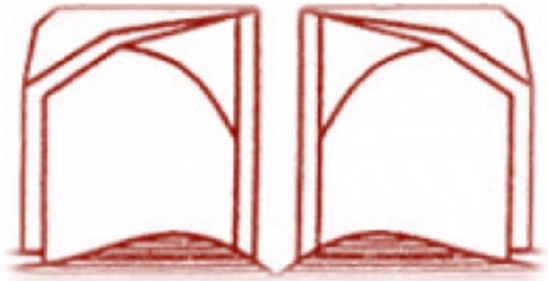


آشنایی با زبان‌شناسی رایانشی

ساخت‌واژه (مورفولوژی)

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین

فهرست

- معرفی ساخت‌واژه
- انواع ساخت‌واژه
  - ساخت‌واژه تصریفی
    - انگلیسی
    - فارسی
  - ساخت‌واژه اشتقاقی
- تجزیه ساخت‌واژی
  - عبارت باقاعده (Regular Expression)
  - Finite State Automata (FSA)
  - Finite State Transducer (FST)
- ریشه‌یابی بدون واژگان (Porter Stemmer)
- خطایاب املایی و روش Minimum Edit Distance

ساخت‌واژه (مورفولوژی) . . .

- **تعریف**
  - چگونه کلمات از واحدهای کوچک‌تر به نام تک‌واژ (Morphemes) تشکیل می‌شوند
- **تک‌واژ (Morpheme)**
  - کوچک‌ترین واحد دربردارنده معنا در زبان
- **دو دستهٔ عمدهٔ تک‌واژها**
  - ریشه‌ها (Stems): تک‌واژ اصلی کلمه که دربردارندهٔ معنای اصلی کلمه است
  - وندها (Affixes): تک‌واژهایی که معناهای دیگری به ریشه اضافه می‌کنند

ساخت‌واژه (مورفولوژی) . . .

- وندها (Affixes)
  - پسوندها (Suffixes/Postfixes)
    - رفته‌ام (فارسی)
    - Books (انگلیسی)
  - پیشوندها (Prefixes)
    - می‌روم (فارسی)
    - Asymmetric (انگلیسی)
  - میان‌وندها (Infixes)
    - در زبان‌هایی مانند اسپانیایی و عربی
    - مثال: Osquítar (اسپانیایی)
  - پسوند-پیشوند (Circumfix)
    - در برخی زبان‌ها مانند آلمانی، مالایی و گرجی
    - مثال: kabaddangan به معنی مفید (فیلیپین)

Affix	Example	Schema	Description
Prefix	un-do	prefix-stem	Appears before the stem
Suffix/postfix	look-ing	stem-suffix	Appears after the stem
Suffixoid[1]**/semi-suffix**[2]	cat-like	stem-suffixoid	Appears after the stem, but is only partially bound to it
Infix	Minne(flippin')sota	st(infix)em	Appears within a stem — common in Borneo-Philippines languages
Circumfix	en(light)en	circumfix(stem)circumfix	One portion appears before the stem, the other after
Interfix	speed-o-meter	$stem_a$-interfix-$stem_b$	Links two stems together in a compound
Duplifix	money~shmoney	stem~duplifix	Incorporates a reduplicated portion of a stem (may occur before, after, or within the stem)
Transfix	Maltese: k(i)t(e)b "he wrote" (compare root *ktb* "write")	s(transfix)te(transfix)m	A discontinuous affix that interleaves within a discontinuous stem
Simulfix	mouse → mice	stem\simulfix	Changes a segment of a stem
Suprafix	**pro**duce (noun) pro**du**ce (verb)	stem\suprafix	Changes a suprasegmental feature of a stem
Disfix	Alabama: tipli "break up" (compare root *tipasli* "break")	st(disfix)m	The elision of a portion of a stem

ساخت‌واژه

- **ساخت‌واژه پیوندی (Concatenative Morphology)**
  - وندها به ریشه متصل می‌شوند
  - مثال: فارسی- کتاب‌ها، می‌خورم
  - مثال: انگلیسی: Notes، Untouchable
- **ساخت‌واژه غیرپیوندی (Non-Concatenative Morphology)**
  - وندها به نحو پیچیده‌تری با ریشه ترکیب می‌شوند
  - مثال: در زبان تاگالو (فیلیپین)- ریشه hingi (قرض) با وند um= humingi (میان‌وند)
  - نوع دیگر: Templatic (root-and-pattern) Morphology
    - زبان‌های عربی و عبری
    - مثال: CCC یک کلمه (مانند lmd = یادگیری) به صورت CaCaC (مانند lamad= او مطالعه کرد)

چرا ساخت‌واژه؟ . . .

- **در زبان‌هایی که ساخت‌واژه پیچیده دارند، لیست کردن تمام اشکال مختلف یک کلمه در عمل غیرممکن است**
  - **مثال از فارسی**
    - انقلابی‌ترینهایشانند: انقلابی + ترین + ها + شان + َند
    - در فارسی یک اسم مانند "کتاب" دارای ۱۸۹ صورت مختلف صرفی است.
  - **مثال از ترکی**
    - Uygarlaştıramadıklarımızdanmışsınızcasına
      - "(behaving) as if you are among those whom we could not civilize"
    - تا ۱۹ وند به یک کلمه وصل می‌شود! (در انگلیسی تا ۵ وند)
    - افعال بدون وندهای اشتقاقی تا ٤٠.٠٠٠ صورت دارند!
      - افزودن وندهای اشتقاقی مانند علت، افزایش تعداد تا بی نهایت حالت
    - زبان‌هایی مانند ترکی که یک رشته طولانی به عنوان وند دارند، زبان‌های Agglutinatibe می‌گویند
  - **انگلیسی**
    - بلندترین واژه autocyberconceptualizations

چرا ساخت‌واژه؟

- تجزیه یک کلمه به تک‌واژهای تشکیل دهندهٔ آن از جمله ریشه
  - Morphological Parsing
    - Give + s ⬅ Gives
  - ریشه‌یابی (Stemming): کاربرد وسیع در بازیابی اطلاعات
    - Fox ⬅ Foxes
  - سرواژه‌سازی (Lemmatization): تشخیص یکسان بودن ریشه چند واژه علی‌رغم ظاهر متفاوت آنها
    - sing ⬅ sings, sang, sung
  - واحدسازی (Tokenization)
    - I + am ⬅ I'm
- لیست کردن تمام شکل‌های مختلف لغات در مجموعه واژگان از لحاظ محاسباتی غیربهینه است
  - تولید صورت‌های مختلف (صرفی و اشتقاقی) یک کلمه با داشتن ریشهٔ آن
- تولید تلفظ خودکار کلمات خارج از واژگان

انواع ساخت‌واژه . . .

ساخت‌واژه تصریفی (Inflectional)

- **ترکیب ریشهٔ کلمه با وندها به طوری که مقولهٔ نحوی و معنای کلی کلمه تغییر نکند**
  - مانند تطابق شخص و شمار و ساختن زمان فعل
  - مثال: cats (در مقابل cat) - بیانگر شکل جمع و اشاره به تعداد نامحدودی گربه
    - مقوله پاره‌گفتاری اصلی (اسم) و معنای پایه (گربه‌سان خانگی) تغییری نکرده‌اند
  - در انگلیسی: وندهای فعلی: -ing, -ed, -s، ی اسمی: -s ، و صفتی -er, -est

ساخت‌واژه اشتقاقی (Derivational)

- **ترکیب ریشه کلمه با وندها به طوری که معنای جدیدی تولید شود و احتمالاً مقولهٔ نحوی کلمه تغییر کند**
  - همراه با تغییرات پاره‌گفتار: مانند racial (نژادی) و racist (نژادپرست)
    - گرفته شده از یک ریشه race - پاره‌گفتارهای متفاوت (صفت در مقابل اسم-صفت) و معنای متفاوت
  - تغییر تلفظ یا جابجایی تکیه مثل electric در مقابل electricity
  - در انگلیسی: پیشوندها و پسوندها مانند - re-, pre-, -ial, -ism, -ish, -ity, -tion, -ness, -ment, -ious, -ify, -ize, و غیره.

انواع ساخت‌واژه . . .

- ساخت‌واژه ترکیبی (Compounding)
  - ترکیب ریشه ها با همدیگر
  - مثال: doghouse
- ساخت‌واژه واژه‌بستی (Cliticization)
  - ترکیب یک ریشه با واژه‌بست (Clitic)
    - واژه‌بست (Clitic): یک تک‌واژ که از نظر نحوی مانند یک واژه عمل می‌کند اما شکل کوتاه شده است و به واژه دیگری متصل می‌شود
  - مثال در انگلیسی: I've
  - مثال در عربی: القسم

Full Form	Clitic	Full Form	Clitic
am	'm	have	've
are	're	has	's
is	's	had	'd
will	'll	would	'd

ابهام (he's)

ال	*Al#*	Definite Article
و	*w#*	Conjunction, coordinating
ف	*f#*	Conjunction, subordinating
ل	*l#*	Preposition
ب	*b#*	Preposition
ك	*k#*	Preposition
س	*s#*	Future verbal particle
ي	*+y*	POSS_PRON_1S/ PRON_1S
ا	*+A*	PRON_1P
ني	*+ny*	IVSUFF_DO/PRON_1S/PVSUFF_DO
ك	*+k*	POSS_PRON_2MS/ PRON_2MS
كما	*+kmA*	POSS_PRON_2D/ PRON_2D
كم	*+km*	POSS_PRON_2MP/ PRON_2MP
كن	*+kn*	POSS_PRON_2FP/ PRON_2FP
ه	*+h*	POSS_PRON_3MS/ PRON_3MS
ها	*+ha*	POSS_PRON_3FS/ PRON_3FS
هما	*+hmA*	POSS_PRON_3D/ PRON_3D
هن	*+hn*	POSS_PRON_3FP/ PRON_3FP
هم	*+hm*	POSS_PRON_3MP/ PRON_3MP
نا	*+nA*	POSS_PRON_1P/ PRON_1P

ساخت‌واژه: تصریفی . . .

- در زبان انگلیسی
  - اسم: اتصال پسوند s (برای جمع) و ‘s (برای مالکیت) برای اسامی با قاعده
    - Regular Noun + s → Plural Noun
    - book + s → books
    - fox + s → foxes
  - اسم: موارد بی قاعده
  - فعل: اتصال پسوندهای ed، ing و s به انتهای فعل
    - Regular Verb stem + ed → past verb
    - walk + ed → walked
    - beg + ed → begged
  - فعل: موارد بی قاعده (حدود ۲۵۰ مورد)
    - تفاوت در شکل گذشته فعل eat/ate و cut/cut

	Regular Nouns		Irregular Nouns	
Singular	cat	thrush	mouse	ox
Plural	cats	thrushes	mice	oxen

Morphological Form Classes	Regularly Inflected Verbs			
Stem	walk	merge	try	map
-s	walks	merges	tries	maps
-ing	Walking	merging	trying	mapping
-ed	walked	merged	tried	mapped

ساخت‌واژه: تصریفی . . .

- در زبان اسپانیایی
  - برای فعل amar (عشق ورزیدن)- ۵۰ صورت مختلف

	Present Indicative	Imperfect Indicative	Future	Preterite	Present Subjnct.	Conditional	Imperfect Subjnct.	Future Subjnct.
1SG	amo	amaba	amaré	amé	ame	amaría	amara	amare
2SG	amas	amabas	amarás	amaste	ames	amarías	amaras	amares
3SG	ama	amaba	amará	amó	ame	amaría	amara	amáreme
1PL	amamos	amábamos	amaremos	amamos	amemos	amaríamos	amáramos	amáremos
2PL	amáis	amabais	amaréis	amasteis	améis	amaríais	amarais	amareis
3PL	aman	amaban	amarán	amaron	amen	amarían	amaran	amaren

اول شخص مفرد

دوم شخص جمع

ساخت‌واژه: تصریفی . . .

- در زبان فارسی
  - اتصال پسوندهای جمع، نکره، اضافه، ضمایر متصل و واژه بست‌های ربطی به انتهای اسم، صفت، ضمیر و عدد
  - شکل‌های مختلف تصریف فعل با زمان‌ها و شخص‌های مختلف

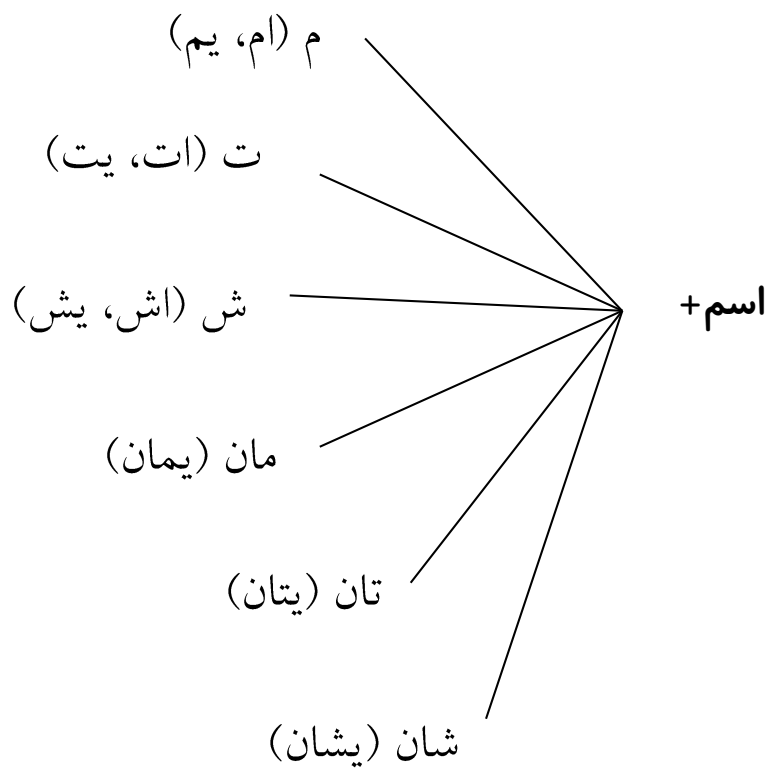
- اتصال تک‌واژ جمع (ان)
  - مرد + ان : مردان
  - گدا + ان : گدایان
  - فرشته + ان : فرشتگان

- اتصال تک‌واژ نکره‌ساز (ای)
  - مرد + ای : مردی
  - گدا + ای : گدایی
  - فرشته + ای : فرشته‌ای

ساخت‌واژه: تصریفی در فارسی . . .

- **کلمات غیرفعل . . .**

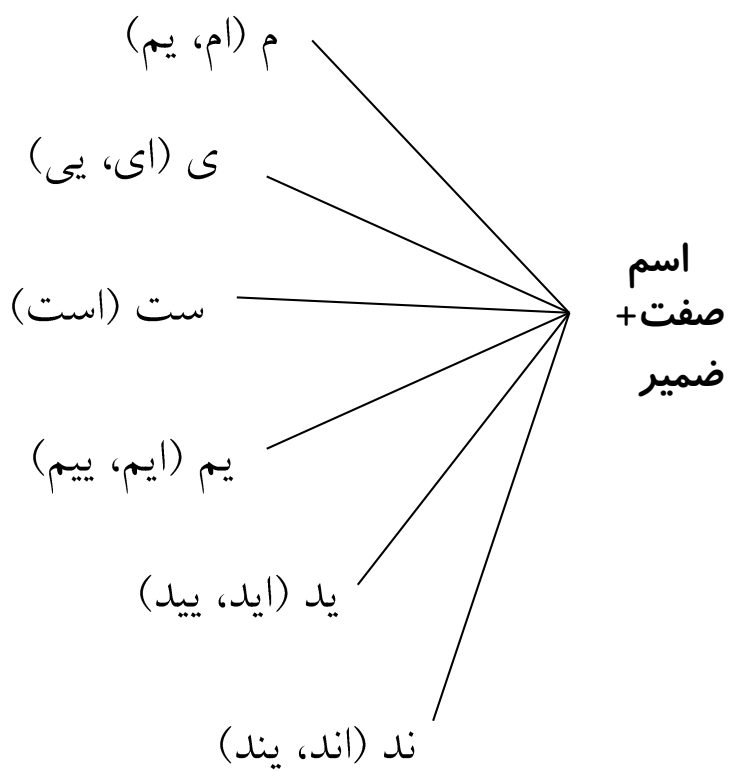
چشمم	خانه‌ام
چشمت	خانه‌ات
چشمش	خانه‌اش
چشممان	خانه‌مان
چشمتان	خانه‌تان
چشمشان	خانه‌شان



ساخت‌واژه: تصریفی در فارسی . . .

- کلمات غیرفعل

مشغولم	خسته‌ام
مشغولی	خسته‌ای
مشغولست	خسته‌است
مشغولیم	خسته‌ایم
مشغولید	خسته‌اید
مشغولند	خسته‌اند



ساخت‌واژه: تصریفی در فارسی . . .

- افعال اخباری/گزارشی (Indicative)
  - انجام گرفتن/نگرفتن فعلی را گزارش می‌کند و مختص رویدادهای واقعی است
  - Infinitival: رفتن
  - PresentParticiple: رونده
  - PastParticiple: رفته
  - Present: می‌روم
  - Preterite: رفتم
  - Imperfect: می‌رفتم
  - Perfect: رفته‌ام
  - Pluperfect: رفته بودم
  - Compound Imperfect: می‌رفته‌ام
  - DoubleCompound: رفته بوده‌ام
  - Future: خواهم رفت

ساخت‌واژه: تصریفی در فارسی . . .

- **افعال التزامی/پیرو (Subjunctive)**
  - زمانی به کار می‌رود که انجام فعل واقعیت ندارد و می‌خواهیم پنداشت، احتمال، میل، قصد، دستور، خواهش و مانند آن را بیان کنیم (امور ناواقعی و ذهنی)
  - Present: بروم
  - CompoundPast: رفته باشم
- **افعال امری (Imperative)**
  - برای بیان دستور، خواهش، درخواست، هشدار، راهنمایی و مانند آن به کار می‌رود
  - زیرشاخه‌ای از وجه التزامی‌ست و از آن ساخته می‌شود با این تفاوت که شناسه‌ی دوم شخص مفرد (تو) پوچ است
  - Present: برو

ساخت‌واژه: اشتقاقی . . .

در انگلیسی

- اتصال پسوندها و پیشوندهای مختلف به فعل، اسم، صفت و قید
  - compute + ation → computation
  - un + believe + able → unbelievable

Nominalization
ساختن اسم از روی فعل/صفت

Suffix	Base Verb/Adjective	Derived Noun
-ation	computerize (V)	computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

Suffix	Base Noun/Verb	Derived Adjective
-al	computation (N)	computational
-able	embrace (V)	embraceable
-less	clue (N)	clueless

ترکیب صورت‌های مختلف ساخت‌واژه

- بلندترین واژه انگلیسی autocyberconceptualizations
  - پیشوند اشتقاقی auto-
  - دو ریشه که با هم ترکیب شده‌اند (cyber و concept، هرچند گاهی cyber را پیشوند می‌گویند)
  - سه پسوند اشتقاقی (-ual, -ize, -ation)
  - پسوند تصریفی جمع برای اسامی، -s

ساخت‌واژه: اشتقاقی

- در فارسی
  - اتصال پسوندها و پیشوندهای مختلف به اسم، صفت و قید
    - کار + گر ← کارگر
    - دانش + مند ← دانشمند
    - دانش + گاه ← دانشگاه
    - زیبا + رو ← زیبارو

تجزیه ساخت‌واژی (Morphological Parsing)

- تجزیه یک کلمه به تک‌واژهای تشکیل دهندهٔ آن

- نیازمندی‌های تجزیه ساخت‌واژی
  - Lexicon: مجموعه‌ای از ریشه‌های کلمات و وندها به همراه اطلاعاتی در مورد آنها
    - کتاب (اسم)
  - Morphotactics: مجموعه‌ای از قوانین مورفولوژی که نوع وندهای قابل اتصال به هر ریشه و ترتیب اتصال آنها را بیان می‌کند
    - شکل جمع اسم‌ها: ریشه (اسم) + (ها | ان)
  - Orthographic rules: مجموعه‌ای از قوانین املایی که نوع تغییر در املای کلمه را هنگام اتصال وندها بیان می‌کند
    - در انگلیسی: y ↤ ie (city ↤ cities)
    - در فارسی: ه ↤ گ (همسایه ↤ همسایگان)

عبارت باقاعده/منظم (Regular Expression)

- زبانی برای بیان مجموعه‌ای از رشته‌ها
  - نوعی نشانه‌گذاری جبری برای بیان دنباله‌های (الگوهای) مشخص از رشته‌ها

- کاربرد: جستجو در متن
  - می‌خواهیم کلمه «کتاب» را در یک متن پیدا کنیم- نحوه جستجو؟
    - هم به صورت «کتاب» و هم به صورت «کتاب‌ها» یا به صورت «کتابم»
  - اگر بخواهیم همه عددهایی را پیدا کنیم که به همراه آن علامت «$» است چی؟
    - مانند $2.3 یا $395000
  - عبارتی مانند «خخخخخخخخخ» در متون شبکه‌های اجتماعی امروزی
    - تعداد «خ»ها هر عددی می تواند باشد: خخ، خخخ، خخخخ، ....

عبارت باقاعده . . .

علائم مربوط به زبان Perl است که در زبان های دیگر (مانند پایتون) به صورت مشابه وجود دارند

موارد عادی

- استفاده از «/عبارت/»

RE	Example Patterns Matched
/woodchucks/	"interesting links to woodchucks and lemurs"
/a/	"Mary Ann stopped by Mona's"
/Claire_says,/	"Dagmar, my gift please," Claire says,"
/DOROTHY/	"SURRENDER DOROTHY"
/!/	"You've left the burglar behind again!" said Nori

انفصال (Disjunction)

- به صورت [عبارت]
- یکی از موارد

در انگلیسی تفاوت در حروف کوچک و بزرگ = Case Sensitive

RE	Match	Example Patterns
/[wW]oodchuck/	Woodchuck or woodchuck	"Woodchuck"
/[abc]/	'a', 'b', *or* 'c'	"In uomini, in soldati"
/[1234567890]/	any digit	"plenty of 7 to 5"

- علامت «|» به معنی «یا»
  - مثال: /گربه|موش/ برای جستجوی «گربه» یا «موش»
- هر عددی /[0123456789]/ که برابر است با /[0-9]/
- هر کاراکتر با حروف بزرگ /[ABCDEFGHIJKLMNOPQRSTUVWXYZ]/

بازه (Range)

RE	Match	Example Patterns Matched
/[A-Z]/	an uppercase letter	"we should call it 'Drenched Blossoms'"
/[a-z]/	a lowercase letter	"my beans were impatient to be hoed!"
/[0-9]/	a single digit	"Chapter 1: Down the Rabbit Hole"

عبارت باقاعده . . .

نفی (Negation)

- به صورت [عبارت ۱ ^]
  - معنی هر چیزی به جز «عبارت ۱»

RE	Match (single characters)	Example Patterns Matched
[^A-Z]	not an uppercase letter	“Oyfn pripetchik”
[^Ss]	neither ‘S’ nor ‘s’	“I have no exquisite reason for’t”
[^\.]	not a period	“our resident Djinn”
[e^]	either ‘e’ or ‘^’	“look up ^ now”
a^b	the pattern ‘a^b’	“look up a^ b now”

- توجه: علامت ^ به صورت /عبارت ۲^/ به معنی اول جمله و در غیر این دو حالت به معنی خود کاراکتر ^ است

اختیاری بودن (Optionality)

- به صورت /?/
- به معنی کاراکتر قبلی یا هیچی
  - صفر یا یکی از کاراکتر قبلی

RE	Match	Example Patterns Matched
woodchucks?	woodchuck or woodchucks	“woodchuck”
colou?r	color or colour	“colour”

عبارت باقاعده . . .

Kleene اسم پیشنهاد دهنده عبارات باقاعده در ۱۹۵۶

Kleene *

- به صورت /*عبارت/
- تکرار کاراکتر قبل از * در «عبارت» به تعداد صفر یا هر تعداد دیگر
- مثال: /a*/
  - تعداد صفر یا هر عدد از کاراکتر a مثل aa، aaaa، ...
- مثال: /[ab]*/
  - مانند aaaaa، ababab، bbb، ...
- مثال: هر عدد صحیح؟؟
  - پاسخ /[1-9][0-9]*/ (اعداد غیرصفر: /[1-9][0-9]*/)

Kleene +

- به صورت /+عبارت/
- تکرار کاراکتر قبل به تعداد یک یا هر تعداد دیگر
- مثال: +خ ⇇ خ، خخ، خخخ، ...

عبارت باقاعده . . .

نویسه عام (Wildcard)

RE	Match	Example Patterns
/beg.n/	any character between *beg* and *n*	begin, beg'n, begun

- به صورت /./
- به معنای هر کاراکتری (به غیر از سطر جدید)
- عبارت «.\» به معنی خود «علامت نقطه» است
- مثال: یافتن پاراگرافی که دو بار کلمه «ایران» در آن تکرار شده باشد
  - پاسخ: /ایران*. ایران/

لنگر (Anchor)

- اول سطر: /^عبارت ۱/
  - مثال: /^Hi/ فقط Hi در اول جمله
- پایان سطر: /عبارت۲$/
- اطراف کلمه یا عبارت: \b
  - مثال: /\bthe\b/ فقط کلمه the و نه کلماتی مانند other

عبارت باقاعده . . .

- پرانتز: برای گروه‌بندی و اولویت‌گذاری
  - مثال: /کتاب(م|ت|ش|ها)/ ← کتابم، کتابش، کتابت، کتاب‌ها
- شمارش تعداد
  - به صورت {n} به معنی تعداد n تکرار از عبارت قبلی
  - به صورت {n, m} به معنی از تعداد n تا m تکرار از عبارت قبلی
  - به صورت {n,} به معنی از تعداد حداقل n تکرار از عبارت قبلی
    - مثال: /.*d\.{1,}/ به معنی هر تعداد کاراکتر که به d ختم می‌شود و پس از آن یک یا بیشتر نقطه است
- اولویت به ترتیب
  - پرانتز
  - شمارشی‌ها: ? * + {}
  - دنباله‌ها و لنگرها
  - انفصال با |

عبارت باقاعده . . .

سایر عملگرها

RE	Expansion	Match	Example Patterns
\d	[0-9]	any digit	Party␣of␣5
\D	[^0-9]	any non-digit	Blue␣moon
\w	[a-zA-Z0-9_]	any alphanumeric or underscore	Daiyu
\W	[^\w]	a non-alphanumeric	!!!!
\s	[␣\r\t\n\f]	whitespace (space, tab)	
\S	[^\s]	Non-whitespace	in␣Concord

RE	Match	Example Patterns Matched
*	an asterisk "*"	"K*A*P*L*A*N"
\.	a period "."	"Dr. Livingston, I presume"
\?	a question mark	"Why don't they come and lend a hand?"
\n	a newline	
\t	a tab	

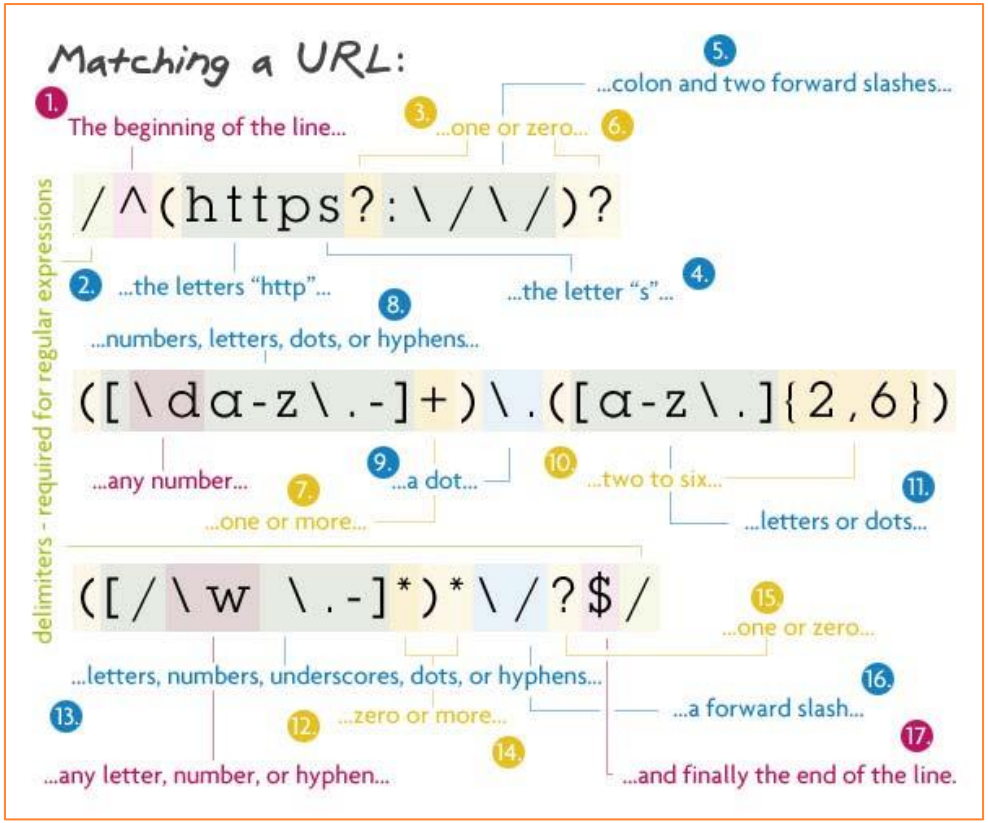
عبارت باقاعده . . .

- مثال . . .
  - /(^ | [^a-zA-Z])[tT]he([^a-zA-Z] | $)/
    - جستجوی the به صورت‌های مختلف
      - آمدن در اول جمله یا وسط جمله
      - حرف اول بزرگ یا کوچک
    - فقط the به تنهایی و عدم اتصال به کاراکتر دیگر مانند در other
  - \b[A-Z0-9._%+\-]+@[A-Z0-9.-]+\.[A-Z]{2,}\b
    - جستجوی آدرس ایمیل در یک متن

علامت - به تنهایی در داخل [ ] معنی دارد و نیاز است خود علامت – مشخص شود

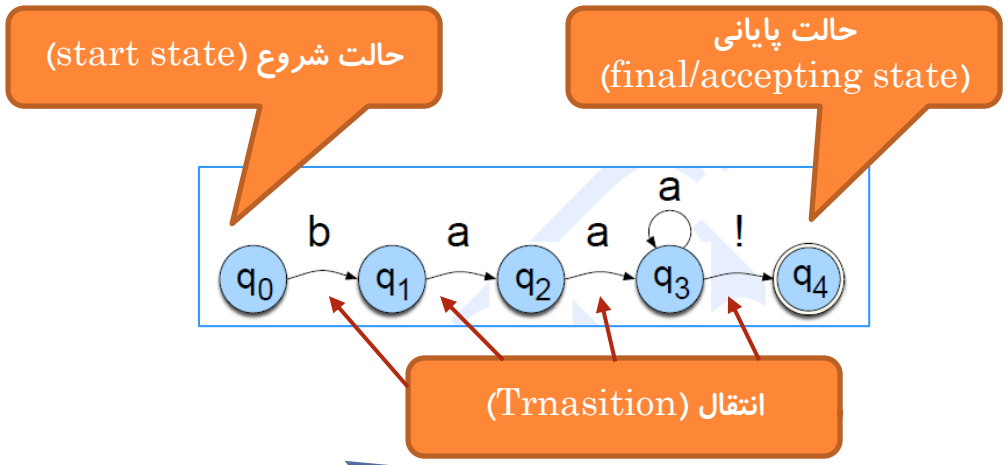
عبارت باقاعده

- مثال
  - /^(https?:\/\/)?([\da-z.\-]+)\.([a-z\.]{2,6})([\/\w .\-]*)*\/?$/
    - جستجوی آدرس اینترنتی (URL)



اتوماتای حالت متناهی (FSA) . . .

- **اتوماتای حالت متناهی Finite-State Automaton (FSA)**
  - عبارت باقاعده/منظم (Regular Expression) روشی برای نمایش این نوع اتوماتا

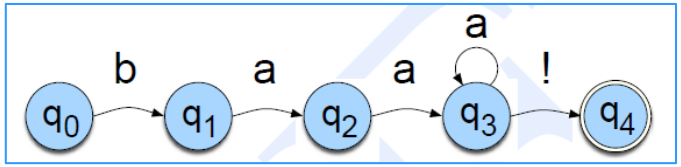


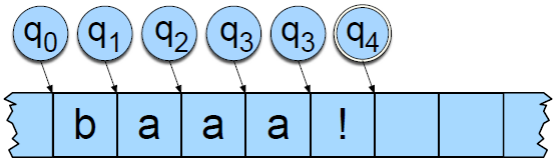
- **مثال: زبان گفتاری گوسفند!**
  - baa!
  - baaa!
  - baaaa!
  - baaaaa!
  - ...
  - معادل عبارت منظم: /baa$^+$!/

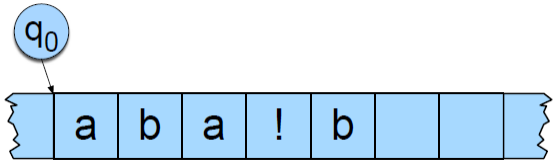
ماشین حالت
اتوماتای حالت متناهی

اتوماتای حالت متناهی (FSA) . . .

- عملکرد اتوماتا (در مثال)
  - کاراکترهای ورودی به صورت نواری متحرک یکی یکی پردازش می شود
  - متناسب با آمدن کاراکترها، بین حالات اتوماتا جابجا می شود
  - در صورت وارد شدن به حالت پایانی، رشته پذیرفته شده است (Accept)
  - در صورت پایان یافتن کاراکترها و عدم وارد شدن به حالت پایانی، رشته ورودی پذیرفته نشده است (Reject)







جدول انتقال حالت
(State Transition Table)

State	Input b	a	!
0	1	0	0
1	0	2	0
2	0	3	0
3	0	3	4
4:	0	0	0

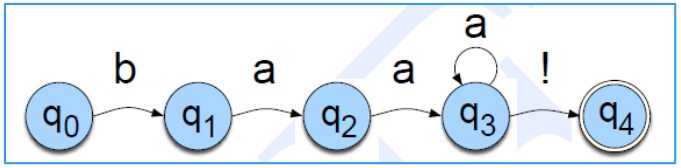
وقتی در حالت ۳ باشیم و وردی a باشد به حالت ۳ می‌رویم و اگر ورودی ! باشد به حالت ۴ می‌رویم

اتوماتای حالت متناهی (FSA) . . .

تعریف

- Q: یک مجموعه متناهی با تعداد N حالت: $q_0, q_1, \ldots, q_{N-1}$
- Σ: مجموعه‌ای متناهی حاوی کاراکترهای ورودی
- q_0: حالت اولیه (شروع)
- F: مجموعه حالات پایانی، $F \subseteq Q$
- δ(q,i): تابع انتقال حالت– انتقال از حالت $q \in Q$ با آمدن ورودی $i \in \Sigma$ و تولید حالت بعدی $q' \in Q$

در مثال قبل



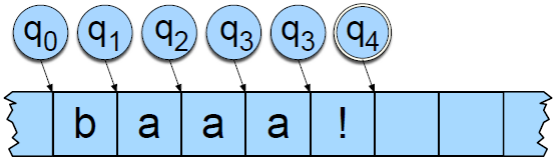
	Input		
State	b	a	!
0	1	∅	∅
1	∅	2	∅
2	∅	3	∅
3	∅	3	4
4:	∅	∅	∅

- $Q=\{q_0, q_1, q_2, q_3, q_4\}$ و $N=4$
- $\Sigma=\{a, b, !\}$
- $F=\{q_4\}$
- δ(q,i)= جدول انتقال حالت

اتوماتای حالت متناهی (FSA) . . .

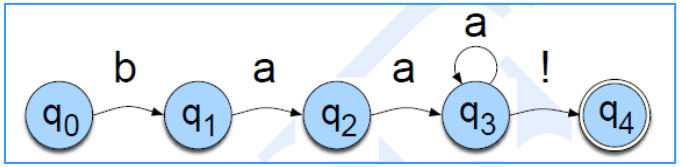
- الگوریتم پذیرش رشته
  - الگوریتم قطعی (Deterministic)
    - می‌داند با هر ورودی به طور قطعی چگونه برخورد کند
    - با آمدن هر ورودی، یک انتخاب بیشتر وجود ندارد

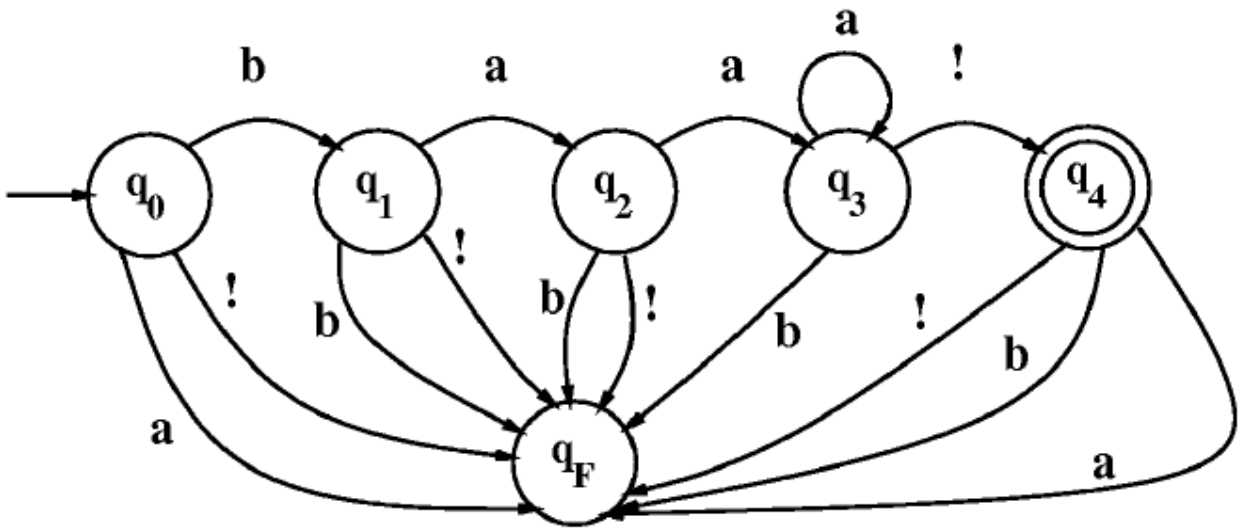
```
function D-RECOGNIZE(tape, machine) returns accept or reject

  index ← Beginning of tape
  current-state ← Initial state of machine
  loop
    if End of input has been reached then
      if current-state is an accept state then
        return accept
      else
        return reject
    elsif transition-table[current-state,tape[index]] is empty then
      return reject
    else
      current-state ← transition-table[current-state,tape[index]]
      index ← index + 1
  end
```



اتوماتای حالت متناهی (FSA) . . .

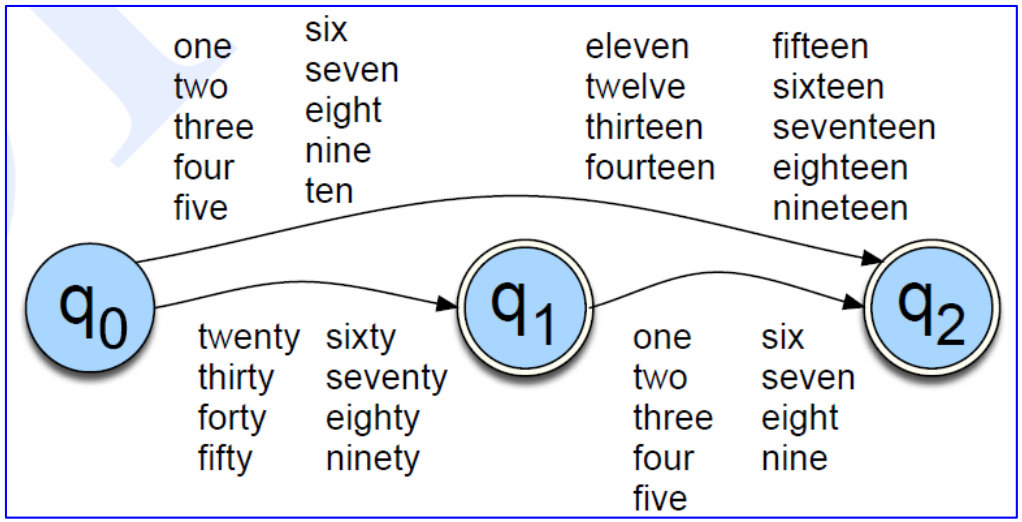
- حالت شکست/جذب (Fail/Sink State)
  - ایجاد پاسخ در آمدن همه نوع ورودی از جمله ورودی‌های غیرمعتبر



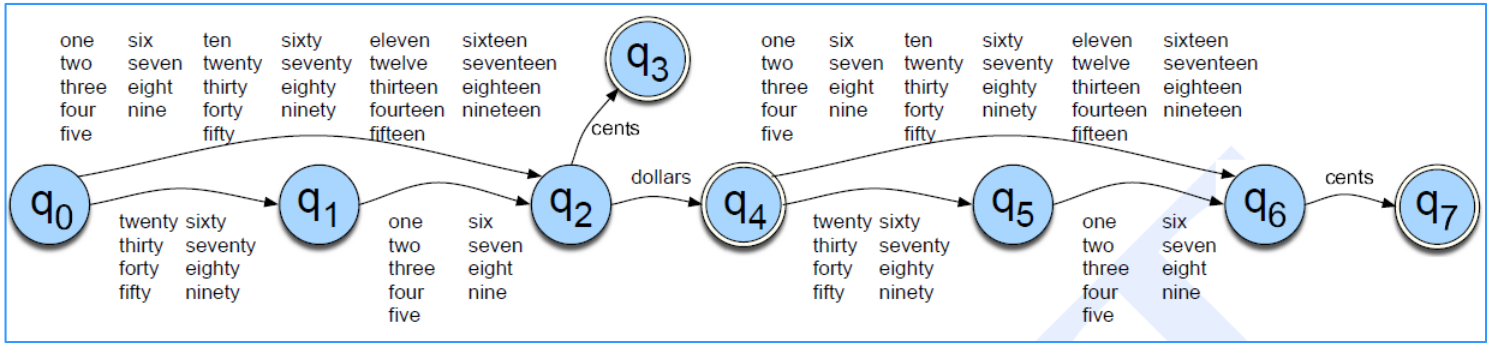


اتوماتای حالت متناهی (FSA) . . .

- مثال: تولید اعداد ۱ تا ۹۹
  - مانند ۵، ۱۲، ۶۸

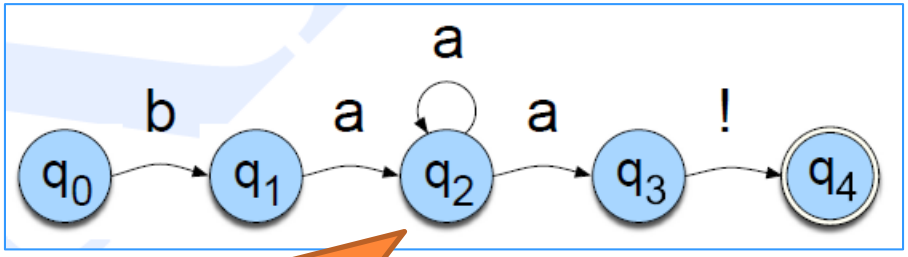


- اضافه کردن عبارت Cent (مشابه ریال) و Dollar (مشابه تومان)
  - مانند ۲۵ دلار و ۷۸ سنت



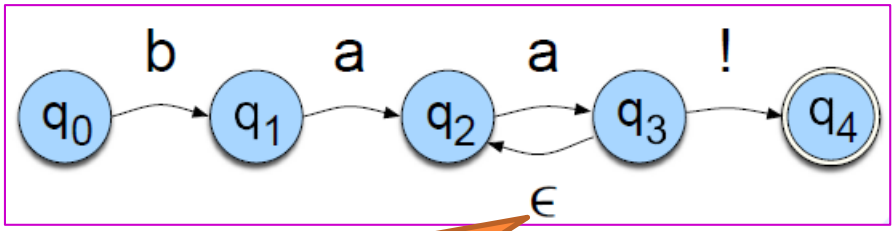
اتوماتای حالت متناهی (FSA) . . .

- اتوماتای غیرقطعی (NFSA: Non-deterministic FSA)
  - عدم قطعیت در انتخاب حالت بعدی با آمدن ورودی (در برخی حالت‌ها)



	Input			
State	b	a	!	ε
0	1	0	0	0
1	0	2	0	0
2	0	2,3	0	0
3	0	0	4	0
4:	0	0	0	0

انتخاب بیش از یک حالت با آمدن ورودی a

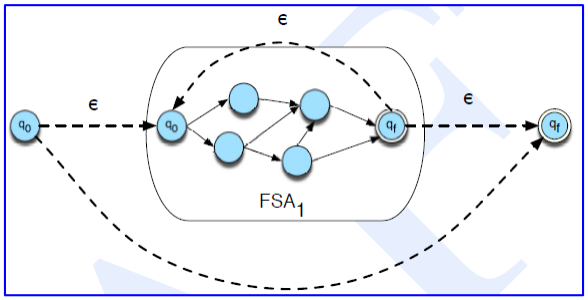


ε-transitions
بدون توجه به ورودی، از حالت q3 به حالت q2 برویم

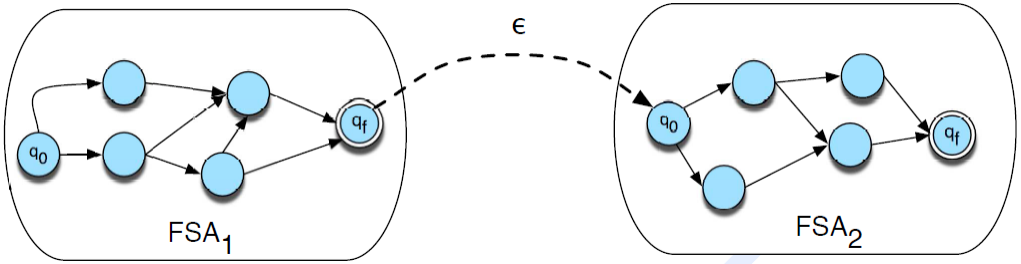
- می‌توان هر اتوماتای غیرقطعی را اتوماتای قطعی تبدیل کرد

اتوماتای حالت متناهی (FSA) . . .

- کاربرد در عبارات با قاعده . . .
  - تعریف زبان‌های باقاعده (Regular Languages)
  - Kleene * (closure)

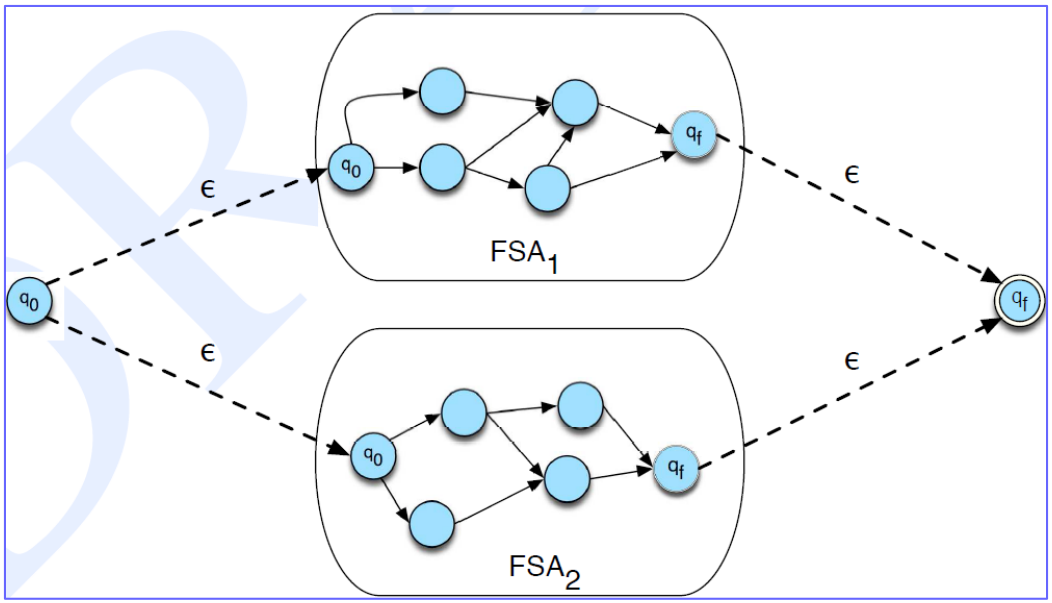


  - اتصال (concatenation) دو اتوماتا



اتوماتای حالت متناهی (FSA)

- کاربرد در عبارات با قاعده
  - Disjunction (union)
    - همان «|»

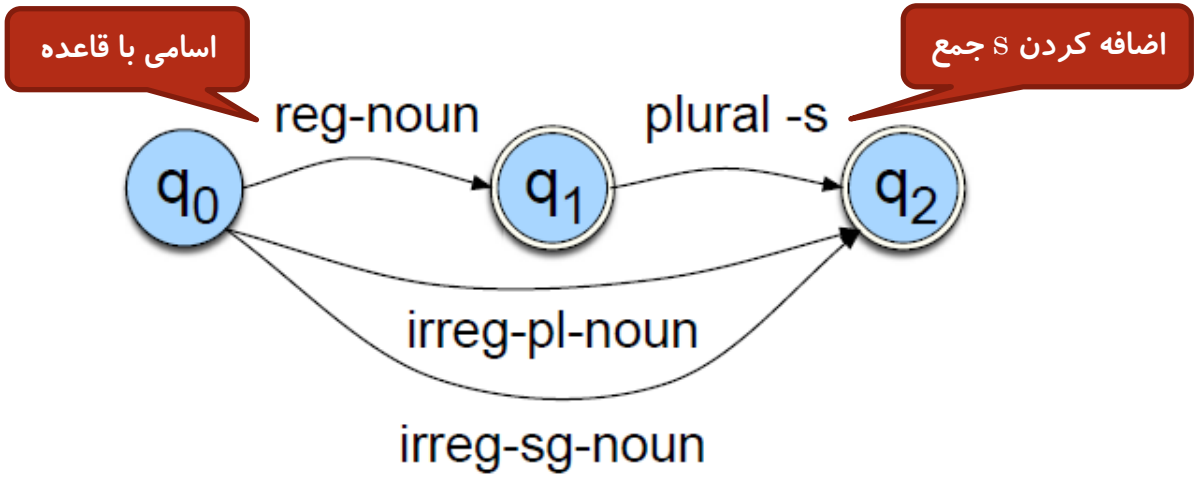


اتوماتای حالت متناهی در ساخت‌واژه . . .

- یادآوری
  - نیازمندی‌های تجزیه ساخت‌واژی
    - Lexicon: مجموعه‌ای از ریشه‌های کلمات و وندها به همراه اطلاعاتی در مورد آنها
      - کتاب (اسم)
    - Morphotactics: مجموعه‌ای از قوانین مورفولوژی که نوع وندهای قابل اتصال به هر ریشه و ترتیب اتصال آنها را بیان می‌کند
      - شکل جمع اسم‌ها: ریشه (اسم) + (ها | ان)
    - Orthographic rules: مجموعه‌ای از قوانین املایی که نوع تغییر در املای کلمه را هنگام اتصال وندها بیان می‌کند
      - در انگلیسی: y ↤ ie (city ↤ cities)
      - در فارسی: ه ↤ گ (همسایه ↤ همسایگان)
- نحوه ساختن morphotacticها؟؟
  - استفاده از اتوماتای حالت متناهی

اتوماتای حالت متناهی در ساخت‌واژه . . .

- یک اتوماتا برای پذیرش اسامی مفرد و جمع انگلیسی
  - تصریفی

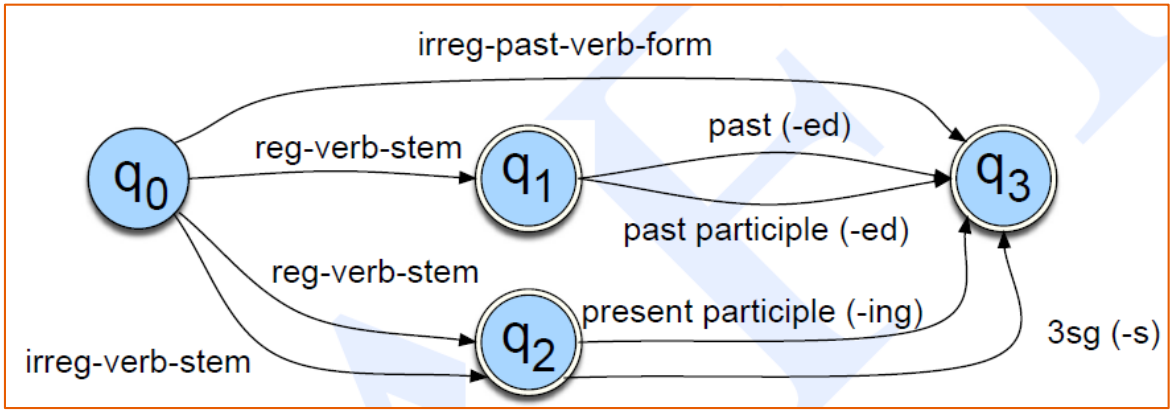


اسامی با قاعده و بی قاعده در واژگان قرار دارند

reg-noun	irreg-pl-noun	irreg-sg-noun	plural
fox	geese	goose	-s
cat	sheep	sheep	
aardvark	mice	mouse	

اتوماتای حالت متناهی در ساخت‌واژه . . .

- یک اتوماتا برای پذیرش افعال انگلیسی
  - تصریفی

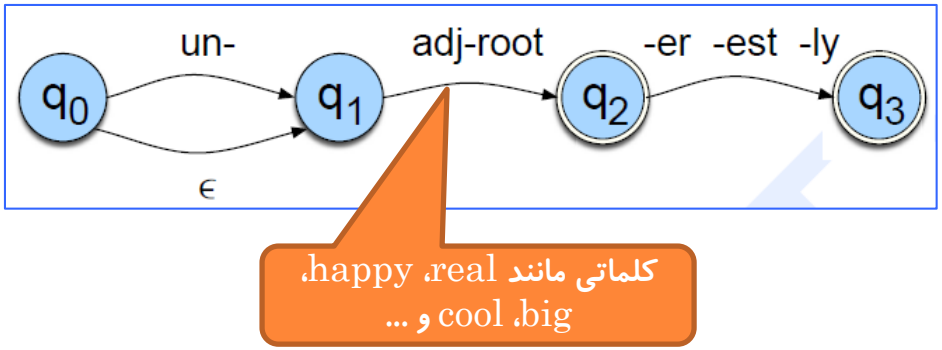


اطلاعات موجود در واژگان: افعال با قاعده، افعال بی قاعده، حالت گذشته افعال بی قاعده

reg-verb-stem	irreg-verb-stem	irreg-past-verb	past	past-part	pres-part	3sg
walk fry talk impeach	cut speak sing	caught ate eaten sang	-ed	-ed	-ing	-s

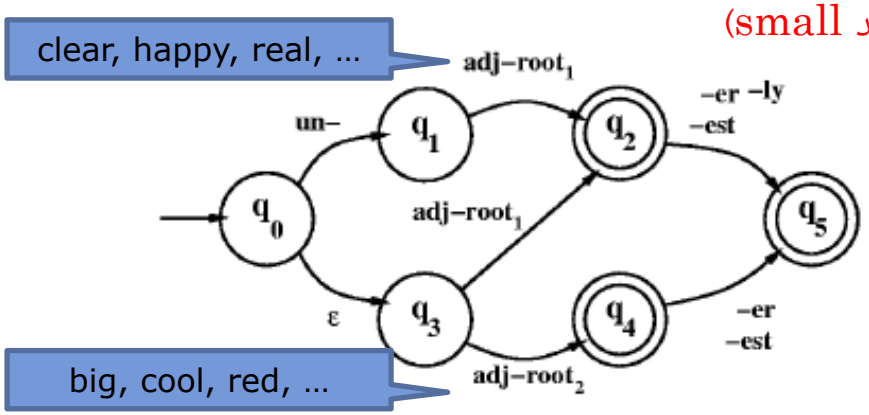
اتوماتای حالت متناهی در ساخت‌واژه . . .

- یک اتوماتا برای پذیرش صفات انگلیسی . . .



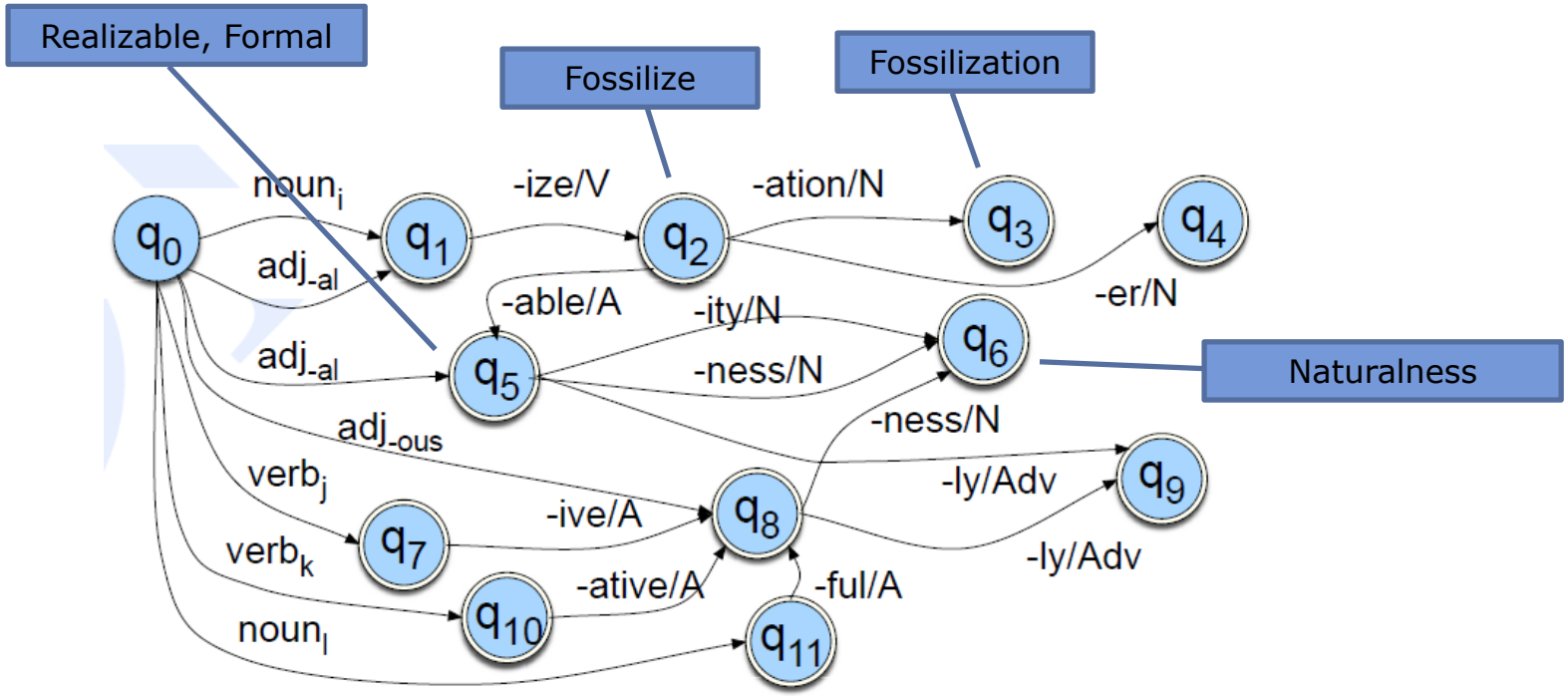
big, bigger, biggest,
happy, happier, happiest, happily
unhappy, unhappier, unhappiest, unhappily
clear, clearer, clearest, clearly, unclear, unclearly
cool, cooler, coolest, coolly
red, redder, reddest
real, unreal, really

- تولید کلمات نادرستی مانند unbig و smally
  - نحوه حل مشکل؟
    - جدا کردن ریشه‌ها به دو دسته ۱- آنهایی که un- و ly– می‌گیرند (مانند real و happy) و ۲- آنهایی که قادر به گرفتن این دو نیستند (مانند small)



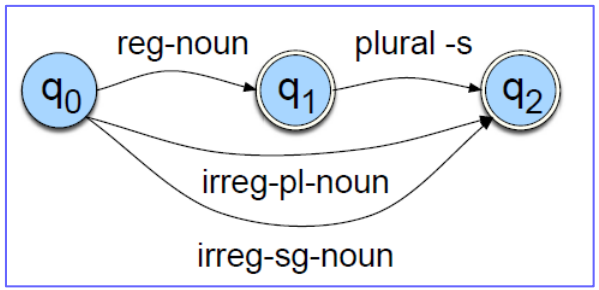
اتوماتای حالت متناهی در ساخت‌واژه . . .

- یک اتوماتا برای پذیرش ساخت‌واژه اشتقاقی انگلیسی

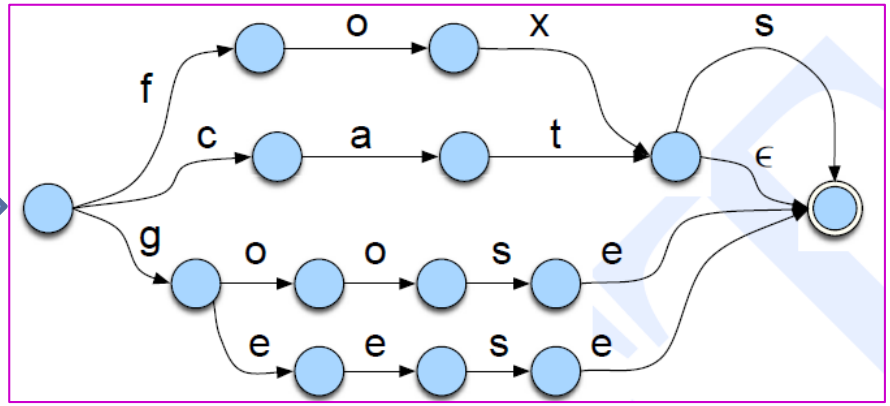


- ساخت‌واژه اشتقاقی دارای اتوماتای پیچیده‌تری نسبت به ساخت‌واژه تصریفی است

اتوماتای حالت متناهی در ساخت‌واژه



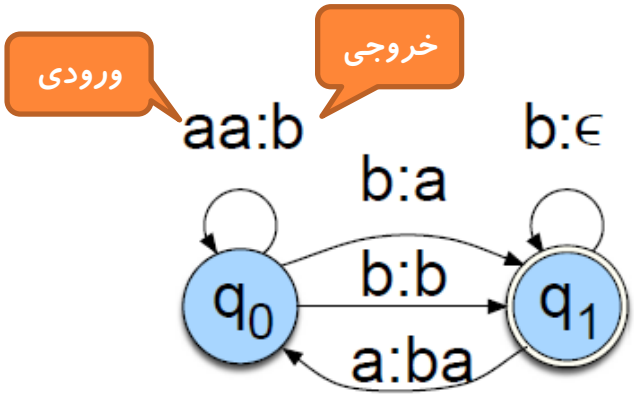
- نحوه استفاده در بررسی کلمات متن
  - ایجاد اتوماتاهای morphotactic
  - توسعه مسیرها (کمان‌ها) به هر نویسه روی هر کمان
    - جهت ساخت کلمات
  - بررسی هر کلمه به صورت نویسه به نویسه



reg-noun	irreg-pl-noun	irreg-sg-noun	plural
fox	geese	goose	-s
cat	sheep	sheep	

مبدل حالت متناهی (FST) . . .

- **مبدل حالت متناهی (FST: Finite State Transducers)**
  - **نوعی از اتوماتای حالت محدود که قادر است یک رشته را به رشته دیگری نگاشت کند**
    - بر روی هر یال یک جفت «ورودی:خروجی» قرار می‌گیرد ← تعریف رابطه بین ورودی و خروجی

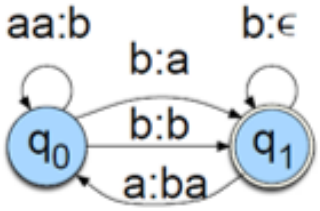


- **انواع کاربردها**
  - تشخیص دهنده (Recognizer): دریافت یک جفت رشته و تشخیص اینکه به هم نگاشت می‌شوند یا نه
  - تولید کننده (Generator): تولید یک جفت رشته به عنوان خروجی
  - ترجمه کننده (Translator): دریافت یک رشته ورودی و تولید یک رشته دیگر به عنوان خروجی
  - نقل دهنده (Relater): محاسبه رابطه بین دو رشته

مبدل حالت متناهی (FST) . . .

تعریف

- Q: مجموعه متناهی حالت‌ها با تعداد N: $q_0, q_1, \ldots, q_{N-1}$
- Σ: مجموعه‌ای متناهی حاوی کاراکترهای ورودی
- Δ: مجموعه‌ای متناهی حاوی کاراکترهای خروجی
- q_0: حالت اولیه (شروع)
- F: مجموعه حالات پایانی، $F \subseteq Q$



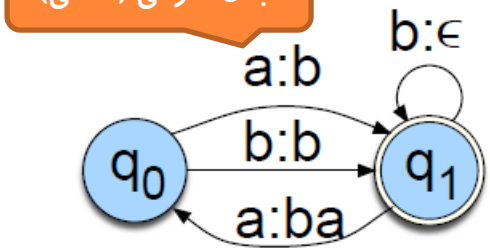
- δ(q,w): تابع انتقال حالت- انتقال از حالت $q \in Q$ با آمدن ورودی $w \in \Sigma$ و تولید حالت بعدی $q' \in Q$ (ممکن است خروجی بیش از یک حالت باشد: حالات ابهام)

- σ(q,w): تابع انتقال تولید خروجی- تولید خروجی $o \in \Delta$ از حالت $q \in Q$ و با آمدن ورودی $w \in \Sigma$

مبدل حالت متناهی (FST) . . .

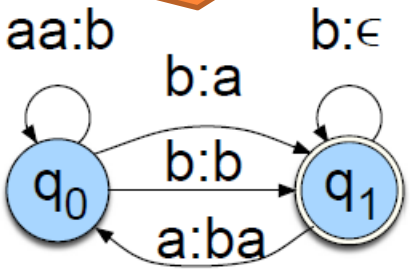
- **مبدل متوالی (Sequential Transducer): قطعی**
  - مبدل قطعی که به ازای هر ورودی در هر حالت فقط یک خروجی مشخص تولید کند
  - می‌توانند خروجی ε داشته باشند اما نمی‌توانند ورودی ε داشته باشند

مبدل متوالی (قطعی)

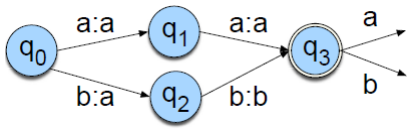


معکوس مبدل = غیرمتوالی (غیرقطعی)

مبدل غیرمتوالی (غیرقطعی)



- معکوس (inversion) مبدل: نگاشت کردن خروجی به ورودی
- مبدل Subsequential: تولید یک رشته اضافی در حالت پایانی و اتصال آن به خروجی
  - تولید بیشتر از یک رشته: p-subsequential (تولید p رشته)



2-subsequential

مبدل حالت متناهی (FST) . . .

خروجی تجزیه‌گر

ساخت تجزیه‌گر ساخت‌واژی . . .

- ورودی–خروجی‌ها
  - ورودی: cats
  - خروجی: cat+N+PL
    - اسم، جمع

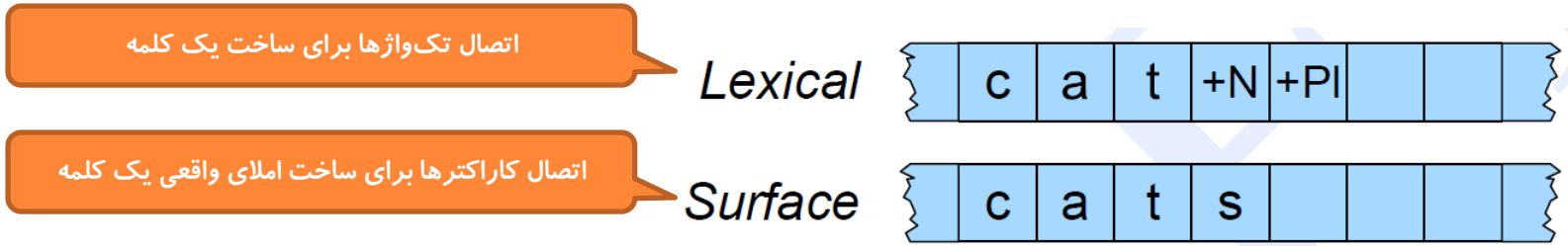
English		Spanish		
Input	Morphologically Parsed Output	Input	Morphologically Parsed Output	Gloss
cats	cat +N +PL	pavos	pavo +N +Masc +Pl	'ducks'
cat	cat +N +SG	pavo	pavo +N +Masc +Sg	'duck'
cities	city +N +Pl	bebo	beber +V +PInd +1P +Sg	'I drink'
geese	goose +N +Pl	canto	cantar +V +PInd +1P +Sg	'I sing'
goose	goose +N +Sg	canto	canto +N +Masc +Sg	'song'
goose	goose +V	puse	poner +V +Perf +1P +Sg	'I was able'
gooses	goose +V +1P +Sg	vino	venir +V +Perf +3P +Sg	'he/she came'
merging	merge +V +PresPart	vino	vino +N +Masc +Sg	'wine'
caught	catch +V +PastPart	lugar	lugar +N +Masc +Sg	'place'
caught	catch +V +Past			

تجزیه‌گر دانشگاه شهید بهشتی

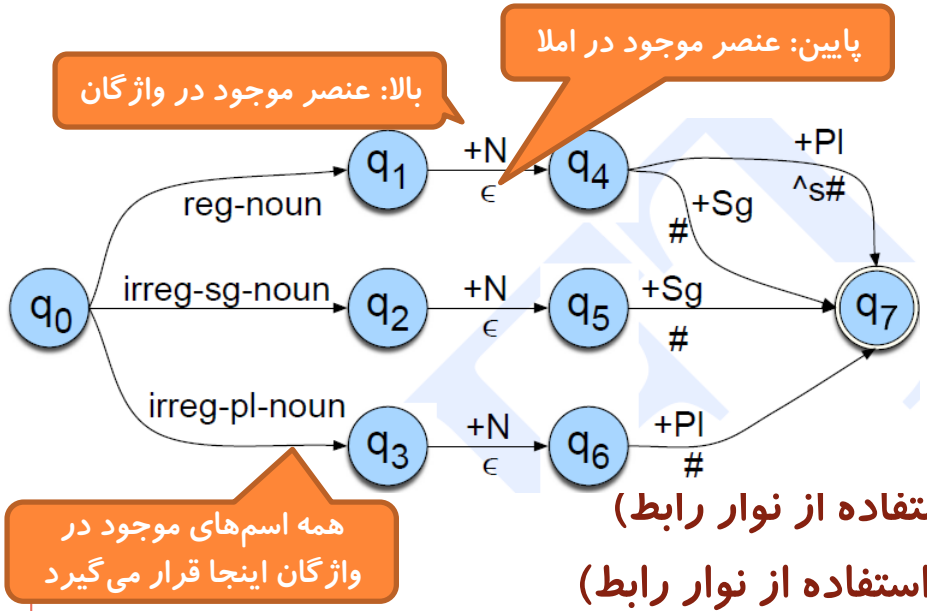
http://step1.nlplab.sbu.ac.ir/stemmer/Stemmer.aspx

مبدل حالت متناهی (FST): ساخت تجزیه‌گر . . .

نوارها (Tapes)



کاراکترهای خاص



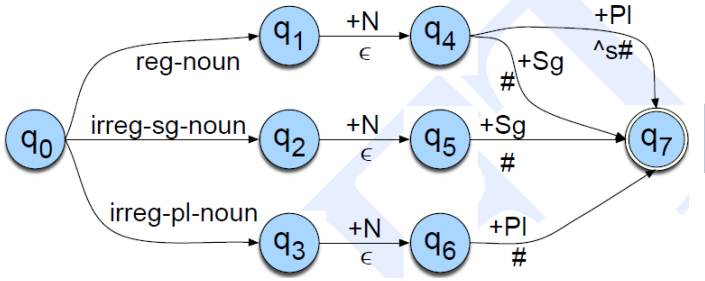
- +N: اسم
- +Sg و +Pl: مفرد و جمع بودن
- علامت ε: هیچ چیز!
- علامت #: مرز کلمه (در عمل نداریم: استفاده از نوار رابط)
- علامت ^: مرز تک‌واژ (در عمل نداریم: استفاده از نوار رابط)

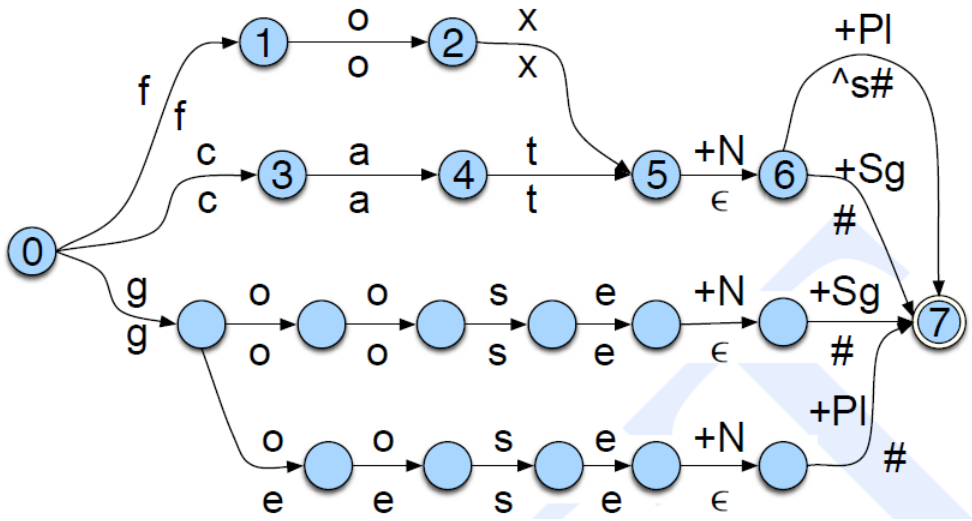
مبدل حالت متناهی (FST): ساخت تجزیه‌گر . . .

- تغییر واژگان
  - حاوی نگاشت‌ها

نگاشت املای geese به goose توسط تجزیه‌گر
o:e معادل نگاشت FST بین ورودی (پایین کمان) و خروجی (بالای کمان) است

reg-noun	irreg-pl-noun	irreg-sg-noun
fox	g o:e o:e s e	goose
cat	sheep	sheep
aardvark	m o:i u:ϵ s:c e	mouse





cats → cat+N+Pl
c:c a:a t:t +N:ε +Pl:^s#

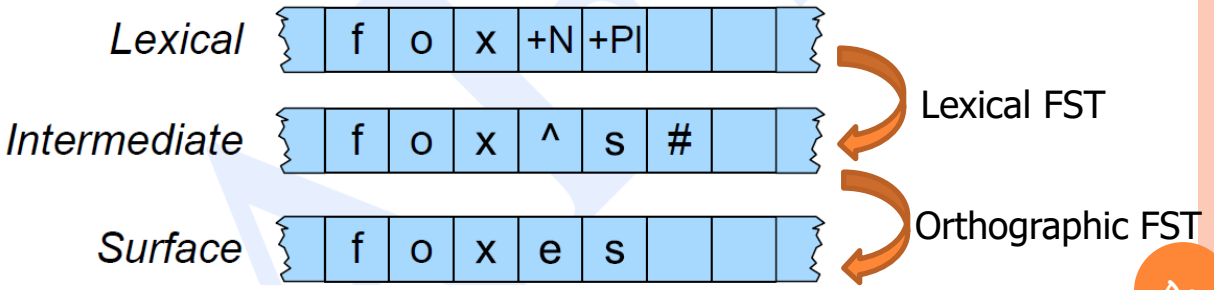
مبدل حالت متناهی (FST): ساخت تجزیه‌گر . . .

- نیازمندی‌های تجزیه ساخت‌واژی
  - Lexicon: مجموعه‌ای از ریشه‌های کلمات و وندها به همراه اطلاعاتی در مورد آنها
    - کتاب (اسم)
  - Morphotactics: مجموعه‌ای از قوانین مورفولوژی که نوع وندهای قابل اتصال به هر ریشه و ترتیب اتصال آنها را بیان می‌کند
    - شکل جمع اسم‌ها: ریشه (اسم) + (ها | ان)
  - Orthographic rules: مجموعه‌ای از قوانین املایی که نوع تغییر در املای کلمه را هنگام اتصال وندها بیان می‌کند
    - در انگلیسی: y ↤ ie (city ↤ cities)
    - در فارسی: ه ↤ گ (همسایه ↤ همسایگان)

- یادآوری
  - نیاز به Orthographic Rules

Name	Description of Rule	Example
Consonant doubling	1-letter consonant doubled before *-ing/-ed*	beg/begging
E deletion	Silent e dropped before *-ing* and *-ed*	make/making
E insertion	e added after *-s,-z,-x,-ch, -sh* before *-s*	watch/watches
Y replacement	*-y* changes to *-ie* before *-s*, *-i* before *-ed*	try/tries
K insertion	verbs ending with *vowel* + *-c* add *-k*	panic/panicked

- حل با FST



مبدل حالت متناهی (FST): ساخت تجزیه‌گر . . .

مبدل برای Orthographic Rules

- برای هر کدام از قانون‌ها یک FST ساخته می‌شود

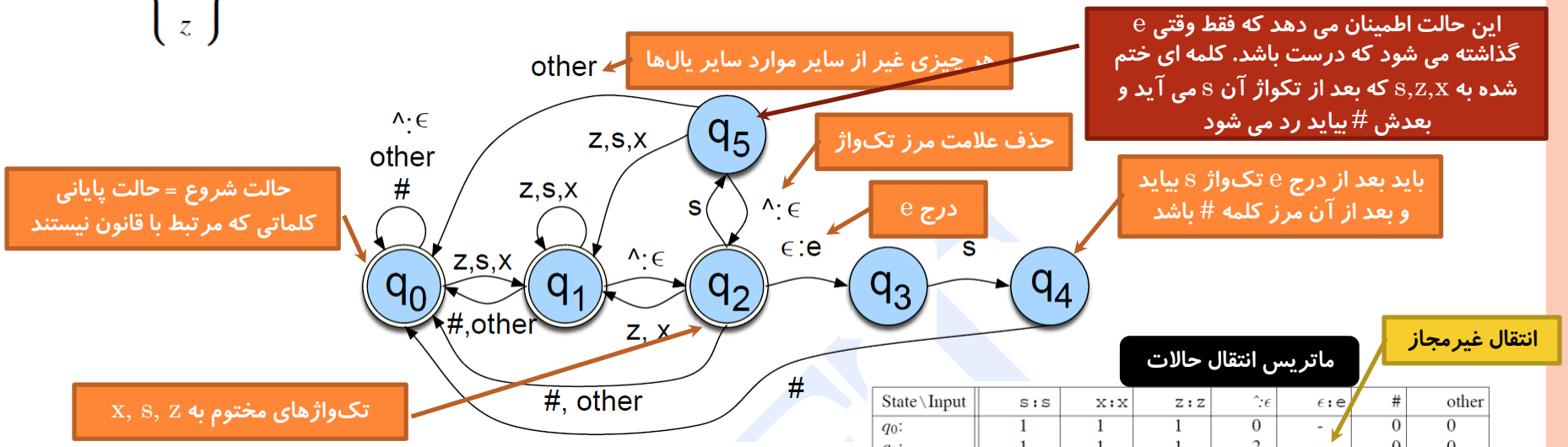
به جای ε بنویسید e
اگر ε بین {x,s,z} و s بیاید

E insertion	e added after -*s*,-*z*,-*x*,-*ch*, -*sh* before -*s*	watch/watches

$$\epsilon \rightarrow e \,/\, \left\{ \begin{matrix} x \\ s \\ z \end{matrix} \right\} \hat{}\_\_\ s\#$$

مبدل برای درج e

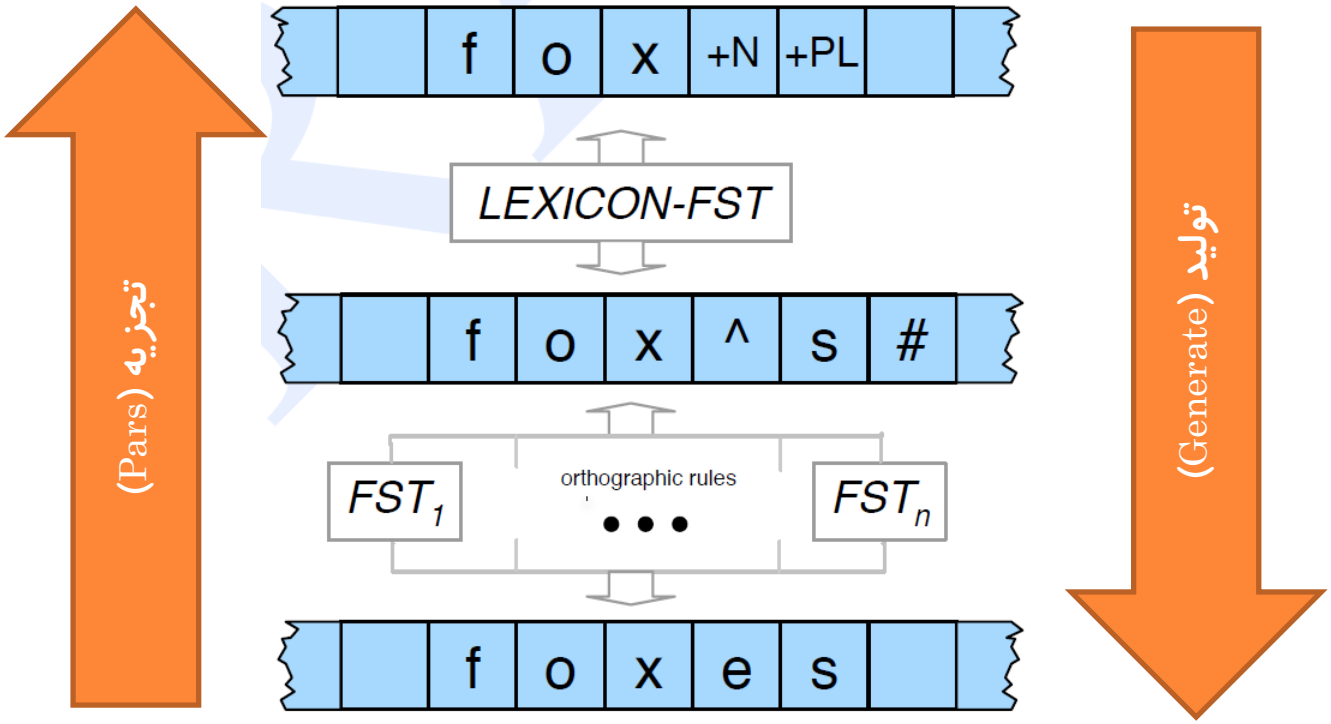
- قانون: گذاشتن e بعد از تک‌واژ ختم شده به x، s و z و تک‌واژ s



State\Input	s:s	x:x	z:z	^:ϵ	ϵ:e	#	other
q_0:	1	1	1	0	-	0	0
q_1:	1	1	1	2	-	0	0
q_2:	5	1	1	0	3	0	0
q_3	4	-	-	-	-	-	-
q_4	-	-	-	-	-	0	-
q_5	1	1	1	2	-	-	0

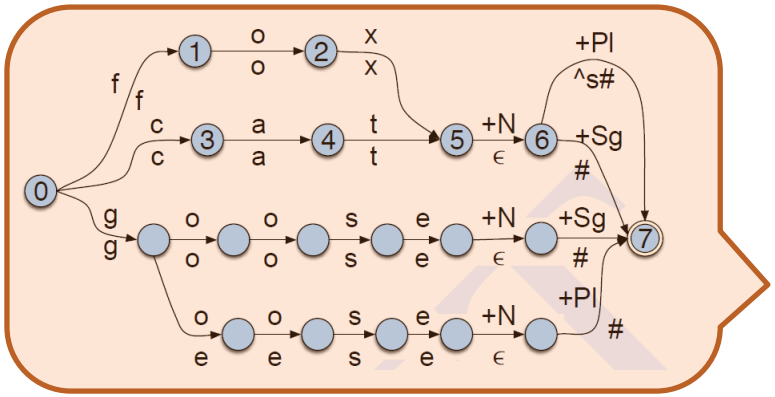
مبدل حالت متناهی (FST): ساخت تجزیه‌گر . . .

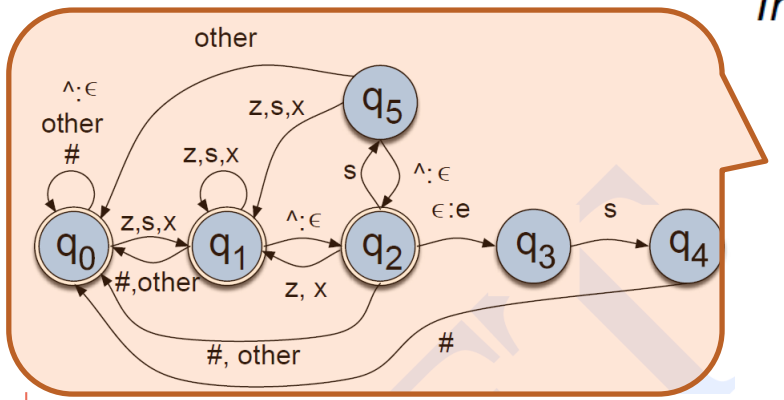
- ترکیب FSTهای واژگان و قوانین

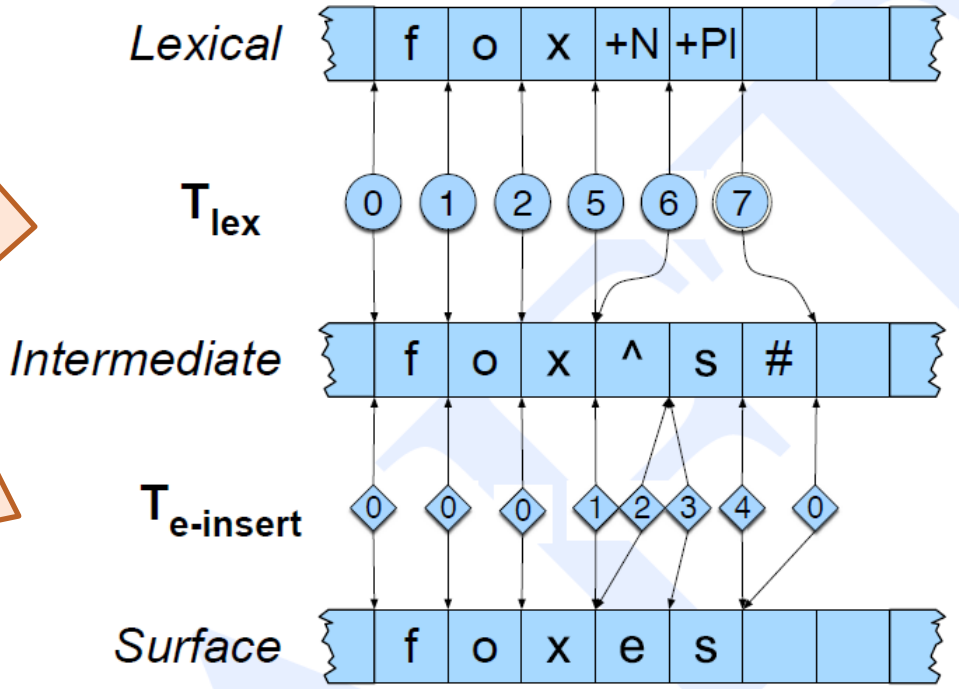


مبدل حالت متناهی (FST): ساخت تجزیه‌گر

- **مثال: پذیرش/تولید**
  - پذیرش نگاشت fox +N +PL به foxes



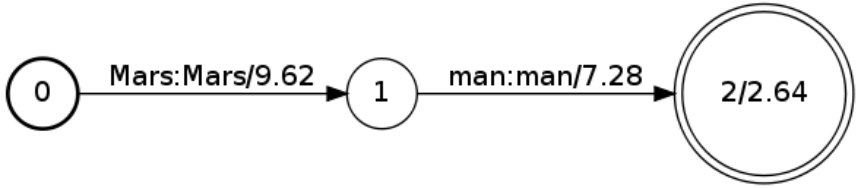




مبدل حالت متناهی (FST): سایر کاربردها

- واحدسازی (Tokenization)
  - کلمه
  - جمله

- درست کردن بزرگ و کوچک کردن حروف (Case Restoration)



ریشه‌یابی بدون واژگان (Porter Stemmer) . . .

- تجزیه‌گر با مبدل حالت متناهی (FST)
  - استفاده از واژگان + مجموعه قوانین (ساخت‌واژی و املایی)
  - استخراج ریشه کلمات = ریشه‌یابی
- ریشه‌یابی بدون واژگان و فقط با مجموعه‌ای از قوانین
  - کاربرد در بازیابی اطلاعات (مانند موتورهای جستجو)
    - حذف وندهای اضافی از مستندات و پرسش
  - استفاده از مجموعه‌ای از قوانین
    - ATIONAL→ ATE (e.g., relational→ relate)
    - ING→ ε if stem contains vowel (e.g., motoring→ motor)
  - همیشه درست عمل نمی‌کند
    - روش پورتر (Porter)

Errors of Commission		Errors of Omission	
ورودی	پاسخ پورتر	ورودی	پاسخ پورتر
generalization	gener	Europe	Europ
numerical	numer	analysis	analysi
policy	polici	noise	nois

ریشه‌یابی بدون واژگان (Porter Stemmer)

- ریشه‌یاب پورتر (Porter)
  - ارائه شده در ۱۹۸۰ توسط Martin Porter
- دسترسی: http://tartarus.org/~martin/PorterStemmer/index.html
  - قوانین و کد با زبان‌های مختلف

Step 1a

sses	→	ss	caresses	→	caress
ies	→	i	ponies	→	poni
ss	→	ss	caress	→	caress
s	→	ø	cats	→	cat

Step 1b

(*v*)ing	→	ø	walking	→	walk
			sing	→	sing
(*v*)ed	→	ø	plastered	→	plaster
…					

Step 2 (for long stems)

ational	→	ate	relational	→	relate
izer	→	ize	digitizer	→	digitize
ator	→	ate	operator	→	operate
…					

Step 3 (for longer stems)

al	→	ø	revival	→	reviv
able	→	ø	adjustable	→	adjust
ate	→	ø	activate	→	activ
…					

- نسخه برخط: http://9ol.es/porter_js_demo.html

خطاياب املايی . . .

- خطای املایی خیلی رایج است
  - در متون ویرایش شده: ۰.۰۵٪
  - در متون تایپ شده در شرایط مشکل (مانند دریافت اطلاعات تلفنی): ۳۸٪
  - در جستجوهای اینترنتی: ۲۶٪

- کاربردها
  - رفع خطاهای تایپ متون توسط انسان
  - رفع اشکال خروجی سیستم‌های تشخیص خودکار
    - نویسه خوان نوری (OCR)
    - تشخیص دست نوشته

خطایاب املایی . . .

سطح بررسی خطا

- املایی «دانسگاه تهران»
- نحوی «من رفتند»
- معنایی «سلاح کار خویش را نمی‌داند»
- ساختار گفتمانی «سه نفر ماندند: حسن و حسین»
- کاربردی «این سمینار زمین‌شناسی است.»

منشا خطا

- حروفچینی «قالب ← فالب» یا «سمینار ← سمنیار»
- خطاهای شناختی «سپاسگزاری← سپاسگذاری» یا «برخاست← برخواست»
- خطاهای آوایی «اجتماعی← اشتماعی»

خطاياب املايی . . .

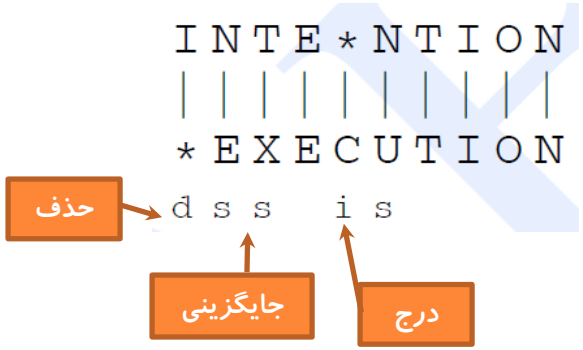
- راه حل: برای تشخیص خطا
  - استفاده از یک فرهنگ لغت حاوی شکل درست املایی کلمات
  - پیاده‌سازی تشخیص درستی کلمه در فرهنگ لغت: استفاده از مبدل حالت متناهی (FST)
    - کاهش تعداد کلمات موجود در فرهنگ لغت: جدا کردن ریشه‌ها و وندها
    - ایجاد انعطاف در افزودن کلمه جدید به فرهنگ لغت: با افزودن ریشه، صورت‌های تصریفی آن به صورت خودکار پوشش داده می‌شود
- برای تصحیح خطا
  - در صورت تشخیص خطا، چگونه می‌توان آن را اصلاح کرد
    - اصلاح کردن کلمه graffe: giraffe, graf, craft, grail
  - جایگزین کردن کلمه نادرست با کلمه درست نزدیک به آن
    - محاسبه فاصله بین کلمه نادرست و کلمات درست
    - یکی از روش‌ها: کمینه فاصله ویرایش (minimum edit distance)
  - نحوه اعمال
    - به صورت خودکار یا به صورت تعاملی با ارائه لیست پیشنهاد

خطاياب املايی . . .

- چهار امکان اصلی و رایج خطای املایی
  - حذف «آد» به‌جای «آدم».
  - درج «آددم» به‌جای «آدم».
  - جایگزینی «آذم» به‌جای «آدم».
  - جابه‌جایی «آمد» به‌جای «آدم».

خطاياب املايی- Minimum Edit Distance . . .

- هدف: یافتن رشته‌هایی با کمترین فاصله با یک رشته دیگر
  - کمترین فاصله = کمترین تعداد اپراتورهای ویرایشی بین دو رشته (درج، حذف یا جایگزینی) برای تبدیل یک رشته به دیگری
  - تزاز (alignment) بین دو رشته

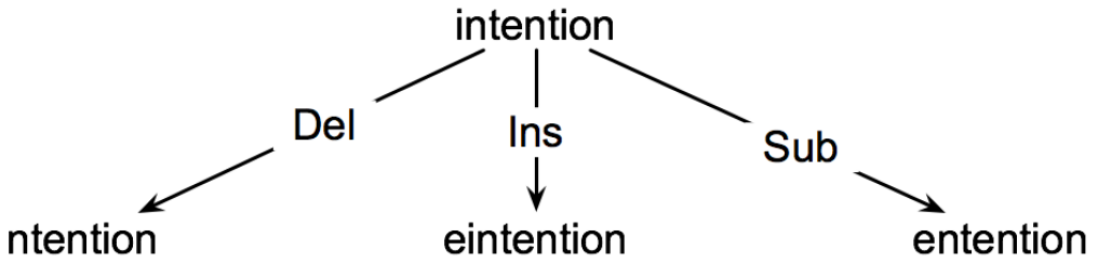


  - روش Levenshtein
    - هر کدام از سه خطا دارای وزن برابر یک هستند: در مثال فوق = ۵
    - نوع دیگر: عدم در نظر گرفتن خطای جایگزینی (هر جایگزینی = یک درج و یک حذف = وزن ۲): در مثال فوق = ۸
  - پیاده‌سازی با روش برنامه‌نویسی پویا (dynamic programming)

خطایاب املایی- Minimum Edit Distance . . .

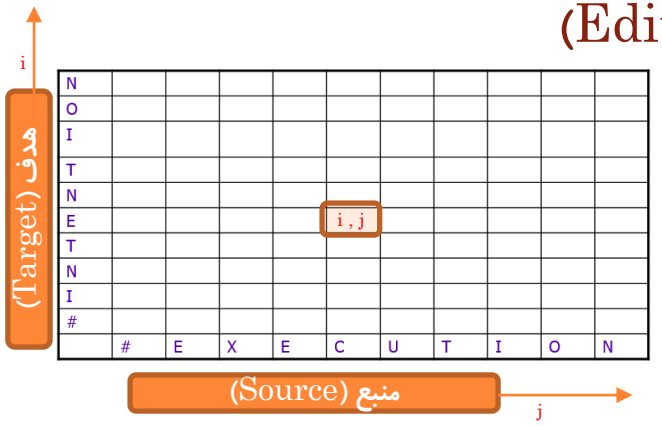
- مثال
  - تراز شدن intention و execution
    - سه نوع خطا: حذف، درج و جایگزینی

	i	n	t	e	n	t	i	o	n
delete i →	n	t	e	n	t	i	o	n	
substitute n by e →	e	t	e	n	t	i	o	n	
substitute t by x →	e	x	e	n	t	i	o	n	
insert u →	e	x	e	n	u	t	i	o	n
substitute n by c →	e	x	e	c	u	t	i	o	n



خطایاب املایی- Minimum Edit Distance . . .

- الگوریتم کمینه فاصله ویرایش (Minimum Edit Distance)
  - ایجاد یک ماتریس فاصله (Edit-Distance Matrix)
    - قرار دادن رشته هدف (Target) در ستون
      - رشته مورد بررسی (اصلی)
    - قرار دادن رشته منبع (Source) در سطر
      - رشته مورد استفاده در مقایسه

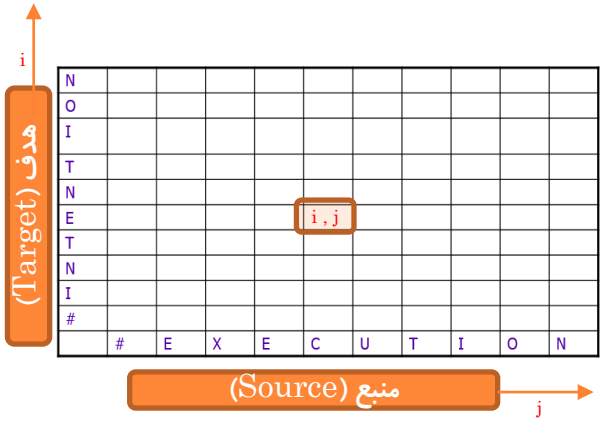


- حرکت در سطرها و ستون‌ها و پر کردن عناصر ماتریس فاصله
  - هر عنصر edit-distance[i,j] بیانگر فاصله i کارکتر اول رشته هدف با j کاراکتر رشته منبع است
  - محاسبه فاصله هر عنصر بر اساس کمینه مقدار ایجاد شده در حرکت از سه عنصر قبلی به آن

$$distance[i,j] = \min \begin{cases} distance[i-1,j] + \text{ins-cost}(target_{i-1}) \\ distance[i-1,j-1] + \text{subst-cost}(source_{j-1}, target_{i-1}) \\ distance[i,j-1] + \text{del-cost}(source_{j-1}) \end{cases}$$

خطایاب املایی- Minimum Edit Distance . . .

- الگوریتم کمینه فاصله ویرایش (Minimum Edit Distance)




```
function MIN-EDIT-DISTANCE(target, source) returns min-distance

  n ← LENGTH(target)
  m ← LENGTH(source)
  Create a distance matrix distance[n+1,m+1]
  Initialize the zeroth row and column to be the distance from the empty string
    distance[0,0] = 0
    for each column i from 1 to n do
        distance[i,0] ← distance[i-1,0] + ins-cost(target[i])
    for each row j from 1 to m do
        distance[0,j] ← distance[0,j-1] + del-cost(source[j])
  for each column i from 1 to n do
    for each row j from 1 to m do
      distance[i,j] ← MIN( distance[i−1,j] + ins-cost(target_{i−1}),
                           distance[i−1,j−1] + subst-cost(source_{j−1}, target_{i−1}),
                           distance[i,j−1] + del-cost(source_{j−1}))
  return distance[n,m]
```

فاصله بین دو رشته

خطایاب املایی- Minimum Edit Distance . . .

- الگوریتم Levenshtein
  - هزینه درج = ۱
  - هزینه حذف = ۱
  - هزینه جایگزینی = ۲ (یک درج و یک حذف)

- Initialization

$$D(i,0) = i$$

$$D(0,j) = j$$

- Recurrence Relation:

For each i = 1…N
  For each j = 1…M

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

- Termination:

$D(N,M)$ is distance

خطایاب املایی- Minimum Edit Distance . . .

- مثال . . .
  - مقداردهی اولیه

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

- Initialization

```
D(i,0) = i
D(0,j) = j
```

هزینه جایگزینی سه کارکتر با INT # (حذف سه کاراکتر)

خطایاب املایی- Minimum Edit Distance . . .

- مثال . . .
  - محاسبه جدول فاصله

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

خطایاب املایی- Minimum Edit Distance . . .

- مثال . . .
  - محاسبه جدول فاصله

N	9									
O	8									
I	7									
T	6									
N	5									
E	4	3								
T	3	4								
N	2	3								
I	1	2								
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

خطایاب املایی- Minimum Edit Distance . . .

- مثال . . .
  - محاسبه جدول فاصله

فاصله بین دو رشته

N	9	8	9	10	11	12	11	10	9	**8**
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

خطایاب املایی- Minimum Edit Distance . . .

- علاوه بر مقادیر هزینه، مسیر حرکت را نیز نیاز داریم
  - برای تراز کردن (Alignment)

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
```

- افزودن اطلاعات موردنیاز برای تراز کردن
  - نگهداری اطلاعات اینکه از کدام سلول به سلول جاری آمده‌ایم
  - وقتی به انتها رسیدیم، عقب‌گرد کنیم تا کل مسیر را بدست آوریم

خطایاب املایی- Minimum Edit Distance . . .

- افزودن اطلاعات مسیر

N	9									
O	8									
I	7									
T	6									
N	5									
E	4	3								
T	3	4								
N	2	3								
I	1	2								
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

خطایاب املایی- Minimum Edit Distance . . .

- افزودن اطلاعات مسیر
  - عقب‌گرد از آخر به اول

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
```

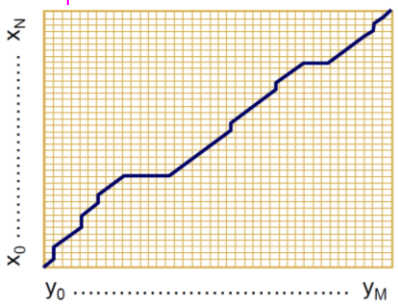
n	9	↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙←↓ 12	↓ 11	↓ 10	↓ 9	↙ **8**	
o	8	↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↓ 10	↓ 9	↙ **8**	← 9	
i	7	↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	↙ **8**	← 9	← 10	
t	6	↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙ **8**	← 9	← 10	←↓ 11	
n	5	↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ **8**	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙↓ 10	
e	4	↙ 3	← 4	↙← **5**	← **6**	← 7	←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	
t	3	↙←↓ 4	↙←↓ **5**	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙ 7	←↓ 8	↙←↓ 9	↓ 8	
n	2	↙←↓ **3**	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↓ 7	↙←↓ 8	↙ 7	
i	**1**	↙←↓ 2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙ 6	← 7	← 8	
#	**0**	1	2	3	4	5	6	7	8	9	
	#	**e**	**x**	**e**	**c**	**u**	**t**	**i**	**o**	**n**	

خطایاب املایی- Minimum Edit Distance . . .

الگوریتم Levenshtein – افزودن اطلاعات مسیر برای تراز کردن

- Base conditions:
  `D(i,0) = i` `D(0,j) = j`
  Termination:
  `D(N,M) is distance`
- Recurrence Relation:

```
For each  i = 1…M
      For each  j = 1…N
                          D(i-1,j) + 1                    deletion
D(i,j)= min               D(i,j-1) + 1                    insertion
                          D(i-1,j-1) +  2; if X(i) ≠ Y(j)   substitution
                                        0; if X(i) = Y(j)

                          LEFT    insertion
ptr(i,j)=                 DOWN    deletion
                          DIAG    substitution
```



خطایاب املایی- Minimum Edit Distance . . .

- حالت وزن‌دار
  - به جای استفاده از مقدار هزینه ۱ (برای حذف و درج) و ۲ (برای جایگزینی) از مقادیر دیگری استفاده می‌شود
  - مقادیر متناسب کاربرد و در قالب یک ماتریس که بیانگر هزینه‌های فوق است
    - برخی کارکترها بیشتر از بقیه به جای هم اشتباه می‌شوند (در تایپ) - هزینه جایگزینی کمتر

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

هرچه دو کاراکتر بیشتر با هم اشتباه شوند، هزینه جایگزینی آنها را باید کمتر در نظر بگیریم

اشتباه شدن y با s

h.veisi@ut.ac.ir)

خطایاب املایی- Minimum Edit Distance . . .

«ب» با «پ» و «ی» بیشتر اشتباه می‌شود تا با «ت» و «ث»

- ماتریس درهم ریختگی فارسی
  - بر اساس خروجی سیستم OCR
  - ماتریس هزینه عکس این ماتریس

	ا	ب	پ	ت	ث	ج	چ	ح	خ	د	ذ	ر	ز	ژ	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ک	گ	ل	م	ن	و	ه	ی	فراوانی کل
ا	0	4	13	7	3	0	0	2	0	7	0	1	3	0	1	15	1	0	1	0	1	0	3	2	0	1	47	1	9	0	5	10	23,073
ب	5	0	99	27	2	1	0	0	0	0	0	2	0	0	8	2	0	0	0	1	13	0	0	0	0	0	8	12	17	0	5	399	6,227
پ	2	51	0	4	0	0	0	5	0	1	0	2	0	0	6	0	1	0	0	0	0	0	0	1	1	0	0	7	2	0	1	67	839
ت	6	11	0	0	102	0	1	1	1	3	0	2	11	3	6	31	4	1	6	0	12	0	3	7	1	1	31	31	276	1	17	25	7,559
ث	1	0	0	37	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	1	0	0	0	1	0	11	0	0	2	210
ج	0	11	4	15	1	0	41	36	15	1	0	2	1	0	22	2	1	0	1	0	0	0	1	0	0	0	2	1	1	0	2	3	1,776
چ	0	6	1	5	0	60	0	3	0	0	0	2	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	7	2	1	8	410
ح	0	2	0	1	0	176	20	0	25	0	0	0	1	0	6	2	3	0	1	0	32	3	2	1	0	0	2	3	5	1	7	3	1,562
خ	1	1	0	11	0	12	0	101	0	0	1	0	0	0	5	2	1	0	1	0	1	1	13	0	0	0	0	0	6	0	1	4	1,813
د	14	9	0	5	0	0	0	4	0	0	1	0	0	0	2	1	1	0	0	0	2	0	1	3	0	0	0	5	6	2	83	1	10,438
ذ	0	0	0	2	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	1	0	0	0	1	1	283
ر	1	6	0	1	0	1	0	0	0	0	0	0	195	0	0	0	2	0	0	4	0	0	3	1	4	0	1	15	6	64	1	11	12,867
ز	0	0	0	2	0	0	0	0	0	1	0	24	0	5	0	0	0	0	1	0	0	0	0	1	0	0	0	4	1	0	0	0	3,321
ژ	0	0	0	0	0	0	0	0	0	0	0	9	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	124
س	0	23	0	8	0	0	2	15	0	14	0	0	0	0	0	21	47	0	0	1	83	0	1	1	0	0	4	136	40	0	29	38	4,500
ش	0	4	0	81	2	1	0	0	3	1	0	2	0	0	11	0	3	0	0	0	2	0	10	11	0	3	15	12	50	5	2	8	3,982
ص	0	22	0	11	0	0	0	11	0	6	0	0	0	0	36	0	0	9	12	0	8	0	0	2	1	0	0	49	2	0	9	3	1,042
ض	0	2	0	3	0	1	0	0	1	1	0	0	0	0	0	2	20	0	0	0	3	0	4	0	0	0	1	1	8	0	0	1	517
ط	2	3	0	1	0	1	0	1	0	0	0	1	0	0	1	2	0	0	0	37	12	0	2	4	0	0	29	2	30	3	0	3	917
ظ	1	3	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	6	0	1	0	0	10	0	0	2	1	0	0	0	0	389
ع	1	28	0	9	0	4	0	55	1	1	0	0	0	0	14	0	3	0	0	0	0	93	12	2	0	0	3	24	17	1	3	14	1,915
غ	0	2	0	2	0	0	0	0	4	1	1	0	0	0	3	0	1	0	0	0	32	0	0	4	3	0	0	2	0	1	0	0	225
ف	9	6	0	6	0	1	0	2	0	10	0	1	0	0	6	20	2	3	1	0	2	6	0	89	3	0	11	1	8	0	7	2	2,071
ق	8	0	0	4	0	0	0	0	21	0	0	0	0	0	1	0	12	0	1	0	4	4	87	0	0	0	24	2	1	0	0	9	2,001
ک	0	2	0	33	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	117	3	1	56	0	1	5	4,062
گ	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	25	0	3	0	0	0	0	0	2,107
ل	40	3	0	37	0	0	0	0	0	2	3	0	0	0	0	22	0	0	1	4	8	0	21	11	7	0	0	3	44	0	2	3	3,600
م	1	34	0	51	0	6	0	20	15	3	0	15	0	0	139	4	12	0	1	0	32	1	3	0	0	0	10	0	61	1	77	45	8,018
ن	3	33	6	732	4	0	0	9	4	9	0	1	8	0	17	16	7	0	0	0	27	1	4	9	1	1	37	13	0	0	10	52	10,775
و	0	3	1	0	0	0	0	0	0	0	0	83	0	0	2	0	0	0	0	0	0	0	1	1	0	0	2	7	5	0	0	1	8,704
ه	2	2	0	5	0	0	0	1	1	58	0	2	0	0	2	1	1	0	4	0	5	0	0	0	1	0	1	16	20	1	0	13	8,937
ی	2	375	171	30	1	7	45	1	1	2	0	23	0	0	20	5	2	0	1	0	8	0	1	2	4	0	3	16	41	8	9	0	14,043

«ز» با «ر» بیشتر اشتباه می‌شود تا با «ن»

هزینه جایگزینی «ز» و «ر» کمتر از هزینه جایگزینی «ز» و «ن» است

خطاياب املايی- Minimum Edit Distance . . .

$$distance[i,j] \leftarrow \text{MIN}(\ distance[i-1,j] + ins\text{-}cost(target_{i-1}),$$
$$distance[i-1,j-1] + subst\text{-}cost(source_{j-1}, target_{i-1}),$$
$$distance[i,j-1] + del\text{-}cost(source_{j-1}))$$

مثال: محاسبه هزینه وزندار

- هزینه جایگزینی «ز» با «ر» برابر با ۰.۵ است (بیشتر با هم جایگزین می‌شوند)
  - دازد و دارد؟
- هزینه جایگزینی «ز» با «ن» برابر با ۲ است (کمتر با هم جایگزین می‌شوند)
  - دازد و داند؟
- سایر هزینه‌ها مانند لونتشتاین محاسبه شده است

0.5	1.5	2	3	4	د
1.5	0.5	1	2	3	ر
2	1	0	1	2	ا
3	2	1	0	1	د
4	3	2	1	0	#
د	ز	ا	د	#	

i: هدف (Target)

j: منبع (Source)

2	1.5	2	3	4	د
1.5	2	1	2	3	ن
2	1	0	1	2	ا
3	2	1	0	1	د
4	3	2	1	0	#
د	ز	ا	د	#	

i: هدف (Target)

j: منبع (Source)

خطایاب املایی- Minimum Edit Distance

- کاربردها . . .
  - یافتن پیشنهادهای مناسب برای اصلاح خطای املایی
  - تراز کردن دو رشته DNA

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

خطایاب املایی- Minimum Edit Distance

کاربردها

- تراز (Align) کردن دو متن در سیستم تشخیص گفتار
  - بین پاسخ سیستم و متن اصلی برای محاسبه خطای سیستم

جملات بازشناسی شده	مرتبه جواب
یک قوطی میخ پریروز خریدم	جواب درست
یک قوطی میخ پریروز خریدم	1
یک قوطی میخ پریروز خرید هم	2
یک قوطی میخ پریروز خرید در	3
یک قوطی میخ پریروز خرید	4
یک قوطی میخ تایید روز خریدم	5
یک قوطی میخ تایید روز خرید هم	6
یک قوطی میخ تایید روز خرید در	7

- تراز (Align) کردن دو متن در سیستم ترجمه ماشینی
  - تراز کردن جملات دو زبان در پیکره‌های موازی یا پاسخ سیستم و پاسخ انسانی

```
R Spokesman confirms    senior government adviser was shot
H Spokesman said    the senior              adviser was shot dead
              S       I                 D                   I
```

منابع مرتبط در زبان فارسی

- Karine Megerdoomian, "Persian Computational Morphology: A Unification-based Approach", NMSU, CRL, Memoranda in Computer and Cognitive Science, MCCS-00-320, 2000.

- Karine Megerdoomian, "Finite-state morphological analysis of Persian“, In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Coling 2004, University of Geneva, 2004.

- محرم اسلامی و همکاران، ”واژگان زایای زبان فارسی“، مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، تهران، ۱۳۸۳.

- سیستم ساخت‌واژی آزمایشگاه پردازش زبان دانشگاه شهید بهشتی
  - http://step1.nlplab.sbu.ac.ir/stemmer/Stemmer.aspx