کمتر باشد، آن وزن از بقیه اوزان دورتر است و همچنین انحراف معیار هرچه بالاتر باشد، دوری بیشتری را نسبت به هر وزن داریم. باید سعی در این باشد که تابعی برای شباهت انتخاب شود که میانگین شباهت را پایین آورده و انحراف معیار را افزایش دهد.

در صورتی که میانگین پایین باشد ولی انحراف معیار به اندازه کافی بزرگ نباشد، بدان معنی است که وزن ما به بعضی اوزان بسیار نزدیک و از بعضی دیگر دور است. در آخر نیز باید میانگین شباهت اوزان به یکدیگر تا حد امکان پایین بیاید، باتوابع شباهتی که ما استفاده کرده‌ایم و همچنین وزن انتخابی برای هر تابع شباهت ۶۵.۶۷ درصد شباهت بین اوزان وجود دارد.

جدول نزدیکی اوزان به یکدیگر و درصد شباهت

کد وزن	نزدیکترین وزن	درصد شباهت	میانگین شباهت	انحراف معیار
۱۰۰۱	۱۰۲۵	۹۳.۷۵	۶۷.۸۱	۹.۵
۱۰۰۲	۱۰۱۴	۹۳.۷۵	۶۷.۲۸	۶.۹۸
۱۰۰۳	۱۰۲۸	۸۰.۲۹	۶۷.۰۵	۶.۰۹
۱۰۰۴	۱۰۳۰	۹۳.۷۵	۶۶.۴۱	۸.۸۵
۱۰۰۵	۱۰۳۰	۸۳.۸۵	۶۴.۲۵	۸.۲۴
۱۰۰۶	۱۰۱۰	۷۹.۰۲	۶۲.۹	۵.۸۴
۱۰۰۷	۱۰۳۱	۸۴.۶۲	۶۶.۳۲	۷.۳۴
۱۰۰۸	۱۰۱۲	۷۹.۶۷	۶۸.۲۱	۷.۶۲
۱۰۰۹	۱۰۲۶	۸۴.۶۲	۶۳.۰۵	۷.۵۱
۱۰۱۰	۱۰۲۱	۸۰.۴۲	۶۲.۹	۶.۳۸
۱۰۱۱	۱۰۱۹	۸۱.۲۵	۶۶.۲۲	۶.۹۵
۱۰۱۲	۱۰۰۸	۷۷.۴۷	۶۴.۹۸	۶.۶۳
۱۰۱۳	۱۰۱۱	۸۲.۶۹	۶۷.۲۵	۶.۷۶
۱۰۱۴	۱۰۰۲	۹۳.۷۵	۶۷.۰۹	۷.۷۵

۸.۴۵	۶۱.۰۸	۹۰.۹۱	۱۰.۲۹	۱۰.۱۵
۷.۵۱	۶۱.۲	۷۸.۴۶	۱۰.۱۵	۱۰.۱۶
۵.۰۱	۶۷.۰۹	۷۷.۴۷	۱۰.۱۹	۱۰.۱۷
۸.۷۹	۶۴.۲۲	۸۶.۰۶	۱۰.۰۴	۱۰.۱۸
۶.۳۸	۶۹.۴۴	۸۲.۲۱	۱۰.۲۴	۱۰.۱۹
۶.۶	۶۴.۷۶	۸۱.۳۲	۱۰.۰۷	۱۰.۲۰
۶.۶۷	۶۴.۹۹	۸۰.۴۲	۱۰.۱۰	۱۰.۲۱
۵.۵۵	۶۷.۲۸	۹۱.۶۷	۱۰.۲۳	۱۰.۲۲
۶.۰۲	۶۴.۷۳	۹۱.۶۷	۱۰.۲۲	۱۰.۲۳
۸.۲۳	۶۷.۰۶	۸۲.۲۱	۱۰.۰۱	۱۰.۲۴
۸.۹۵	۶۸.۱۵	۹۳.۷۵	۱۰.۰۱	۱۰.۲۵
۸.۰۶	۶۲.۴۷	۸۴.۶۲	۱۰.۰۹	۱۰.۲۶
۷.۱۹	۶۷.۶۳	۸۴.۶۲	۱۰.۲۵	۱۰.۲۷
۶.۷۷	۶۶.۶۶	۸۱.۲۵	۱۰.۰۱	۱۰.۲۸
۷.۵۲	۶۲.۵۵	۹۰.۹۱	۱۰.۱۵	۱۰.۲۹
۹.۳۶	۶۵.۹۹	۹۳.۷۵	۱۰.۰۴	۱۰.۳۰
۶.۲۶	۶۶.۸۴	۸۴.۶۲	۱۰.۰۷	۱۰.۳۱
۷.۲۳	۶۵.۶۷	۸۵.۹۶	---	میانگین

۶-۳- در این فصل...

در فصلی که گذشت، آزمون‌هایی که بر روی برنامه انجام شده به همراه نتیجه آنها آورده شد. نتایج به صورت کلی و به تفکیک وزن شرح داده شدند. همچنین بر اساس این آمار تحلیل‌هایی انجام شد و علت ضعیف بودن برخی نتایج توجیه شد. با بررسی وزن‌های نزدیک به هم این تحلیل‌ها روشن‌تر شدند.

پس از روشن شدن مساله، روش حل مساله و آزمون برنامه، نوبت به نتیجه‌گیری و جمع‌بندی مطالب گفته شده می‌رسد. فصل بعد به این موضوع می‌پردازد.

فصل هفتم

نتایج

۷- نتایج

در فصل‌های پیشین به شرح آنچه می‌خواهیم انجام دهیم و آنچه انجام شد، پرداختیم. به تشریح علم عروض و نیاز به برنامه‌ای برای به دست آوردن وزن عروضی پرداختیم. پیش‌نیازهای طرح را بررسی کرده و برای رفع این نیازها روش‌هایی را ارائه دادیم. الگوریتمی برای مسئله پیشنهاد دادیم و مراحل شش‌گانه آن را به صورت مبسوط شرح دادیم. پس از پایان یافتن مراحل، آزمونی برای سیستم طراحی کرده و نتایج آنها را تحلیل کردیم. و اکنون پس از طی این مراحل باید آنچه انجام داده‌ایم را جمع‌بندی نموده و راه را برای ادامه این پژوهش روشن نماییم.

۷-۱- عروض

شاعری را که حتی یک بیت شعر ترکی نگفته در دنیا به عنوان شاعری ترک می‌شناسند و اشعارش به جای زبان مادری از ترکی به زبان‌های دیگر دنیا ترجمه می‌شود. اگر جمله قبل برای شما آشنا و دردآور باشد، به اهمیت مطالعه در زمینه ادبیات فارسی واقف هستید. مولانا جلال الدین حتی یک بیت شعر به زبان ترکی ندارد، لقب رومی چند قرن بعد به او داده شد. زمانی که بر روی چنین مسالهی واضحی باید بحث کرد و قسم خورد که مولانا شاعری فارسی زبان است، چه انتظاری از مسائل دیگر می‌توان داشت؟ عروض هم یکی از این مسائل است.

هم اکنون تقریباً تمامی اساتید ادبیات بر این نکته اتفاق نظر دارند که عروض سنتی دارای اشکالات فراوانی است و باید روشی نو برای احیای این علم ساخت. البته کارهای بسیار بزرگی برای احیای این علم انجام شده ولی تا کنون هیچ کاری برای رایانه‌ای کردن این علم انجام نشده. با توجه به اهمیت این علم در تحقیقات ادبی، باید این علم را از علمی سحری به علمی عمومی دریاوریم.

در این بخش بر آن نیستیم که به تشریح علم عروض بپردازیم، برای اطلاع از این علم و ریزه‌کاری‌های آن می‌توانید به کتب عروض مراجعه نمایید، به علاوه آن که در فصل دوم هم توضیحاتی مختصر در این باب داده شده. در این بخش فقط می‌خواستیم بگوییم که عروض هرچند علمی است قدیمی، ولی چون پیرمرد صد ساله‌ای که تازه دندان درمی‌آورد، کودکی است که نیاز به مراقبت دارد تا از نو شروع به رشدی کند.

۷-۲- اهمیت و لزوم وجود این سیستم

هرچند تلاش فراوانی برای آسان‌سازی عروض انجام شده - که انصافا بسیار هم موثر بوده است - ولی هنوز هم به دست آوردن وزن عروضی در برخی ابیات برای بعضی اساتید هم مشکل‌زا و مورد اختلاف است. همچنین نمی‌توان به سرعت یک کامپیوتر وزن عروضی یک دیوان شعر را به دست آورد. این سیستم سعی دارد این نواقص را حل کند، هرچند که در ابتدای راه است و نیاز به تحقیقات بیشتری دارد.

همانگونه که در فصل پیش هم گفتیم ما توانسته‌ایم بیش از ۶۵ درصد جواب بگیریم. و این بدان معنی است که شما می‌توانید در حدود ۶۵ درصد احتمال دهید که جواب داده شده به شما صحیح است. در ادامه به تشریح پیاده‌سازی این سیستم خواهیم پرداخت.

در زیر شمایی از پیاده‌سازی این سیستم را می‌توانید ببینید.

صفحه نخست | آزمون | مقایسه | کشف وزن | ابزار

وبگاه آزمایشی عروض

میت راب تعلیک مصرع بدون حرکت گذاری وارد نمایید:

مصرع اول:

مصرع دوم:

تایید

میت راب تعلیک مصرع بدون حرکت گذاری وارد نمایید: (با استفاده از مجموعه کلمات)

مصرع اول:

مصرع دوم:

تایید

میت راب تعلیک مصرع همراه حرکت گذاری وارد نمایید:

مصرع اول:

مصرع دوم:

تایید

صامت و مصوت را به صورت CV وارد نمایید:

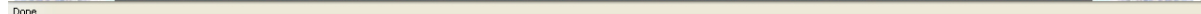
تایید

شکل های چهار اواورد نماید:

تایید

صفحه نخست | آزمون | مقایسه | کشف وزن | ابزار

پس از وارد کردن بیت شعر، برنامه شش مرحله‌ای را که در فصل سوم و چهارم شرح دادیم، انجام خواهد داد و نتیجه را به صورت زیر اعلام می‌نماید:



برنامه رشته cv و Udash هریک از مصراع‌ها را جداگانه نمایش می‌دهد و سپس مرحله ششم را اجرا کرده و وزنی را به عنوان برنده اعلام می‌دارد. سپس برنامه مشخصات وزن تشخیص داده شده (برنده) به همراه درصد برازندگی و توضیحات مربوط به وزن را از پایگاه داده خوانده و به کاربر نمایش می‌دهد.

در مثالی که مشاهده کردید، ما بیت زیر را از طریق گزینه سوم به برنامه داده‌ایم:

دارم از زلف سیاهش گله چندان که می‌رس که چنان ز او شده ام بی سر و سامان که می‌رس

برنامه هر مصراع را جداگانه به رشته cv و Udash تبدیل کرده و در جدولی نمایش داده است. سپس وزن تشخیصی را که «فعلاتن فعلاتن فعلتن فعلن» می‌باشد را به همراه نام وزن، نمونه‌ای شعر از این وزن، رشته Udash این وزن و درصد برازندگی نمایش می‌دهد. بر طبق نظر خبره وزن تشخیص داده شده در مثال ما صحیح است. همچنین ۹۰/۰۴ درصد برازندگی می‌تواند به ما اطمینان بیشتری دهد که وزن تشخیصی صحیح است.

البته برای آنکه بتوان برنامه را آزمایش نمود و همچنین برای کارهای تخصصی‌تر گزینه‌های دیگری هم قرار داده‌ایم. کاربر می‌تواند رشته CV را خود به دست آورد و سپس از برنامه بخواهد وزن این رشته را تشخیص دهد. در صورتی که از این گزینه استفاده شود، مراحل اول و دوم انجام نشده و رشته ورودی مستقیماً به مرحله سوم یعنی اصلاح رشته صامت-مصوت داده شده و تا مرحله‌ی ششم رشته را پردازش می‌کند. همچنین کاربر می‌تواند یک رشته Udash به برنامه دهد، که در این صورت برنامه فقط مراحل پنجم و ششم را انجام خواهد داد.

تمامی پیاده‌سازی برنامه تحت وب بوده و بزودی بر روی اینترنت قرار داده خواهد شد. البته به دلیل سربار زیاد آزمون برنامه برای چندین بیت شعر، و برخی قسمت‌های توضیح داده شده، از قرار دادن این بخش در وب به صورت عمومی معذور می‌باشیم. ولی برای آنکه محققان بتوانند از نتایج برنامه استفاده لازم را ببرند، مجموعه داده آزمایش شده به همراه فایل‌های تولیدی بر روی وبسایت قرار داده خواهد شد.

۷-۳- پیشنهادات

همانطور که در مقدمه گفته شد، ادبیات زمینه‌ای بکر -و شاید مهجور- می‌باشد و تحقیقات کامپیوتری در این زمینه به ندرت انجام می‌شود. این پژوهش که اولین پژوهش کامپیوتری در زمینه عروض می‌باشد، بسیار ناقص و در ابتدای کار می‌باشد، توسعه آن مستلزم صرف زمان و انرژی بیشتری است و شاید بدون همکاری اساتید ادبیات و کامپیوتر نتواند به راه خود ادامه پیدا کند.

نتایجی که از این پژوهش در این مدت کم به دست آمد، نشان می‌دهد که می‌توان با صرف زمان بیشتری نتایج بسیار بهتری کسب کرد. برای توسعه و بهبود این پژوهش می‌توان کارهایی از قبیل کارهای ذیل انجام داد:

- ۱- استفاده از الگوریتم ژنتیک برای به دست آوردن وزن توابع شباهت
- ۲- استفاده و آزمایش دیگر توابع شباهت رشته
- ۳- استفاده از ساختار سلسله مراتبی برای به دست آوردن وزن‌ها
- ۴- استفاده از ساختاری دیگر مانند شبکه عصبی برای یافتن شباهت
- ۵- افزودن تمامی ۳۰۰ وزن فارسی

فصل هشتم

پیوست‌ها

۸- پیوست‌ها

۸-۱- آشنایی با طبقه بندی زبان های جهان و زبان اوستایی

زبان شناسان زبان های جهان را به سه دسته ی زیر تقسیم کرده اند:

۱- زبان های یک هجائی (تک هجایی)

در این زبان ها کلمات فقط از یک هجای تغییر ناپذیر درست شده اند. یعنی نه ریشه ی کلمه تغییر می کند و نه بدان پیشوند یا پسوندی متصل می شود. این زبان ها در حالت ابتدایی باقی مانده اند، مثل: زبان های چینی، تبتی و سیامی

۲- زبان های الصاقی (پیوندی)

در این زبان ها کلمات از یک یا چند هجای تغییر ناپذیر درست شده اند، ولی می تواند به هم بچسبند یا پیشوند و پسوندی بگیرند تا معنی نویی به دست آید، مثل: زبان های فنلاندی، ترکی، مغولی، تاتاری، ژاپنی و کره ای

۳- زبان های منصرف (صرفی)

در این زبان ها کلمات از یک یا چند هجای تغییر پذیر درست شده اند. یعنی ریشه ی کلمه ها برعکس زبان های بالا تغییر می کند و به صورت های مختلف درمی آید. علاوه بر این پیوند پیشوند و پسوند با کلمه ها در این زبان ها فراوان است. زبان های مهم دنیا از این نوعند، مثل زبان های هندی، ایرانی، یونانی، لاتین، ژرمنی، اسلاوی، ارمنی و سامی.

الف- زبان های هندی و اروپایی

۱- زبان های هندی: سانسکریت، ودا، ...

۲- زبان های ایرانی: پارسی باستان (پارسی کهن - فرس قدیم)، اوستایی، پهلوی (زبان

ساسانیان)، پارتی (زبان اشکانیان)، دری، سُغدی، خوارزمی، سکایی، ...

پارسی باستان با زبان اوستای و سانسکرت از یک ریشه اند. اوستا زبانی است که برای نوشتن اوستا کتاب دینی زردشتیان ایران باستان به کار رفته است، به این سبب آن را زبان اوستایی نامند. پارسی باستان و اوستا فرزندان یک پدرند.

زبان های هند و ایرانی یکی از مهم ترین شاخه های زبان هند و اروپایی است و آریایی نامیده می شوند.

۳- زبان یونانی

۴- **زبان لاتین**: زبان لاتین زبان رومیان قدیم بوده است که امروزه جزو زبان های مرده ی دنیا محسوب می شود.

زبان های زیر مشتق از زبان لاتین هستند: فرانسه، ایتالیایی، اسپانیایی، پرتغالی و رومانی.

۵- **زبان های ژرمنی**: آلمانی، انگلیسی، هلندی، سوئدی، نروژی، ...

۶- **زبان های اسلاوی**: روسی، چکی، لهستانی، ...

ب- زبان های سامی

زبان های سامی عبارتند از: عربی، عبری، آشوری، فینیقی، کلدانی، سریانی، ...

زبان های سامی بالا با چند زبان نزدیک به آنها در شاخه ی بزرگ تری بنام زبان های سامی و حامی قرار می گیرند.

فارسی زبانی است که در ایران، افغانستان، تاجیکستان، قسمتی از هندوستان، ترکستان، قفقاز و عراق بدان صحبت می کنند. زبان فارسی به جا مانده ی زبان های قدیمی ایران است. این زبان ها عبارتند از:

۱- فارسی باستان

فارسی باستان (پارسی کهن) زبانی است که در زمان **هخامنشیان** بدان صحبت کرده اند و سنگ نبشته های بیستون، الوند و تخت جمشید بدان زبان است. خط این زبان را میخی نامند، زیرا شبیه به میخ است و از چپ به راست نوشته می شده است.

۲- فارسی میانه

فارسی میانه که آن را **پهلوی** هم می خوانند، به خصوص در زمان **اشکانیان** معمول بوده است و به دو قسمت می شود، یکی **پهلوی اشکانی (پارتی)** و دیگری **پهلوی ساسانی**. خط این زبان را نیز پهلوی نامند. پهلوی در شمال شرق ایران معمول بوده است و از راست به چپ نوشته می شده.

۳- فارسی دری

فارسی دری در زمان **ساسانیان** در مشرق و جنوب ایران به خصوص در پایتخت آنان مدائن معمول بوده است و پس از ورود اسلام به ایران به تدریج تغییراتی در آن داده شده است و به صورت زبان بعد از اسلام ایران در آمده و از نیمه دوم قرن سوم هجری تا به حال ادامه دارد.

چندی پس از ورود اسلام به ایران کم کم لغت‌های عربی به مقدار کم وارد زبان فارسی شدند، چنان که در قدیمی‌ترین کتاب‌های فارسی تعداد لغت‌های عربی از پنج درصد تجاوز نمی‌کند. ولی در دوره‌های بعد نویسندگان فارسی لغت‌های عربی فراوانی وارد زبان فارسی کردند، به حدی که در دوره‌های بعد در بعضی از کتاب‌های قدیمی تعداد آنها به بیش از شصت درصد رسید.

زبان اوستایی از شاخه‌ی شرقی زبان‌های ایران باستان است. اما ناحیه آن را بدرستی روشن نکرده‌اند و هنوز معلوم نیست دقیقاً کجای شرق ایران نسک‌های اوستا را بدین زبان نوشته‌اند. این زبان هم ریشه با سانسکریت و نزدیک به پارسی باستان می‌باشد. تنها اثر به جا مانده به آن اوستا است. از مطالعه اوستا بر می‌آید که دو لهجه مختلف از این زبان در اوستا به کار رفته است که یکی کهنه‌تر می‌نماید. نخستین لهجه گاهانی است که سروده‌های شخص زرتشت به آن است. به علاوه یسن‌های ۳۵ تا ۴۱ (هفت‌هات) و نیز چهار دعا از یسن ۲۷ به این لهجه است. لهجه پسین‌تر لهجه سایر قسمت‌های اوستای امروزی است. این الفبا برای هر یک از آواهای اوستایی یک نشانه دارد. نشانه‌های الفبایی جدا از هم نوشته می‌شده و در هر جای کلمه که می‌آمده یک صورت داشته است.

برخی از حروف خط اوستایی

ردیف	اوستایی	آوا	۹	د	ای	۱۸	ځ	د	۲۷	و	ک
۱	𐬀	آ	۱۰	پ	ای کشیده	۱۹	𐬁	ذ	۲۸	ڤ	گ
۲	𐬁	آ	۱۱	ل	پ	۲۰	𐬂	ر	۲۹	𐬃	ل
۳	𐬂	!	۱۲	و	پ	۲۱	𐬄	ز	۳۰	𐬅	م
۴	𐬃	! کشیده	۱۳	ص	ت	۲۲	𐬅	ژ	۳۱	𐬆	ن
۵	𐬄	أ	۱۴	ځ	ث	۲۳	𐬆	س	۳۲	𐬇	و
۶	𐬅	أ کشیده	۱۵	ج	چ	۲۴	𐬇	ش	۳۳	𐬈	و میان دو
۷	𐬆	او	۱۶	چ	چ	۲۵	𐬈	ف	۳۴	𐬉	ه
۸	𐬇	او کشیده	۱۷	ځ	خ	۲۶	𐬉	غ	۳۵	𐬊	ی

۸-۲- جدول ارکان مشهور

مشهورترین و پر استعمالترین ارکان عروض فارسی بیست رکن زیر است.

- : هجای بلند

U : هجای کوتاه

جدول ارکان مشهور

یک هجایی		دو هجایی		سه هجایی		چهار هجایی		پنج هجایی	
فَع	**	-	فَعْل**	U-	فَعْلُن	UU-	فاعلاتن	-U--	مُسْتَفْعَلَاتُنْ
			فَعْلُن	--	فاعِلن	-U-	*فاعلاتْ	-U-U	مُتَفَاعِلُنْ
					فَعولن	U--	فعلاتن	UU--	
					مفعولن	---	*فعلاتْ	UU-U	
					مفعول*	--U	مفاعیلن	U---	
							*مفاعیلْ	U--U	
							مفاعِلنْ	U-U-	
							مُسْتَفْعِلنْ	--U-	
							*مستفعلْ	--UU	
							مفتعلنْ	-UU-	

** فقط در پایان مصراع می آیند

* هیچگاه در آغاز مصراع قرار نمی گیرند

۸-۳- جدول اوزان پر کاربرد

بیش از ۳۰۰ وزن در شعر فارسی موجود است. خوش‌آهنگ‌ترین و رایج‌ترین اوزان، همان‌هاست که حافظ و مولوی در دیوان خود به کار برده‌اند [۱]. (جدول به ترتیب کاربرد در دیوان حافظ و مولانا مرتب شده)

جدول اوزان پر کاربرد

کد	نماد	وزن	نام
۱۰۰۱	UU--UU--UU--UU-	فعلاتن فعلاتن فعلاتن فعطن	رمل مثنی مخبون محذوف
۱۰۰۲	U-U-UU--U-U-UU-	مفاعیلن فعلاتن مفاعیلن فعطن	مجتث مثنی محذوف
۱۰۰۳	--U-U-UU--U-U-	مفعول فاعلاتن مفاعیلن فاعطن	مضارع مثنی اخرج مکفوف محذوف
۱۰۰۴	-U---U---U---U-	فاعلاتن فاعلاتن فاعلاتن فاعطن	رمل مثنی محذوف
۱۰۰۵	U---U---U---U--U	مفاعیلن مفاعیلن مفاعیلن مفاعیلن	هزج مثنی سالم
۱۰۰۶	U---U---U--	مفاعیلن مفاعیلن فعولن	هزج مسدس محذوف
۱۰۰۷	--U-U----U-U--	مفعول فاعلاتن مفعول فاعلاتن	مضارع مثنی اخرج
۱۰۰۸	--UU--UU--UU--	مفعول مفاعیلن مفاعیلن فعولن	هزج مثنی اخرج مکفوف محذوف
۱۰۰۹	UU--U-U-UU-	فعلاتن مفاعیلن فعطن	خفیف مخبون محذوف
۱۰۱۰	-U---U---U-	فاعلاتن فاعلاتن فاعطن	رمل مسدس محذوف
۱۰۱۱	-UU-U-U--UU-U-U-	مفتعلن مفاعیلن مفتعلن مفاعیلن	رجز مثنی مطوی مخبون
۱۰۱۲	--UU-----UU---	مفعول مفاعیلن مفعول مفاعیلن	هزج مثنی اخرج
۱۰۱۳	UU-U-U--UU-U-U--	فعلات فاعلاتن فعلات فاعلاتن	رمل مثنی مشکول
۱۰۱۴	U-U-UU--U-U-UU--	مفاعیلن فعلاتن مفاعیلن فعلاتن	مجتث مثنی مخبون
۱۰۱۵	--UU-U-U--	مفعول مفاعیلن فعولن	هزج مسدس اخرج مقبوض محذوف
۱۰۱۶	--U---U--	مستفعلاتن مستفعلاتن	مقارب مثنی اثلیم
۱۰۱۷	-UU--U-U-UU--	مفتعلن فاعلات مفتعلن فع	منسرح مثنی مطوی منحور
۱۰۱۸	--U---U---U---U-	مستفععلن مستفععلن مستفععلن مستفععلن	رجز مثنی سالم
۱۰۱۹	-UU--U--UU--U-	مفتعلن فاعطن مفتعلن فاعطن	منسرح مثنی مطوی مکشوف
۱۰۲۰	-U-U---U-U---	فاعلات مفعولن فاعلات مفعولن	مقتضب مثنی مطوی مقطوع
۱۰۲۱	-UU--UU--U-	مفتعلن مفتعلن فاعطن	سریع مطوی مکشوف
۱۰۲۲	U--U--U--U--	فعولن فعولن فعولن فعولن	مقارب مثنی سالم
۱۰۲۳	U--U--U--U-	فعولن فعولن فعولن فعل	مقارب مثنی محذوف
۱۰۲۴	-UU--UU--UU--UU-	مفتعلن مفتعلن مفتعلن مفتعلن	رجز مثنی مطوی
۱۰۲۵	UU--UU--UU--UU--	فعلاتن فعلاتن فعلاتن فعلاتن	رمل مثنی مخبون

کد	نماد	وزن	نام
۱۰۲۶	UU--UU--UU-	فعلاتن فعلاتن فعلن	رمل مسدس مخبون محذوف
۱۰۲۷	U--UU--UU--UU--	مفاعیل مفاعیل مفاعیل فعولن	هزج مثنی مکفوف محذوف
۱۰۲۸	UU--U-U-UU--U-U-	فعلاتن مفاعیلن فعلاتن مفاعیلن	خفیف مثنی مخبون
۱۰۲۹	--UU-U-U---	مفعول مفاعیلن مفاعیلن	هزج مسدس اخرب مقبوض
۱۰۳۰	-U---U---U---U--	فاعلاتن فاعلاتن فاعلاتن فاعلاتن	رمل مثنی سالم
۱۰۳۱	--U-UU---U-UU-	مستفعلن فعلن مستفعلن فعلن	بسیط مثنی مخبون

۸-۴- اوزان مثنوی‌های مشهور

مثنوی‌ها بخش مهمی از شعر فارسی را تشکیل می‌دهند، با این همه از تمام اوزان شعر فارسی، تنها از هفت وزن جهت سرودن مثنوی‌های مهم استفاده شده:

۱- فعولن فعولن فعولن فعل (از گروه کد ۱۰۲۲)

۱۰۲۲	U--U--U--U--	فعولن فعولن فعولن فعولن	مقارب مثنی سالم
------	--------------	-------------------------	-----------------

وزن حماسی و رزمی است. شاهنامه فردوسی در این وزن سروده شده است.
 چو فردا برآید بلند آفتاب من و گرز و میدان افراسیاب (فردوسی)
 این وزن گاه در پند و اندرز نیز به کار رفته است. مانند بوستان سعدی.

۲- مفاعیلن مفاعیلن فعولن (کد ۱۰۳۳)

۱۰۳۳	U---U---U--	مفاعیلن مفاعیلن فعولن
------	-------------	-----------------------

بیشتر مناسب داستان‌های عاشقانه است مانند ویس و رامین فخرالدین گرجانی و خسرو شیرین نظامی.
 شب افروزی چو مهتاب جوانی سیه چشمی چو آب زندگانی (نظامی)

۳- فاعلاتن فاعلاتن فاعلن (۱۰۱۰)

۱۰۱۰	-U---U---U-	فاعلاتن فاعلاتن فاعلن	رمل مسدس محذوف
------	-------------	-----------------------	----------------

مناسب مضامین پندآمیز و عارفانه است. منطق الطیر عطار و مثنوی مولوی در این وزن است.
 بشنو از نی چون حکایت می‌کند از جدایی‌ها حکایت می‌کند (مولوی)

۴- مفتعلن مفتعلن فاعلن (۱۰۲۱)

۱۰۲۱	-UU--UU--U-	مفتعلن مفتعلن فاعلن	سریع مطوی مکشوف
------	-------------	---------------------	-----------------

گرچه به قول نیما یوشیج، تند و رقص آور است و مناسبتی با معانی پند و حکمت ندارد، اما نظامی مخزن الاسرار را -که در معنی پند و حکمت است- در این وزن سروده.
ای همه هستی ز تو پیدا شده خاک ضعیف از تو توانا شده (نظامی)

۵- فعلاتن مفاعلن فعلن (۱۰۰۹)

۱۰۰۹	UU--U-U-UU-	فعلاتن مفاعلن فعلن	خفیف مخبون محذوف
------	-------------	--------------------	------------------

وزنی شاد و نسبتاً ضربی است و مناسب مضامین بزمی مانند هفت پیکر نظامی است. اما سنایی، حدیقه را که در بیان معارف و حقایق و حکمت است در این وزن سروده است.
هست بود همه درست به تو بازگشت همه به توسست به تو (نظامی)

۶- فعلاتن فعلاتن فعلن (۱۰۲۶)

۱۰۲۶	UU--UU--UU-	فعلاتن فعلاتن فعلن	رمل مسدس مخبون محذوف
------	-------------	--------------------	----------------------

جامی، سبجه الابرار خود را -که پند آمیز و عرفانی است- در این وزن سروده
شمع شو شمع که خود را سوزی تا بدان بزم کسان افروزی (جامی)

۷- مفعول و مفاعلن فعولن (۱۰۱۵)

۱۰۱۵	--UU-U-U--	مفعول مفاعلن فعولن	هزج مسدس اخرب مقبوض محذوف
------	------------	--------------------	---------------------------

وزنی ضربی و مناسب مضامین عاشقانه است. لیلی و مجنون نظامی و مجنون و لیلی جامی و هاتفی در این وزن سروده شده. اما خاقانی تحفه العراقین را که سفرنامه‌ای است در این وزن سروده است.
غافل منشین نه وقت بازی است وقت هنر است و سرفرازی است

جدول اوزان مثنوی‌های مشهور

نام مثنوی	شاعر	قرن	کد گروه وزن	فرع	وزن
مثنوی	مولوی	۷	۱۰۱۰	X	فاعلاتن فاعلاتن فاعلن
شاهنامه	فردوسی	۵	۱۰۲۲	√	فعولن فعولن فعولن فعول

نام مثنوی	شاعر	قرن	کد گروه وزن	فرع	وزن
بوستان	سعدی	۷	۱۰۲۲	√	فعولن فعولن فعولن فعول
مخزن الاسرار	نظامی	۶	۱۰۲۱	X	مفتعلن مفتعلن فاعلن
خسرو شیرین	نظامی	۶	۱۰۳۳	X	مفاعیلن مفاعیلن فعولن
هفت پیکر	نظامی	۶	۱۰۰۹	X	فعلاتن مفاعلن فعلن
لیلی و مجنون	نظامی	۶	۱۰۱۵	X	مفعول مفاعلن فعولن
منطق الطیر	عطار	۷	۱۰۱۰	X	فاعلاتن فاعلاتن فاعلن
حدیقه‌ی	سنایی	۶	۱۰۰۹	X	فعلاتن مفاعلن فعلن
سبحه الابرار	جامی	۹	۱۰۲۶	X	فعلاتن فعلاتن فعلن
ویس و رامین	فخرالدین گرجانی	۵	۱۰۳۳	X	مفاعیلن مفاعیلن فعولن
مجنون و لیلی	جامی	۹	۱۰۱۵	X	مفعول مفاعلن فعولن
تحفه العراقین	خاقانی	۶	۱۰۱۵	X	مفعول مفاعلن فعولن

۸-۵- نویسه‌های فارسی و همانندها

در این بخش اطلاعاتی از نویسه‌های فارسی، نویسه‌های عربی شبیه فارسی و نویسه‌های خاص گنجانده شده. این اطلاعات برای کار با یونی‌کد در متون فارسی بسیار مفید بوده و می‌تواند راهنمای خوبی برای حل کردن برخی مشکلات ناشی از استانداردهای مختلف باشد. این اطلاعات از سایت <http://www.nic.ir> برداشته شده است.

۸-۵-۱- صفحه کلید فارسی

صفحه کلید فارسی به طور پیش فرض در اکثر نسخه‌های جدید لینوکس و نیز در مکینتاش موجود است، ولی در ویندوز مایکروسافت نیاز به نصب دارد. راهنمای نصب صفحه کلید استاندارد فارسی در ویندوز ۲۰۰۰ و XP را می‌توانید در نشانی‌های زیر بیابید:

http://www.p30download.com/archives/utility/persian_tool/download_traylayout_12/

<http://www.dadkhahi.net/wiki/index.php/%D8%AF%D8%A7%D9%86%D8%B4%D9%86%D8%A7%D9%85%D9%87:%D9%81%D8%A7%D8%B1%D8%B3%DB%8C%E2%80%8C%D9%86%D9%88%DB%8C%D8%B3%DB%8C>

۸-۵-۲- نویسه‌های فارسی و یونی‌کد آنها

در جدول زیر تمامی حروفی که در فارسی امروز به کار برده می‌شود، به همراه یونی‌کد آنها آورده شده.

نام نویسه	شکل نویسه	کد یونی‌کد	مثال
حرف آی با کلاه	آ	U+0622	آب، مآخذ
حرف الف	ا	U+0627	امید، بام
حرف همزه تنها	ء	U+0621	شیء
حرف الف همزه	أ	U+0623	مبدأ، یأس
حرف ب	ب	U+0628	بال، تب، طبل، ناب
حرف پ	پ	U+067E	پلک، سپید، چپ، توپ
حرف ت	ت	U+062A	تور، چتر، دست، بیات
حرف ث	ث	U+062B	ثابت، مثل، ارث، بعث

نام نویسه	شکل نویسه	کد یونی کد	مثال
حرف جیم	ج	U+062C	جوی، پنجره، گنج، کاج
حرف چ	چ	U+0686	چرم، پنچر، گچ، پارچ
حرف ح	ح	U+062D	حمید، محله، صبح، مدح
حرف خ	خ	U+062E	خرم، سخن، ملخ، چرخ
حرف دال	د	U+062F	دام، مد
حرف ذال	ذ	U+0630	ذرت، جذر
حرف ر	ر	U+0631	راد، مرد
حرف ز	ز	U+0632	زنگ، بزم
حرف ژ	ژ	U+0698	ژاله، مژده
حرف سین	س	U+0633	سیب، بست، مس، درس
حرف شین	ش	U+0634	شب، کشت، آتش، موش
حرف صاد	ص	U+0635	صبر، عصر، شاخص، خاص
حرف ضاد	ض	U+0636	ضیا، مضر، قبض، عرض
حرف طا	ط	U+0637	طناب، مطب، ضبط، افراط
حرف ظا	ظ	U+0638	ظهر، نظم، حفظ، حفاظ
حرف عین	ع	U+0639	عماد، معلم، ربع، شجاع
حرف غین	غ	U+063A	غم، مغز، تیغ، باغ
حرف ف	ف	U+0641	فرهاد، نفر، زلف، برف
حرف قاف	ق	U+0642	قند، نقد، زنبق، برق
حرف کاف	ک	U+06A9	کفش، شکر، پلک، خاک
حرف گاف	گ	U+06AF	گرم، شگرد، زنگ، رگ
حرف لام	ل	U+0644	لاله، قلک، مبل، بال
حرف میم	م	U+0645	مرز، کمک، سم، نام
حرف نون	ن	U+0646	نرم، لنگر، آهن، نان
حرف واو	و	U+0648	ورق، لبو
حرف واو همزه	ؤ	U+0624	رؤسا، مؤمن
حرف ه	ه	U+0647	هلو، مهر، خانه، راه
حرف ت گرد	ة	U+0629	دایرةالمعارف، ثقةالاسلام

نام نویسه	شکل نویسه	کد یونی کد	مثال
حرف ی	ی	U+06CC	یاس، کیف، جارچی، چای
حرف ی همزه	ئ	U+0626	رئیس، مسئله، متألئ

۸-۵-۳- نویسه‌های عربی و همانند آن در فارسی

یکی از مشکلات رایج برنامه نویسی برای فارسی وجود حروف همانند فارسی در عربی و کد متفاوت آنهاست. به دلیل آنکه برخی و بلکه بسیاری از متون تایپ شده فارسی این موارد را رعایت نمی‌نمایند و مثلاً با ی عربی (ي) مطالب خود را نوشته‌اند، در پردازش متون فارسی حتماً باید به این نکته توجه کرد. بنابراین یا قبل از پردازش متون باید این حروف را تبدیل کرد و یا در هنگام پردازش هر دو حالت این حروف را در نظر گرفت. در زیر جدول دو حرف مشترک ولی با یونی کد متفاوت آورده شده. در این جدول اعداد عربی و انگلیسی و معادل آنها گنجانده نشده.

نویسه عربی	یونی کد نویسه	محل در صفحه کلید استاندارد فارسی	محل در صفحه کلید فارسی مایکروسافت	نویسه فارسی معادل	یونی کد نویسه فارسی	محل در صفحه کلید استاندارد فارسی	محل در صفحه کلید فارسی مایکروسافت
حرف کاف عربی (ك)	U+0643	Shift+ Z	ممکن نیست	حرف کاف فارسی	U+06A9	;	;
حرف ی عربی (ي)	U+064A	Shift+ D	Shift+ X	حرف ی فارسی	U+06CC	D	D

۸-۵-۴- فاصله مجازی

در فارسی بعضی حروف هیچگاه به حرف بعدی نمی‌چسبند مثلاً «الف» یا «د»، و بعضی دیگر بسته به نوع استفاده در کلمه گاهی می‌چسبند و گاهی نمی‌چسبند، مانند «ه» در «کلمه‌ای» و «کلمه‌ای» که دو لغا متمایزند. برای وارد کردن عبارت «کلمه‌ای» در رایانه باید از نویسه فاصله مجازی میان حروف «ه» و «الف» استفاده کرد. فاصله مجازی را می‌توان به صورت‌های زیر وارد کرد.

الف) در صفحه کلید استاندارد فارسی، با کلید تبدیل به همراه فاصله (Shift+Space) یا کلید تبدیل به همراه «ذ» (Shift+B). ترکیب Shift+Space ممکن است در بعضی صفحه کلیدهای استاندارد کار نکند.

ب) در صفحه کلید فارسی میکروسافت، با پایین نگه داشتن همزمان کلیدهای مهار و تبدیل و فشردن کلید ۲ (Ctrl+Shift+2)

ج) با هر صفحه کلیدی در ویندوز میکروسافت، پایین نگه داشتن کلید دگرساز و فشردن ۰۱۵۷ (Alt+0157). این روش در همه نسخه‌های ویندوز قابل استفاده است.

مثال	کد نویسه	نام نویسه
خانه‌ها	U+200C	فاصله مجازی (ZERO-WIDTH NON-JOINER)
ه ق	U+200D	اتصال مجازی (ZERO-WIDTH JOINER)

۸-۶- توضیحات مربوط به Data Set تولید شده

DSVN-v1.00

(Data Set Versa None-token)

DS: Data Set

V: Verse

N: None Token

v1.00: Version 1.00

حاوی ابیات بدون حرکت‌گذاری می‌باشد، این فایل توسط خبره تولید شده. هر بیت در یک خط قرار دارد، به عبارت بهتر، جداکننده ابیات از هم NewLine می‌باشد. هر مصراع از مصراع دوم توسط * جدا شده است.

DSCV-v1.00

(Data Set Constant Vowel string)

DS: Data Set

CV: CV String

v1.00: Version 1.00

حاوی رشته CV تولید شده توسط برنامه می‌باشد. مانند قبل هر بیت از بعدی با NewLine و هر مصراع از مصراع دوم با * جدا شده است. (C نشانه صامت و V نشانه مصوت می‌باشد)

DSUS-v1.00

(Data Set Udash String)

DS: Data Set

US: U dash String

v1.00: Version 1.00

حاوی رشته UDash تولید شده توسط برنامه می‌باشد. مانند قبل هر بیت از بعدی با NewLine و هر مصراع از مصراع دوم با * جدا شده است. U نشانه هجای کوتاه و Dash(-) نشانه هجای بلند می‌باشد)

DSA-v1.00

(Data Set Answer)

DS: Data Set

A: Answer

v1.00: Version 1.00

حاوی پاسخ (وزن صحیح) ابیات نسخه یکسان فایل DSVN به صورت کد گذاری می‌باشد. این فایل توسط خبره تولید می‌شود. کدهای این فایل را در جدول «اوزان پرکاربرد» در پیوست سوم می‌توانید ببینید. جداسازی ابیات به وسیله NewLine می‌باشد.

نسخه فایل‌ها

نسخه فایل به صورت vx.xx نشان داده می‌شود.

- v نشانه نسخه (Version) است.
- عدد قبل از ممیز نشانه‌ی نسخه ابیات می‌باشد، بدان معنی که فقط زمانی تغییر می‌کند که تعداد و یا ترتیب ابیات در فایل DSVN عوض شود.
- عدد بعد از ممیز نشانه‌ی نسخه برنامه است. بدان معنی که ممکن است فایل DSVN تغییری نکنند، ولی برنامه تغییر داشته و در نتیجه فایل‌های تولید شده (DSCV, DSUS) تغییر کنند. به عنوان مثال، اگر شما دونسخه v1.21 و v1.10 را داشته باشید، فقط ممکن است فایل‌های DSCV, DSUS تغییر کرده باشند. ولی دو فایل DSVN و DSA در هر دو نسخه یکی هستند. ولی در دونسخه v2.01 و v1.01 ابیات تغییر کرده و در نتیجه همه فایل‌ها دچار تغییر شده‌اند.

فصل نهم

واژه نامه

۹- واژه نامه

به دلیل آنکه این پژوهش، یک پژوهش میان رشته‌ای می‌باشد، و طبیعتاً اکثر دانشجویان کامپیوتر اطلاع چندانی از اصطلاحات ادبیات و عروض ندارند، برآن شدیم بعضی اصطلاحات به کار برده شده در این پژوهش را به صورت واژه‌نامه در پایان بیاوریم، تا برای این دسته از دانشجویان مشکلی پیش نیاید. بخاطر آنکه استفاده از این واژه نامه آسان‌تر شود، در توضیحات هر کلمه، زیر واژه‌گانی که خود در واژه‌نامه باشند، خط کشیده شده.

آهنگ: ر.ک وزن

اجزا: (= اصول = افاعیل) از ترکیب سبب و وتد و فاصله، حاصل می‌شود. در عروض جدید به آن ارکان گفته می‌شود.

ارکان: همانطور که در صرف عربی، قالب‌هایی برای کلمات اختراع کرده‌اند و مثلاً می‌گویند شاعر بر وزن فاعل، در عروض نیز برای (یا مجموعه چند هجای کوتاه و بلند) قالب‌هایی وضع کرده‌اند و مثلاً می‌گویند «می‌آیم» (---) بر وزن مفعولن. بنای این قالب بر «ف»، «ع»، و «ل» است. البته می‌توان به جای آن هر معادل دیگری به کار برد. مانند «ت» و «تن» که در این صورت «مفاعیلن» می‌شود: «ت تن تن تن».

البته ارکان در عروض قدیم معنای دیگری می‌دهد: علمای عروض قدیم به جای صامت و مصوت از ساکن و متحرک و به جای هجا از رکن که مجموع یک ساکن و یک یا چند متحرک بود استفاده می‌کردند، آنان به سه رکن اصلی قائل بودند: سبب، وتد و فاصله که هر کدام دو نوع بودند. از ترکیب این ارکان، اجزا یا اصول یا افاعیل حاصل می‌شود. در عروض جدید به جای اجزا، همان ارکان گفته می‌شود.

اصوات شعر: ر.ک هجا

اصول: ر.ک اجزا

افاعیل: ر.ک اجزا

املای عروضی: (= آوا نویسی) نوشتن کلمات به خطی است که بتوان با آن مقطع اصوات یا هجا را به آسانی تشخیص داد. نگارش دقیق اصوات اعم از صامت و مصوت است. برای این کار می‌توان، با اندک تغییر و اضافاتی هم از خط فارسی استفاده کرد و هم از خط لاتین. (توجه: در این پژوهش از هر دو خط استفاده شده)

بحر: قدما به اوزان مستقل معتقد نبودند و اوزان شبیه به هم را از یک وزن اصلی متفرع (فرع بر وزن اصلی) می‌دانستند. مجموعه این چند وزن شبیه به هم را که اختلاف چندانی با هم نداشتند، بحر می‌گفتند.

بحر سالم: (= وزن سالم) اوزانی که دارای ارکان (با تعریف قدما) سالم هستند، بحر سالم گویند. مانند فاعلاتن چهار بار = بحر رمل مثنی سالم.

بحر طویل: اسم معروف‌ترین وزن در شعر عربی است که مبنی بر تکرار فعولن و مفاعیلن است:
اذا یئس الانسان طال لسانه کسنور مغلوب یصول علی الکلب (گلستان)
اما در ادبیات فارسی اسم نوعی شعر غیر رسمی است که در آن تعداد ارکان حد معینی ندارد و لذا شعر یک نفس خوانده می‌شود:

«رنگ رخسار و دو گوش و خط و خد و قد و عارض و خال و لب ای سرو پری روی سمن‌بر، شفق و کوکب و شام و سحر و طوبی و گلزار بهشت است و هلال و طرف چشمه کوثر» (عصمت بخارایی)
مداحی معروف نزار القطری که در محرم سال پیش در صدا و سیمای جمهوری اسلامی ایران پخش شد، در این بحر سروده شده بود:
«منم شهپر جبریل، منم آیه تقدیر منم نخله خوبان، منم زاده زهرا منم عرش و منم فرش منم کرسی و لوح و قلم و باعث ایجاد دو عالم»
در شعر نو اگر قواعد پایان‌بندی رعایت نشود به بحر طویل شبیه می‌شود.

بخش: ر.ک هجا

بیت مزاحف: بیتی است که در ارکان وزن آن زحافی رخ داده باشد، یعنی ارکان آن مزاحف باشد. مانند فاعلاتن فاعلاتن فاعلن = بحر رمل مسدس محذوف.

تقطیع: در لغت به معنای قطعه قطعه کردن است و در اصطلاح عروض به معنای قطعه قطعه کردن شعر به هجای کوتاه و هجای بلند است.

حرف: واژه از واحد کوچکتری به نام حرف تشکیل شده. منظور از حرف آن چیزی است که شنیده می‌شود نه آنکه نوشته می‌شود. بنابراین «خواهر» از پنج حرف تشکیل شده (خ ا ه - ر). برخی حروف در خط دارای چند نشانه هستند. مانند «ت» و «ط» که هر دو یک حرف هستند. البته بعضی علامات خط نشانه‌ی چند حرف هستند مانند «و» در «سود»، «تو» و «وجد» که به ترتیب نشانه یک مصوت بلند (VV)، یک مصوت کوتاه و یک صامت (VC)، یک صامت (C)، می‌باشد.

رجز: وزن مبتنی بر «مستفعلن» را گویند.

رجز مخبون: وزن مبتنی بر «مفاعeln» را گویند. به جای آن هزج مخبون نیز به کار می‌برند.

رجز مطوی: وزن مبتنی بر «مفتعلن» را گویند.

رکن: ر.ک ارکان

رشته صامت - مصوت: (= رشته CV) رشته‌ای متشکل از صامت (C) و مصوت (V). به دست آوردن این رشته - که از روی صورت ملفوظ شعر به دست می‌آید - یکی از مراحل رسیدن به وزن شعر است.

رشته CV: ر.ک رشته صامت - مصوت

رشته Udash: رشته‌ای متشکل از «U» و «-» که در آن «U» نشانه هجای کوتاه و «-» نشانه هجای بلند است. این رشته از روی رشته صامت-مصوت به دست می‌آید. به این صورت که «CV» تبدیل به «U» شده و «CVV» یا «CVC» تبدیل به «-» می‌شوند.

رکن بندی: بعد از تقطیع اشعار به هجای کوتاه و بلند (رشته Udash) با توجه به جدول ارکان، به جای هر چند هجا، یک رکن قرار می‌دهیم

رمل: وزن مبتنی بر «فاعلاتن» را رمل گویند.

رمل مخبون: وزن مبتنی بر «فعلاتن» را گویند.

رمل مشکول: وزن حاصل از تناوب «فعلات فاعلاتن» را گویند.

زحاف: اگر در رکن (با تعریف قدما) تغییری حاصل شود آن را مزاحف یا غیر سالم گویند. مانند «فعلاتن» یا «فاعلات» که بر اثر تغییری (به اعتقاد قدما) از «فاعلاتن» منشعب شده‌اند بنابراین از زحافات «فاعلاتن» هستند.

زحافات: ر.ک زحاف

ساکن: در قدیم به جای صامت ساکن می‌آوردند و علامت آن «ا» بوده است. مصوت بلند نیز ساکن محسوب می‌شد.

سالم: اگر رکن (با تعریف قدما) شکل اصلی خود را حفظ کرده و تغییری در آن راه نیافته باشد، آن را سالم گویند. مانند فاعلاتن.

سبب: به معنای طناب است و وجه تسمیه آنست که همچنان که طناب خیمه را گاهی تمام بکشند و گاه کوتاه‌تر نمایند، این رکن (در تعریف قدما) در بعضی اجزا گاه تمام بیاورند و گاهی کوتاه کنند. بر دو گونه است: سبب خفیف و سبب ثقیل.

سبب خفیف: از یک حرف متحرک (ه) و یک حرف ساکن (ا) تشکیل شده. مانند «من» (-).

سبب ثقیل: دو حرف متحرک متوالی را گویند. (ه ه) مانند «همه» (UU)

سیلاب: ر.ک هجا

شعر: سخنی خیال انگیز که موزون و قافیه‌دار است. خواجه نصیرالدین طوسی وزن را به دلیل خیال انگیز بودن از فصول ذاتی شعر دانسته. شعر نو نیز موزون است و تنها تفاوت آشکار آن با شعر سنتی، کوتاهی و بلندی مصراع‌ها می‌باشد.

صامت: (= همخوان = Const) زبان فارسی دارای ۲۳ صامت است. برخی صامت‌های زبان فارسی در خط دارای چند علامت هستند. مانند: «ز = ظ = ذ = ض».

ضرب: آخرین رکن مصراع دوم.

عروض: یکی از اقسام علوم ادبی است که موضوع آن بحث در وزن شعر است و در مورد چگونگی ایجاد وزن، انواع وزن، صحت و سقم آن و برخی از شگردهایی که مخصوص کلام منظوم است بحث می‌کند. امروزه عروض جزو یکی از نظام‌های زبان شناسی محسوب می‌شود و در مطالعات مربوط به فنولوژی (واج شناسی) و سبک شناسی کاربرد وسیع یافته است. در معنای لغوی عروض گفته‌اند: کلمه‌ای است در معنی اسم مفعول: یعنی شعر را بر آن عرضه می‌کنند، تا موزون و ناموزون آن از هم باز شناخته شود.

آخرین رکن مصراع اول را نیز عروض گویند.

عروض جدید: (= عروض علمی) روش به دست آوردن وزن شعر در قدیم، غالباً سماعی بوده و از سبب و وتد و فاصله برای تقطیع استفاده می نمودند که روشی مشکل و نادقیق بوده ولی در ۵۰ سال اخیر، با الگوبری از عروض لاتین و یونان باستان روش به دست آوردن وزن تغییر کرد. از متولیان عروض جدید و یا عروض علمی در زبان فارسی می توان به مسعود فرزند، دکتر پرویز ناتل خانلری و ابوالحسن نجفی اشاره کرد.

عروض علمی: (= عروض جدید) ر.ک عروض جدید.

عروض قدیم: ر.ک عروض جدید.

فاصله: جدایی میان دو دامنه خیمه را گویند. این رکن بر دو قسم است: فاصله صغری و فاصله کبری.

فاصله صغری: از سه حرف متحرک (ه ه ه) و یک ساکن (ا) تشکیل شده. مانند «شنوم» (UU-)

فاصله کبری: از چهار حرف متحرک (ه ه ه ه) و یک ساکن (ا) تشکیل شده. مانند «ببرمش» (UUU-)

متحرک: در قدیم به جای مصوت متحرک به کار می بردند و نشانه‌ی آن نیز «ه» بوده است.

متفق الارکان: اگر رکنی در یک مصراع چندبار (۲، ۳، ۴) تکرار شود، وزن را متفق الارکان گویند.

متقارب: وزن مبتنی بر فعولن را گویند.

متحد الارکان: اگر ارکان مصراع یک در میان تکرار شوند، وزن را متحد الارکان گویند.

مثنی: اگر تعداد ارکان شعر ۸ باشد، وزن را مثنی گویند.

مجثت مخبون: وزن حاصل از تناوب ارکان «مفاعیلن فعلاتن» را گویند.

مربع: اگر تعداد ارکان شعر ۴ باشد، وزن را مربع گویند.

مزاحف: ر.ک زحاف

مسدس: اگر تعداد ارکان شعر ۶ باشد، وزن را مسدس گویند.

مصوت: (= واکه = Vowel) زبان فارسی دارای سه مصوت کوتاه (- -) و سه مصوت بلند (ا و ی) است.

مصوت بلند: زبان فارسی دارای سه مصوت بلند (ا و ی) است. مصوت بلند در این پژوهش به وسیله «۷۷» نشان داده می‌شود.

مصوت کوتاه: زبان فارسی دارای سه مصوت کوتاه (اَ - اِ - اُ) است. مصوت کوتاه در این پژوهش به وسیله «۷» نشان داده می‌شود.

مضارع: وزن مبتنی بر ترکیب «مفاعیلن و فاعلاتن» را بحر مضارع گویند.

مضارع اُخرب: وزن مبتنی بر تکرار ارکان «مستفعِلن فعولن» را گویند.

واکه: ر.ک مصوت

وتد: به معنای میخ است و بدین جهت این رکن را وتد نامند که میخ هر کجا رود ثابت و استوار است. این رکن نیز در اغلب اجزای عروضی ثابت است و تغییراتی که در سبب‌ها وارد می‌گردد بدان راه نمی‌یابد. این رکن بر دو نوع است: وتد مقرون و وتد مفروق

وتد مفروق: متحرک + ساکن + متحرک (ه ا ه). مانند «نامه» (U-)

وتد مقرون: از دو حرف متحرک (ه ه) و یک ساکن (ا) تشکیل شده. مانند «چمن» (-U)

وزن: (= آهنگ) آهنگی که از ترکیب و توالی منظم و مخصوصی از هجاها به وجود می‌آید. این نظم و تناسب اصوات به انحنای گوناگون نزد ملل مختلف مبین نوعی آهنگ و موسیقی است. واحد وزن (یا حداقل وزن) در زبان فارسی، رکن است.

وزن تقطیعی: وزنی که بدون اعمال اختیارات شاعری به دست آمده است.

وزن سالم: ر.ک بحر سالم

هجا: (= سیلاب = بخش = اصوات شعر = Syllable) هجای فارسی مجموعه‌ای از یک مصوت و یک یا دو یا سه صامت است. بنابراین تعداد هجاهای هر کلمه مساوی تعداد مصوت‌های آن است.

هجای کوتاه: از صامت + مصوت کوتاه، تشکیل شده است و در این پژوهش با علامت «U» نشان داده می‌شود.

هجای بلند: از صامت + مصوت کوتاه + صامت و یا صامت + مصوت بلند تشکیل شده است و در این پژوهش با علامت «-» نشان داده می‌شود.

هجای کشیده: دارای چهار یا پنج حرف است و با علامت « U- » نشان داده می‌شود و در واقع ترکیب یک هجای کوتاه و یک هجای بلند است، بنابراین در این پژوهش از این هجا استفاده‌ای نشده و به جای آن از هجای کوتاه و بلند استفاده می‌شود.

هزج: به اوزان مبتنی بر «مفاعیلن» هزج گویند.

هزج اخرب: وزن حاصل از تکرار «مفعول مفاعیلن» یا «مستفعل مفعولن» را گویند

هزج مقبوض: ر.ک رجز مخبون

همخوان: ر.ک صامت

منابع

۱۰- منابع

- [۱] شمیسا، س، آشنایی با عروض و قافیه، ویراست چهارم، نشر میترا، ۱۳۸۳
- [۲] عظیمی، م، مقایسه موسیقی شعر حافظ، خواجه و سلمان با استفاده از کامپیوتر، دومین کارگاه پژوهشی زبان فارسی و رایانه، ص ۲۲۶، ششم و هفتم تیر ۱۳۸۵ ج
- [۳] وحیدیان کامیار، ت، فنون و صنایع ادبی (عروض)، نشر ایران، ۱۳۶۵
- [۴] ماهیار، ع، عروض فارسی شیوه‌ای نو برای آموزش عروض و قافیه، نشر قطره، چاپ پنجم ۱۳۷۴
- [۵] فرزاد، م، مجموعه شعر فارسی، مجله خرد و کوشش، شیراز ص ۶۵۰، بهمن ۱۳۴۹
- [۶] خطیب رهبر، خ، دیوان غزلیات خواجه حافظ شیرازی، صفی علی شاه، چاپ ششم، ۱۳۶۹
- [۷] شفیعی کدکنی، م، گزیده غزلیات شمس (مولانا جلال الدین) چاپ سپهر، چاپ پنجم ۱۳۶۳
- [۸] تجلیل، ج، عروض، نشر همراه، پاییز ۱۳۶۸
- [۹] یوسفان، ا، یک سیستم بازیابی اطلاعات متنی برای زبان فارسی بر پایه نمایه گذاری معانی پنهان، پایان‌نامه کارشناسی ارشد، دانشگاه شیراز ایران، دانشکده کامپیوتر، شهریور ۱۳۸۲
- [۱۰] رضی، ه، دستور زبان اوستایی، سازمان انتشارات فروهر، ۱۳۶۸
- [۱۱] نگاری، ل، زبان فارسی و نابسامانی های تلفظی آن، آذرماه ۱۳۸۴
- [12] Pablo Gervas. Wasp: Evaluation of different strategies for the automatic generation of Spanish verse. In Proceedings of the AISB-00 Symposium on Creative& Cultural Aspects of AI, pages 93-100, 2000
- [13] Pablo Gervas. An expert system for the composition of formal Spanish poetry. Journal of Knowledge-Based Systems, 14(3-4):181-188, 2001.

[14] H. M. Manurung, G. Ritchie, and H. Thompson. A flexible integrated architecture for generating poetic texts. Informatics Research Report EDI- INF-RR-0016. University of Edinburgh, 2000.

[15] H. M. Manurung, G. Ritchie, and H. Thompson. Towards a computational model of poetry generation. In Proc. of the AISB-00 Symposium on Creative and Cultural Aspects of AI. 2001. CAEPIA 2001

[16] Elwell, S. The Persian Metters. Cambridge, 1976

[17] Iloff E., and Lehnert W. *Information extraction as a basis for high-precision text classification*. ACM Transactions on Information Systems, No.3, Vol. 12, pp. 296-333, 1994.

[18] Ville V, Proceedings of the Twelfth ESSLLI Student Session, Dublin, Ireland, 2007

[19] Knuth, D. Sorting And Searching, The Art Of Computer Programming, vol.3 Addison-Wesley, pp.391-392, 1973

[20] <http://www.nic.ir>

[21] <http://www.phpclasses.com>