مقایسهٔ ساخت آوایی گونه رسمی و محاوره‌ای زبان فارسی

وحید مواجی[۱]
محرم اسلامی[۲]

چکیده

گونهٔ رسمی و گونهٔ محاوره‌ای زبان‌ها غالباً تفاوت‌هایی با هم دارند و این تفاوت‌ها در همهٔ سطح‌های زبانی دیده می‌شود. میزان تفاوت بین گونهٔ رسمی و گونهٔ محاوره‌ای، که گاهی از آنها با عنوان تفاوت گفتار و نوشتار یاد می‌شود، از زبانی به زبان دیگر متفاوت است. زبان فارسی از جملهٔ زبان‌هایی است که در آن تفاوت گونهٔ رسمی و گونهٔ محاوره‌ای بسیار زیاد است. در این تحقیق تفاوت‌های آوایی یا به عبارتی فرایندهای آوایی‌ای بررسی می‌شود که در زبان فارسی در تبدیل گونهٔ رسمی به گونهٔ محاوره‌ای رخ می‌دهد. پیکره پژوهش حاضر دادگان گفتاری «فارسدات تلفنی» زبان فارسی(بی‌جن‌خان و همکاران، ۲۰۰۳) است که در آن گفتار پیوسته[۳] در دو سطح واجی[۴][۵] و آوایی[۶] در قالب دو زنجیرهٔ مستقل برچسب خورده است. هم‌گذاری این دو رشته از داده‌ها روشن می‌سازد که در مقایسهٔ این دو گونهٔ زبانی کدام فرایندهای آوایی در تبدیل زنجیرهٔ واجی به زنجیرهٔ آوایی دخیل‌اند. در انطباق دو رشتهٔ واجی و آوایی از الگوریتم لونشتاین[۷] استفاده می‌شود که مناسب و رایج در انطباق تقریبی رشته‌های متفاوت جهت یافتن فاصلهٔ بین آنها است. در نتیجه تفاوت دو رشتهٔ واجی و آوایی به صورت آماری به دست می‌آید. از نتایج این پژوهش می‌توان به لحاظ نظری در توصیف‌های زبان‌شناختی درباره نظام آوایی زبان فارسی، تهیهٔ منابع محاوره‌ای زبان فارسی و آموزش زبان فارسی به خصوص به غیرفارسی‌زبانان سود جست. از سوی دیگر در فن‌آوری‌های گفتار مانند بازشناسی و بازسازی گفتار، استخراج اطلاعات از متن‌های محاوره‌ای، تبدیل متن به زنجیرهٔ واجی گونهٔ محاوره‌ای زبان فارسی و امکان تبدیل آن به گونهٔ رسمی می‌توان از نتایج این تحقیق استفاده کرد.

---

[۱]-کارشناسی ارشد زبان‌شناسی رایانشی، دانشگاه صنعتی شریف mavaji@alum.sharif.edu
[۲] - دانشیار، دانشگاه زنجان meslami@znu.ac.ir

[۳]- continuous :گفتار پیوسته در علم گفتار به گفتار ضبط‌شده اطلاق می‌شود که به دو روش خوانده‌شده از روی متن و یا مستقیماً به‌دست می‌آید.

[۴]– در طراحی فارسدات گفتار پیوسته در دو سطح مجزای واجی و آوایی برچسب خورده است.

[5]- phonemic
[6]- phonetic
[7]- Levenshtein

واژه‌های کلیدی: ساخت آوایی، گونه رسمی، گونه محاوره‌ای، الگوریتم لونشتاین، فارسدات تلفنی فارسی

۱- مقدمه

گونه‌های رسمی و محاوره‌ای زبان فارسی در تمام سطح‌های زبان تفاوت‌های زیادی با هم دارند که این امر در مقایسه با برخی زبان‌ها بسیار چشمگیر است. در همهٔ زبان‌ها بین گونه‌های رسمی و محاوره‌ای تفاوت‌هایی وجود دارد و در زبان‌هایی که مسبوق به داشتن نظام نوشتاری دیرینه‌اند، این تفاوت بیشتر به چشم می‌خورد. بنابراین، زبان‌هایی که زبان علم و ادب باشند، که طبیعتاً خط نیز دارند، بین گونه‌های رسمی و محاوره‌ای آنها فاصله بیشتر خواهد بود. از دلایل این فاصله می‌توان به محافظه‌کارانه عمل کردن خط در انعکاس تغییرات زبان اشاره کرد که نهادهایی مانند فرهنگستان زبان، عامدانه در قالب برنامه‌ریزی زبانی متولی انجام تغییرات در خط می‌شوند. دلیل دیگر این است که در محیط‌های علمی افراد تحصیل‌کرده کمتر حاضر می‌شوند خلاف سنت نوشتاری زبان بنویسند. این موضوع باعث می‌شود به مرور زمان، گونه‌های رسمی و محاوره‌ای مثلاً از حیث نظام آوایی فاصله بگیرند، وضعیتی که در زبان فارسی شاهد آن هستیم.

در این پژوهش تفاوت‌های آوایی گونه‌های رسمی و محاوره‌ای زبان فارسی به صورت آماری با استفاده از پیکرهٔ زبانی «فارسدات تلفنی»استخراج و دسته‌بندی شد. در فارسدات تلفنی گفتار پیوسته در دو سطح واجی و آوایی برچسب خورد. مقایسهٔ زنجیره نشان می‌دهد گفتار رسمی و محاوره‌ای زبان چه تفاوت‌های آوایی با هم دارند. هم‌گذاری خودکار زنجیره‌های واجی و آوایی امر دشواری است، چرا که تحولات آوایی در همه جایگاه‌های صدایی کلمه رخ می‌دهد. اینکه کدام واحد صدایی به کدام واحد صدایی تبدیل شده و کدام واحد صدایی حذف یا درج شده در بررسی خودکار دشواری‌هایی را ایجاد می‌کند. تغییرات آوایی را می‌توان به چند دسته یا عمل تقسیم کرد:

الف. عمل درج: مثلاً /mehrban/ با درج /a/ به /mehraban/ تبدیل می‌شود.

ب. عمل حذف: مثلاً /dast/ با حذف همخوان پایانی به /das/ تبدیل می‌شود.

پ. عمل جانشینی: مثلاً /?ejtemâ?/ با جانشینی /š/ به جای /j/ و حذف همزهٔ پایانی به /?eštemâ/ تبدیل می‌شود.

چنانچه در «پ» می‌بینیم چندین عمل می‌توانند هم‌زمان روی یک رشته، اِعمال شوند. مثلاً در مثال «پ» عمل حذف، همزهٔ پایانی کلمه را نیز حذف کرده است. عمل‌های ذکرشده منحصر به یک آوای واحد نیستند و می‌توانند روی گروهی از آواها نیز اِعمال شوند، مثلاً «می‌گویند» تبدیل می‌شود به «می‌گن» که در آن گروه رشتهٔ آوایی /-uyand/ کلاً با گروه /-an/ جانشین شده است.

برای این که فرایندهای آوایی عمل کنند، باید بافتِ آوایی نیز در نظر گرفته شود. در این مقاله، از واج قبلی و واج بعدی به عنوان بافت آوایی استفاده شد. پیکرهٔ فارسدات تلفنی حاوی صورت واجی و آوایی کلماتی است که از ضبط مکالمات تلفنی افراد به دست آمده است. اگر صورت‌های واجی و آوایی را به صورت مجموعه یا رشته‌هایی از نویسه‌ها در نظر بگیریم، مسئله به تعیین میزان فاصلهٔ دو رشته کاهش می‌یابد. یعنی هر چه فاصلهٔ دو رشته کمتر باشد مشابهت بیشتری به هم دارند. بنابراین تغییر واجی از روی فاصلهٔ کمینهٔ بین دو رشته به دست خواهد آمد.

الگوریتم‌های مختلفی برای سنجش فاصلهٔ دو رشته وجود دارد. در این پژوهش از الگوریتم لونشتاین(۱۹۶۵) استفاده شد. البته این الگوریتم فاصلهٔ کمینهٔ دو رشته را محاسبه می‌کند، ولی مراحل اِعمال فرایندهای مختلف (درج، حذف، جانشینی) را مشخص نمی‌کند. لذا با تغییری که در این الگوریتم داده شد، مراحل اِعمال فرایندهای واجی هم به دست آمد. در این مقاله گونهٔ گفتاری معیار[۱] زبان فارسی از فارسدات تلفنی مدنظر است.

۲- معیارهای مشابهت

برای ارزیابی میزان شباهت یا عدم شباهت (فاصله) دو رشته نسبت به یکدیگر از معیارهای مشابهت رشته‌ای استفاده می‌شود. معیار مشابهت، عدد اعشاری است که درجه مشابهت دو رشته از نویسه‌ها را مشخص می‌سازد. مثلاً فاصلهٔ بین دو رشتهٔ «رایانه» و «پایانه» کمتر از فاصلهٔ بین دو رشتهٔ «رایانه» و «یارانه» است.

از معیارهای مشابهت در زمینه‌های مختلف مانند تطابق تقریبی رشته‌ها(ناوارو[۲]،۲۰۰۱)، مقایسه(استویلو[۳]، ۲۰۰۵) و جستجوی فازی رشته‌ها(جین[۴]، ۲۰۰۵) استفاده می‌شود. یکی از پرکاربردترین زمینه‌های آن، خطایاب‌های املایی(لی[۵]، ۲۰۰۸) است، بدین صورت که متنی به عنوان عنوان ورودی به خطایاب داده می‌شود و خطایاب باید کلمات اشتباه را با نزدیک‌ترین کلمات (کلماتی با بیشترین میزان مشابهت و کمترین فاصله) جایگزین سازد. در ادامه پرکاربردترین معیارهای فاصلهٔ دو رشته معرفی می‌شوند.

---

[۱]- در طراحی فارسدات تلفنی پیش‌بینی شده بود که بخش برگی از داده‌ها از گفتار گویشورانی اخذ شود که به گونهٔ رسمی زبان فارسی (معیار) سخن می‌گویند. در این پژوهش فقط به بررسی فرایندهای آوایی در این گونهٔ زبانی پرداخته‌ایم.

[2]- Navarro,G.
[3]- Stoilos, G.
[4]- Jin, L.
[5]- Li, C.

۲-۱- فاصلهٔ لونشتاین

فاصلهٔ لونشتاین[1] به صورت حداقل میزان تغییرات مورد نیاز برای رشتهٔ S به رشتهٔ T تعریف می‌شود که فرایند مجاز در آن عبارتند از درج، حذف یا جانشینی یک نویسهٔ واحد. برای مثال فاصلهٔ لونشتاین دو رشتهٔ« apple» و «orange » با یک ماتریس ساده در شکل ۱ محاسبه شده است. فاصلهٔ این دو رشته برابر ۵ است که در قسمت پایین و سمت راست ماتریس نشان داده شده است. این فاصله بدین معناست که رشتهٔ « apple» را با درج o، جانشینی r به جای a، a به جای p، n به جای p و g به جای l به رشتهٔ «orange» تبدیل می‌شود. یک فرایند درج و چهار فرایند جانشینی که مجموعاً برابر ۵ فرایند می‌شود.

		A	P	P	L	E
	0	1	2	3	4	5
O	1	1	2	3	4	5
R	2	2	2	3	4	5
A	3	2	3	3	4	5
N	4	3	3	4	4	5
G	5	4	4	4	5	5
E	6	5	5	5	5	5

شکل ۱: مثالی از فاصلهٔ لونشتاین

مراحل ساخته شدن ماتریس شکل ۱ به شرح ذیل است:

۱. ماتریس d با M سطر و N ستون ساخته می‌شود.
۲. به سطر اول مقادیر از 0 تا M و به ستون اول مقادیر از 0 تا N اختصاص داده می‌شود.
۳. هر نویسه‌ای از S (i از 1 تا M) و هر نویسه‌ای از T (j از 1 تا N) را بررسی می‌کنیم.
۴. اگر S[i] با T[j] برابر بود، هزینه برابر 0 و در غیر این صورت برابر 1 است.
۵. مقدار عنصر d[i, j] ماتریس برابر کمینه مقادیر زیر است:

الف. عنصر بالایی به علاوهٔ یک: d[i-1, j]+1

ب. عنصر سمت چپ به علاوهٔ یک: d[i, j-1]+1

ج. عنصر بالا و سمت چپ به علاوهٔ هزینه:d[i-1, j-1]+cost

[1]-Levenshtein distance

۶. بعد از این که تکرار مراحل ۳، ۴ و ۵ به پایان رسید، مقدار فاصله در عنصر d[N,M] قرار دارد.

شبه‌کد فاصلهٔ لونشتاین در شکل ۲ آمده است.

```
intLevenshteinDistance(char S[1..M], char T[1..N]){
        declare intd[0..M, 0..N]
        for i from 0 to M
                d[i,0] := i //the distance of any first string to an empty
second string
        for j from 0 to N
                d[0, j] := j //the distance of any second string to an empty
first string
        for j from 1 to N
        {
                for i from 1 to M
                {
                        if S[i] = T[j] then
                                d[i, j] := d[i-1, j-1] //no operation
required
                        else d[i, j] := minimum(d[i-1, j] + 1, //a deletion
                                                d[i, j-1] + 1, //an
insertion
                                                d[i-1, j-1] + 1) //a
substitution
                }
        }
        return d[M, N]
}
```

شکل ۲: شبه‌کد فاصلهٔ لونشتاین

۲-۲- فاصلهٔ همینگ

فاصلهٔ همینگ[1](۱۹۵۰) فقط فرایند جانشینی را برای تبدیل S به T مجاز می‌شمارد. بنابراین طول S و T باید برابر باشد. از این الگوریتم بیشتر برای تشخیص و تصحیح خطا برای دو رشته با طول مساوی استفاده می‌شود. پیاده‌سازی فاصلهٔ همینگ در شبه‌کد شکل ۳ آمده است. این الگوریتم، دو رشتهٔ S و T را به عنوان ورودی گرفته و فاصلهٔ همینگ آنها را برمی‌گرداند.

[1]- Hamming distance

```
intHammingDistance(S[1..M], T[1..N])
{
        If M !=N then return -1
        declare intHammingDistance=0
        for i from 1 to M{
                if S[i] != T[i] then HammingDistance++
        }return HammingDistance
}
```

شکل ۳- شبه‌کد فاصلهٔ همینگ

۲-۳- فاصلهٔ دامرو-لونشتاین

فاصلهٔ دامرو-لونشتاین[1](۱۹۶۴) مشابه فاصلهٔ لونشتاین است. تنها فرق آن این است که فاصلهٔ دامرو-لونشتاین یک فرایند دیگر را مجاز می‌شمارد: ترانهش[2] دو نویسهٔ مجاور. شبه‌کد محاسبهٔ فاصلهٔ دامرو-لونشتاین در شکل ۴ آمده است.

```
intDamerauLevenshteinDistance(char S[1..M], char T[1..N])
{
        declare intd[0..M, 0..N]
        declare inti, j, cost
        for i from 0 to M
                d[i,0] := i //the distance of any first string to an empty
second string
        for j from 0 to N
                d[0, j] := j //the distance of any second string to an empty
first string
        for i from 1 to M
        {
                for j from 1 to N{
                if S[i] = T[j] then cost := 0
                                else cost := 1
                        d[i, j] := minimum(d[i-1, j] + 1, //a deletion
                                                d[i, j-1] + 1, //an
insertion
```

[1]- Damerau–Levenshtein distance
[2]-Transposition

```
                                                    d[i-1, j-1] + 1) //a
substitution
                    if(i > 1 and j > 1 and S[i] = T[j-1] and S[i-1] = M[j])
then
                            d[i, j] := minimum(
                                                    d[i, j],
                                                    d[i-2, j-2] + cost) //
transposition
                    }
            }
            return d[M, N]
}
```

شکل ۴- شبه‌کد فاصلهٔ دامرو-لونشتاین

در این مقاله از روش فاصلهٔ لونشتاین استفاده و در آن تغییراتی ایجاد شد تا جابجایی یک گروه با گروه دیگر به دست آید و نیز مشخص شود که چه مجموعه نویسه‌ای به مجموعه نویسه‌ای دیگر تبدیل شده است. در الگوریتم استاندارد لونشتاین، تغییرات یک نویسه درنظرگرفته می‌شود و درنهایت نیز خروجی الگوریتم برابر فقط حداقل مراحل تغییرات است که در خانه پایین و راست ماتریس موردنظر قرار دارد.

۳- روش کار

در پیکرهٔ زبانی این تحقیق (فارسدات تلفنی) مجموعه‌ای شامل ۲۸۲۵۵۳ ردیف داده قرار دارد. اطلاعات موجود در هر سطر این پیکره شامل موارد زیر است:

- صورت واجی مانند hastid
- صورت آوایی[۱] مانند hastin
- صورت نوشتاری فارسی مانند« هستید»
- زمان شروع و زمان پایان گفتار.

برای این کار، همهٔ سطرهای این پیکره دارای ارزش اطلاعاتی نبودند، چرا که برخی اطلاعات مانند مکث، تپق، خنده، سرفه و ... نیز در این پیکره ذخیره شده است. لذا در مرحله اول، پیکره از این گونه اطلاعات پالایش شد و ۲۱۹۵۸۶ سطر از اطلاعات باقی ماند. از این مقدار نیز، مقادیر زیادی دارای صورت واجی و آوایی یکسانی بودند که این گونه اطلاعات نیز پالایش شد و در نتیجه ۱۶۹۵۰ سطر دارای زنجیره‌های واجی و آوایی متفاوت باقی ماند. سپس الگوریتم تفاوت‌یابی رشته‌هایی که در مرحله قبل توضیح داده شد روی این تعداد داده اعمال گردید و تعداد ۱۳۵۲۷ تغییر به دست آمد.

[۱]- از آنجایی‌که این تفاوت‌ها حاصل اعمال فرایندهای آوایی در یک زبان واحد است، لذا به این صورت‌ها که اشتقاق روساخت از زیرساخت به‌حساب می‌آیند، صورت واجی و آوایی اطلاق می‌شود.

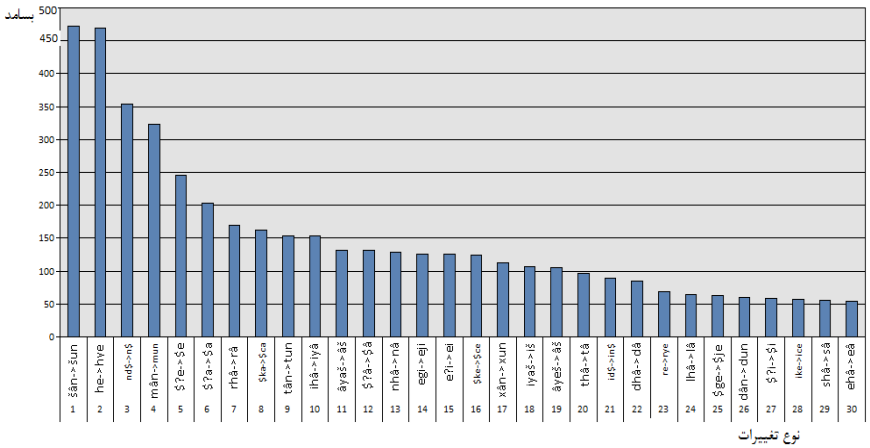
البته بعضی از این تغییرات تکراری نیز بودند و در بررسی دقیق‌تر، تعداد تغییرات برابر ۲۷۹۳ نوع شد. برای بررسی بسامد تغییرات آوایی، تغییرات با تکرار نیز در نظر گرفته شد تا به تحلیل آماری از میزان تغییرات در آواها هنگام صحبت روزمره دست یابیم.

۴- نتایج به دست آمده

تعداد زیادی از این تغییرات بسامد کم(حتی یک بار) و تعداد کمی از این تغییرات بسامد زیاد داشتند. همانطور که در نمودارهای زیر مشاهده می‌شود، درصد کمی از تغییرات، بیشترین میزان توزیع بسامدی را اشغال کردند، یعنی تقریباً ۳۰ تغییر اول بیشترین بسامد را دارند و مابقی تغییرات دارای بسامد و احتمال وقوع کمتری‌اند. نمودار توزیع فراوانی ۳۰ تغییر اول در شکل ۵ و نمودار توزیع فراوانی ۹۰ تغییر اول در شکل ۶ آمده است. جدول ۱ نمودار توزیع فراوانی ۳۰ تغییر اول را با جزئیات بیشتری مشخص می‌سازد. مثالی از تغییرات پربسامد به شرح زیر است[۱]:

- šân→šun
- nd→ n
- mân→mun
- ?e → e
- ?a → a
- rh→ r
- ka → ca
- tân→tun
- ihâ→iyâ
- âyaš→âš
- ?â → â
- nhâ→nâ

[۱]- علامت $ در شکل‌ها و جدول ۱ پیش یا پس از تبدیل آوایی بیانگر جایگاه فرایند آوایی در اول یا آخر کلمه است.

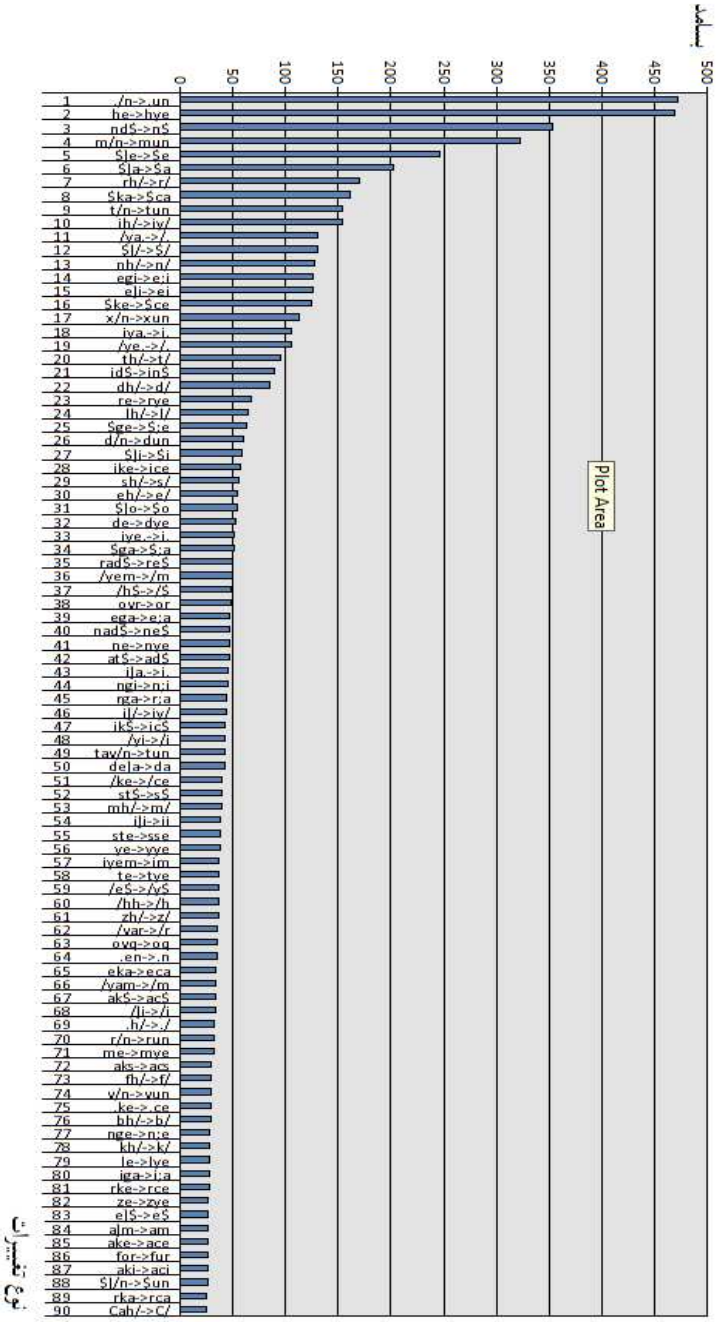


شکل ۵- نمودار توزیع فراوانی ۳۰ تغییر اول

ردیف	تبدیل آوایی	تعداد	ردیف	تبدیل آوایی	تعداد	ردیف	تبدیل آوایی	تعداد
1	šân->šun	472	11	âyaš->âš	131	21	id$->in$	89
2	he->hye	469	12	$?â->$â	131	22	dhâ->dâ	85
3	nd$->n$	354	13	nhâ->nâ	128	23	re->rye	68
4	mân->mun	323	14	egi->eji	126	24	lhâ->lâ	65
5	$?e->$e	246	15	e?i->ei	126	25	$ge->$je	63
6	$?a->$a	203	16	$ke->$c[1]e	124	26	dân->dun	60
7	rhâ->râ	170	17	xân->xun	113	27	$?i->$i	59
8	$ka->$ca	162	18	iyaš->iš	106	28	ike->ice	57
9	tân->tun	154	19	âyeš->âš	105	29	shâ->sâ	55
10	ihâ->iyâ	154	20	thâ->tâ	96	30	ehâ->eâ	54

جدول ۱- نمودار توزیع فراوانی ۳۰ تغییر اول

[1]- در فارسدات تلفنی /k/ صورت زیرساختی قلمداد شده است که در روساخت به دو صورت [k] پسکامی و [c] پیشکامی بازنمایی می‌شود.



شکل ۶: نمودار توزیع فراوانی ۹۰ تغییر اول

۵- نتیجه‌گیری

در این تحقیق ساخت آوایی گونهٔ رسمی و محاوره‌ای زبان فارسی براساس پیکرهٔ زبانی فارسدات تلفنی بررسی شد. در نتیجهٔ این بررسی نوع و بسامد تحولات آوایی در زبان فارسی در چارچوب پیکرهٔ زبانی مورد نظر مشخص شد. نوع تحولات آوایی محدود به فرایندهای جانشینی، حذف و درج بود. این تحولات از مقایسهٔ فاصلهٔ نویسه‌های مربوط به کلمه‌ها در دو سطح واجی و آوایی موجود در فارسدات تلفنی به دست آمد. با انجام تغییراتی از الگوریتم لونشتاین برای مقایسهٔ خودکار فرایندهای جانشینی، حذف و درج استفاده شد. در ارزیابی نتیجهٔ پژوهش مشخص شد که تعداد زیادی از این تغییرات دارای بسامد کم و تعداد کمی از تغییرات دارای بسامد زیاد بودند. شکل‌های ۵ (نیز در جدول ۱) و ۶ به ترتیب توزیع فراوانی ۳۰ و ۹۰ تغییر اول را با جزئیات بیشتری مشخص می‌سازند.

از نتایج این پژوهش می‌توان به لحاظ نظری در توصیف‌های زبان‌شناختی از نظام آوایی زبان فارسی، تهیهٔ منابع محاوره‌ای زبان فارسی و آموزش زبان فارسی به خصوص به غیرفارسی‌زبانان سود جست. از سوی دیگر در فن‌آوری‌های گفتار مانند بازشناسی و بازسازی گفتار، استخراج اطلاعات از متن‌های محاوره‌ای، تبدیل متن به زنجیرهٔ واجی گونهٔ محاوره‌ای زبان فارسی و امکان تبدیل آن به گونهٔ رسمی می‌توان از نتایج این تحقیق استفاده کرد.

منابع

- Bijankhan, M. and others(2003), Tfarsdat – the Telephone Farsi Speech Database. In: *Proc. of the Eurospeech 2003*, pp. 1525–1528.
- Damerau, F.J. (1964)."A Technique for Computer Detection and Correction of Spelling Errors", *Communications of the ACM*, pp. 171-176.
- Hamming, W. R. (1950)." Error Detecting and Error Correcting Codes", *Bell System Technical Journal 29*, pp. 147–160.
- Jin, L. and C. Li(2005)." Selectivity Estimation for Fuzzy String Predicates in Large Data Sets", *VLDB 05: Proceedings of the 31st International Conference on Very* Large Data Bases, pp. 397-408.
- Levenshtein, V.I. (1965)."Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", *Soviet Physics Doklady*, Volume 10, pp. 707–710.

- Li, C. , J. Lu and Y. Lu(2008)." Efficient Merging and Filtering Algorithms for Approximate String Searches", *Proceedings of IEEE 24th International Conference*, pp. 257-266.
- Navarro, G. (2001)." A Guided Tour to Approximate String Matching", *ACM Computing Surveys (CSUR)*, Volume 33, Issue 1, pp. 31-88.
- Stoilos, G. , G. Stamou and S. Kollias(2005)." A String Metric for Ontology Alignment", *Proceedings of International Semantic Web Conference' 2005*, Volume 3729, pp. 624-637.