Processamento lexical, morfologia e morfossintaxe

SCC5908 Introdução ao Processamento de Língua Natural

SCC0633 Processamento de Linguagem Natural

Prof. Thiago A. S. Pardo

Preâmbulo

- Em processamento de texto, é comum
  - Substituir uma palavra por outra
  - Procurar por uma informação, como data, nome, etc.
  - Analisar determinadas palavras
  - Mais genericamente, procurar por padrões no texto
    - Padrões simples: palavras
    - Padrões mais complexos: expressões, segmentos maiores

Processamento textual

- Útil para
  - Tarefas particulares: buscar algo que leu, corrigir automaticamente um texto, etc.
  - Tarefas científicas: associar sintomas e tratamentos de uma doença, identificar opiniões sobre produtos em análise de sentimentos, ciência de dados, etc.
  - Tarefas comerciais: sistemas on-line de comparação de preços, levantamento de dados de concorrentes, etc.

Apple iPad 3G 64 GB - Ta...

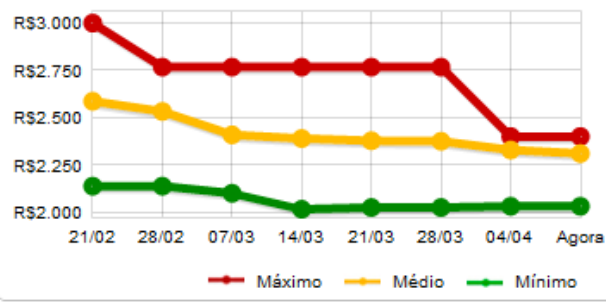
www.bondfaro.com.br/preco--tablet--apple-ipad-3g-64-gb.html

Onde comprar (10 lojas)

Popularidade | Menor Preço | Loja

Saraiva .com.br	R$ 2.399,00	12x de R$ 199,92	IR À LOJA
siciliano .com.br	R$ 2.159,10		IR À LOJA
apetrexo .com	R$ 2.031,42	12x de R$ 199,16	IR À LOJA
fnac	R$ 2.399,00	12x de R$ 199,92	IR À LOJA
Carrefour .com.br	R$ 2.399,00	12x de R$ 199,92	IR À LOJA
magazineluiza	R$ 2.159,10	12x de R$ 199,92	IR À LOJA

Histórico de preços



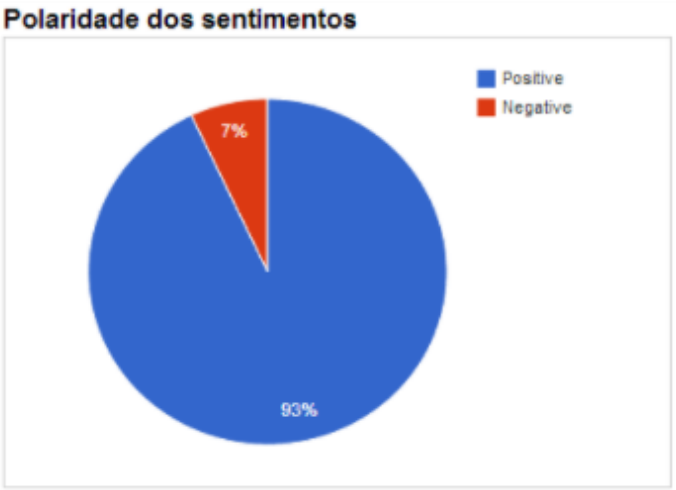
Alerta de preços

Me avisar por e-mail:

- (•) Toda vez que o menor preço do site mudar
- ( ) Quando o produto atingir um preço abaixo de R$ 2.031,42

BuscaOpiniões

(Balage Filho et al., 2014)



Opiniões Positivas:

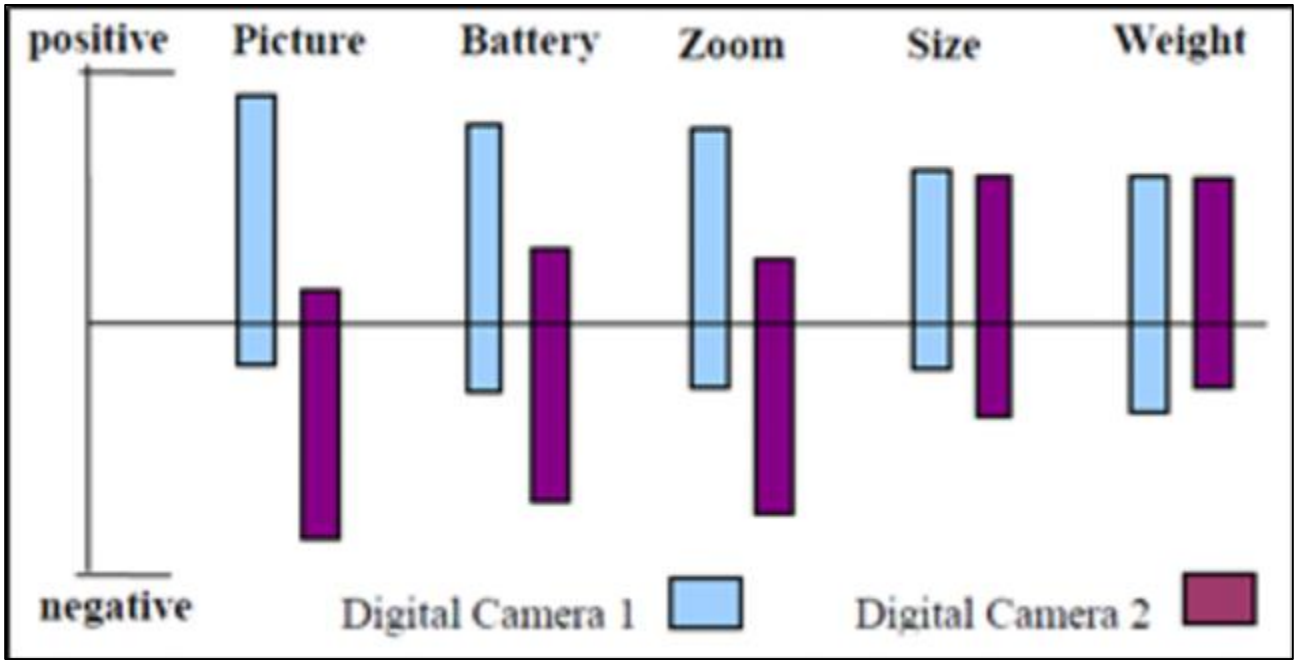
iPad mini: leve (312 gramas) e **perfeito** para segurar sem se cansar.
O formato do **iPad** mini **excelente** para ler livros, documentos e textos na web.
Menor e mais leve do que seus irmos mais velhos, ele **perfeito** para ler livros e textos em geral.
Tablet HP 8 faz **bonito** tanto no design, quanto no desempenho
Sendo pequeno e leve ele **perfeito** para ler livros e textos em geral.
E-bit **Excelente**
A cmera frontal est localizada no centro da borda **superior.**
- **bom** desempenho;
A cmera est posicionada no canto **superior** esquerdo.
positivo

Opiniões Negativas:

A tela oferece espao para posicionar 20 aplicativos e mais 6 fixos na linha **inferior.**

Opinion Observer

(Liu et al., 2005)



Exemplo

- Busca por todos os valores monetários em um texto

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

Como fariam?

Expressões Regulares (ER)

- Notação tradicional para caracterizar segmentos textuais de todo tipo
  - Especificam sequências de símbolos a serem buscados/caracterizados
  - Vários sistemas de busca de expressões regulares
    - grep, no Linux/UNIX
    - Lex/flex
    - Mecanismos próprios das linguagens de programação
      - Há variações de sistema para sistema, mas são muito parecidas

ER: notação

- Exemplos
  - Casamento direto: **preço**
  - Letra maiúscula ou minúscula no início: **[Pp]reço**
    - [ ] indicam disjunção, ou seja, um único elemento do conjunto
  - Identificação de um único dígito do texto: **[0123456789]**
  - Identificação de uma letra em um intervalo de letras: **[a-z]**
  - Qualquer caractere diferente de a: **^a**

ER: notação

- Exemplos
  - Singular ou plural: **preços?**
  - 1 ou mais ocorrências (+) de algum elemento: **Aa+i+!**
    - Aai!, Aaaaaaiiiii!
  - 0 ou maiz ocorrências (*) de algum elemento: **Aa*i*!**
    - Aaaaiii!, Aaaiiiii!, Ai!, Aaaa!
  - Caractere curinga (.): **beg.n**
    - begin, began, begun
  - Alternativa (|): **preço|os** ou **(gato)|(cão)**
    - O que acontece se tivermos **gato|cão** sem parênteses?

Exercícios

- Como identificar nomes próprios?

- E e-mails?

Exercícios

- Como identificar nomes próprios?
  - [A-Z][a-z]+
- E e-mails?
  - [a-z0-9_]+@[a-z\.]+
    - Cuidado: alguns caracteres são especiais e, para serem usados em seu sentido original, precisam de \ ou “”
      - Exemplos: . $ -

Exercício

- Expressão regular para reconhecer os valores monetários?

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

Exercício

- Expressão regular para reconhecer os valores monetários?

Levantamento da consultoria Economática aponta que empresas brasileiras de capital aberto tiveram os maiores lucros em 2010, considerando somente o setor de construção civil das Américas.

No topo da lista está a PDG Realty, com um lucro líquido de US$ 473,9 milhões apurados em 2010, bem acima da segunda colocada, a americana Walter Industries, com ganhos de US$ 385,8 milhões para o mesmo período.

As companhias brasileiras ocupam as próximas cinco posições (da 3ª à 7ª) no ranking preparado pela Economática, a saber: MRV, Cyrela, Gafisa, Brookfield e Rossi Residencial. Numa lista de dez posições, sete são ocupados por companhias nacionais.

A consultoria também preparou uma lista baseada em volume de vendas, desta vez com presença mais expressiva das construtoras americanas, a exemplo do primeiro lugar -- a Pulte Homes-- com um volume de US$ 4,44 bilhões em imóveis comercializados, e do segundo lugar --a Horton-- com uma cifra de US$ 3,96 bilhões.

...

US\$ [0-9]+,[0-9]+ [mb]ilhões

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão regular para os substantivos e os verbos?

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão regular para os substantivos e os verbos?

[A-Za-z][a-z]*_N|V

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão para substantivos seguidos de verbos?

Exercício

Sentença:

O *homem* viu a mulher de binóculos na montanha.

Análise automática:

O_DET homem_N viu_V a_DET mulher_N de_PRP binóculos_N em_PRP a_DET montanha_N .

Expressão para substantivos seguidos de verbos?

[A-Za-z][a-z]*_N [a-z]+_V

Exercício prático

- Passo a passo
  - Baixar, do projeto Gutenberg, o livro *Dom Casmurro*, de Machado de Assis (https://www.gutenberg.org/)
  - Se no Windows, instalar Cygwin e executar (https://cygwin.com/install.html)
    - Copiar o arquivo do Dom Casmurro para a pasta de usuário dentro de Cygwin (deve ser algo como C:\cygwin\home\seu_usuario)
  - Se no Linux, abrir linha de comando
    - Copiar o arquivo do Dom Casmurro para uma pasta de preferência

Exercício prático

- Passo 1: testar o grep

  grep –E –n “[Gg]utenberg” DomCasmurro.txt

- Passo 2: interpretar esse comando

  grep --help
  - Ver também as opções –c, –o e –w

Exercício prático

- Passo 3: buscar todas as ocorrências de palavras terminadas em “mento”, “agem” e “ção” e determinar qual ocorre mais?
  - grep –E –w –c “[A-Za-z]*mento” DomCasmurro.txt
  - grep –E –w –c “[A-Za-z]*agem” DomCasmurro.txt
  - grep –E –w –c “[A-Za-z]*ção” DomCasmurro.txt

- Passo 4: buscar todos os anos citados no texto
  - grep –E –w “[0-9][0-9]+” DomCasmurro.txt
    - Quais os problemas dessa expressão?

Exercício prático

- Baixar córpus MAC-MORPHO (http://nilc.icmc.usp.br/macmorpho/)
- Descompactar e acessar arquivo de teste
- Checar conteúdo e etiquetas usadas (http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf)

Exercício prático

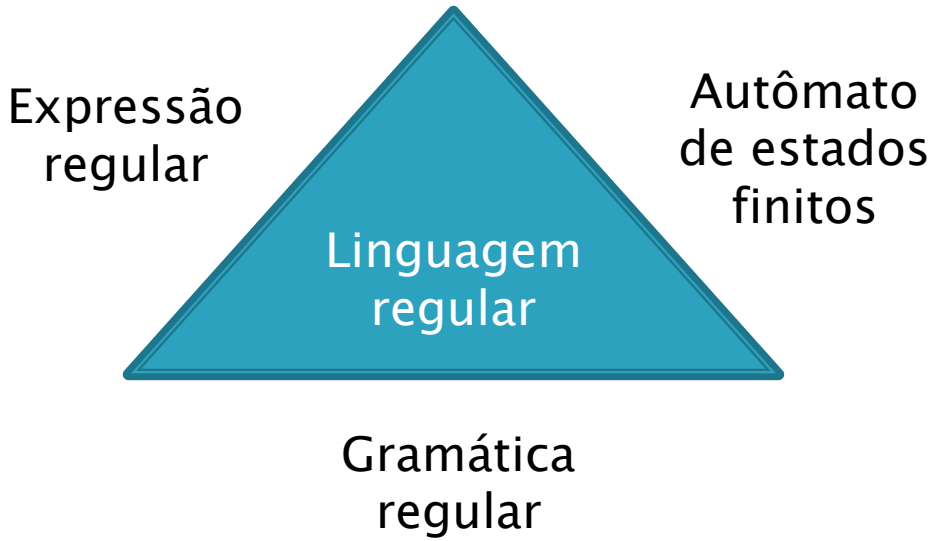
- Buscar todos os substantivos próprios
  - grep -E -w -o "[A-Za-z]+_NPROP" macmorpho.txt
- Buscar todos as construções do tipo substantivo+preposição+substantivo
  - grep -E -w -o "[A-Za-z]+_N [a-z]+_PREP [A-Za-z]+_N" macmorpho.txt
- Determinar qual classe aberta ocorre mais
  - grep -E -c "[A-Za-z]_N" macmorpho.txt
  - Mesma coisa para V, ADJ e ADV

Autômatos

- Expressões regulares implementadas como autômatos de estados finitos
  - Autômato: modelo matemático eficaz e elegante para lidar com expressões regulares
- Autômatos utilizados para revisão ortográfica, síntese e reconhecimento de fala, extração de informação, tradução automática, análise morfológica, análise morfossintática, etc.

Autômatos

- Poder representacional equivalente

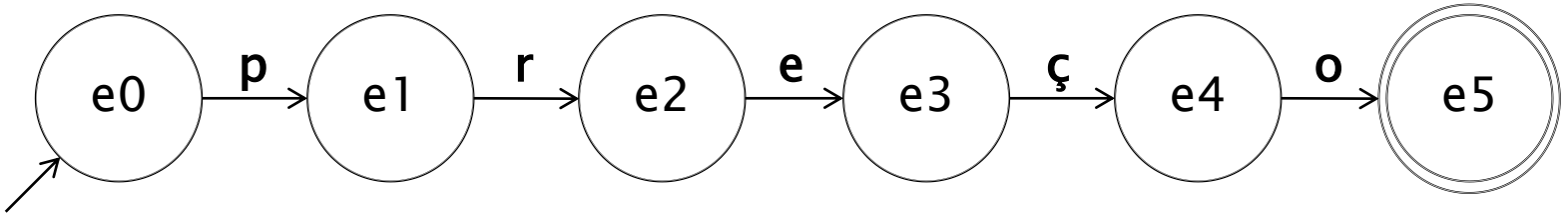


Autômatos

- **Componentes**
  - Estados que modelam o “sistema”
    - Pontos da análise sendo realizada, por exemplo
  - Símbolos de entrada
    - Letras das palavras, números, símbolos, etc.
  - Estados inicial e final
    - Início e fim do processo
  - Transições entre estados

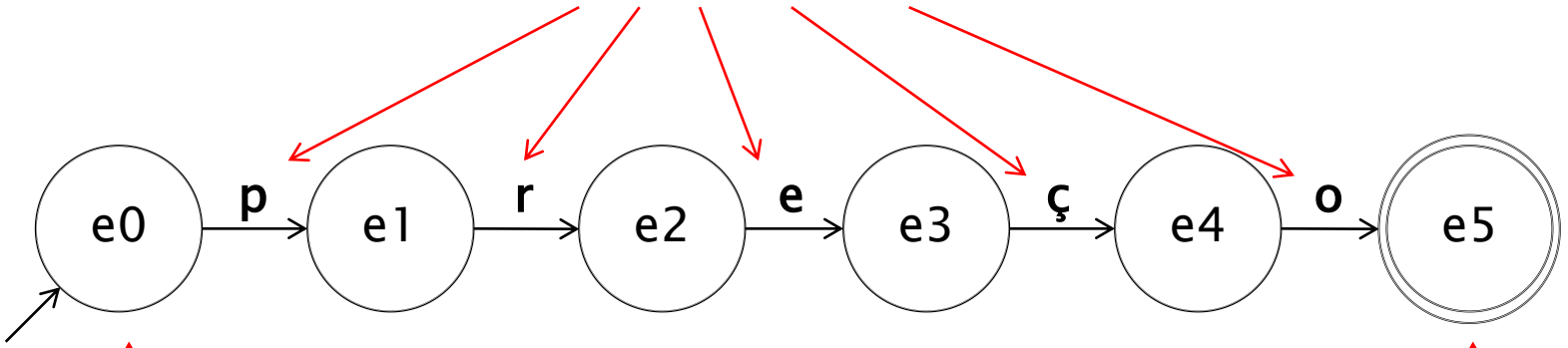
Exemplo

- preço



Exemplo

- preço



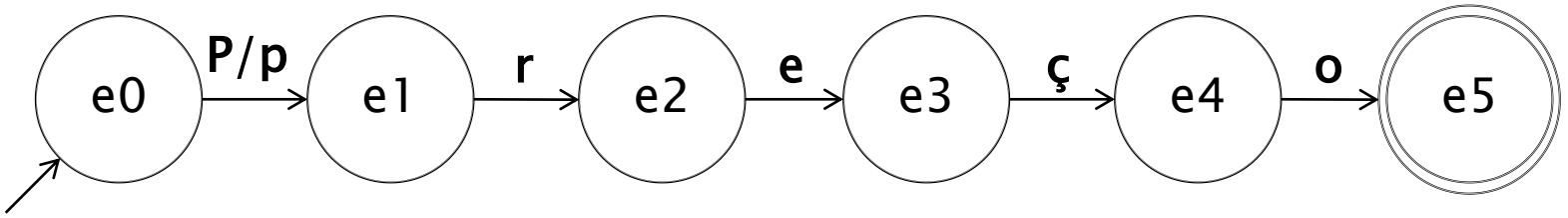
Estado inicial (indicado pela seta, em geral)

Começa-se no e0:
a cada transição, percorre-se uma letra da palavra de entrada;
se atingiu estado final, palavra reconhecida

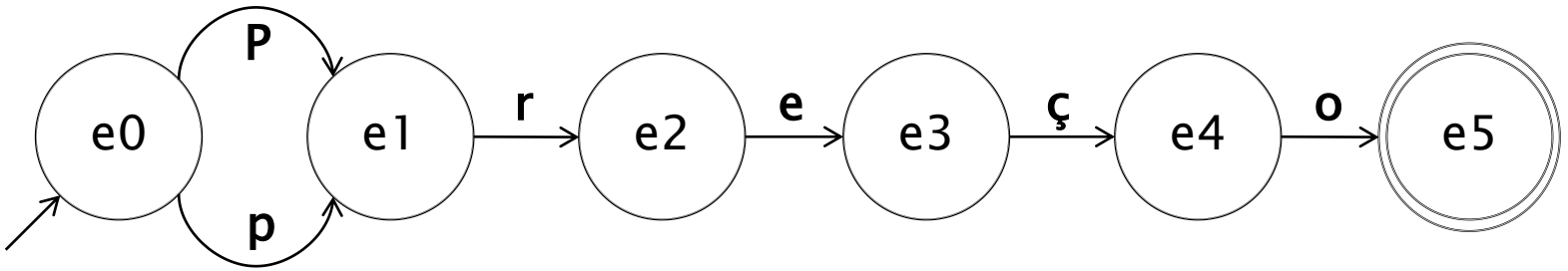
Estado final (indicado pelo contorno duplo, em geral)

Exemplo

- [Pp]reço

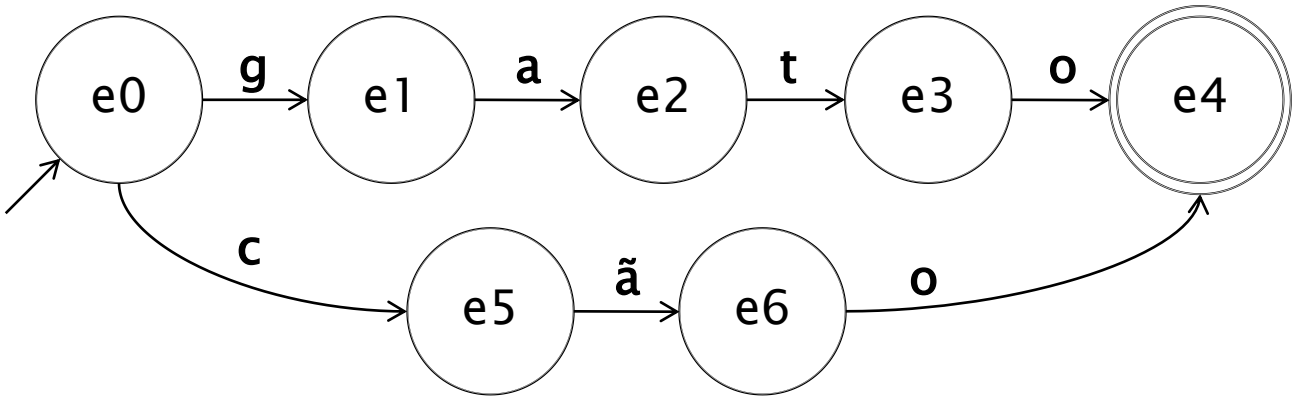


ou

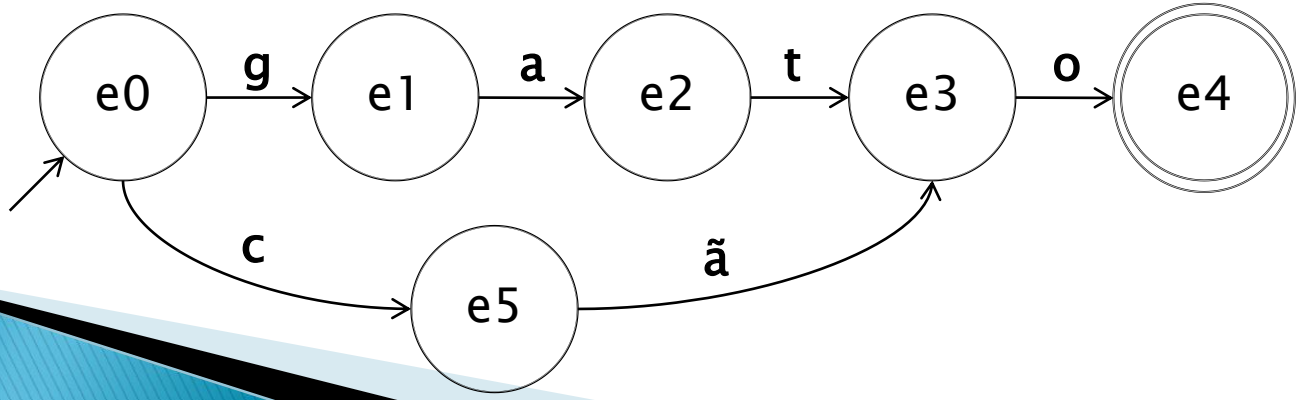


Exemplo

- (gato)|(cão)

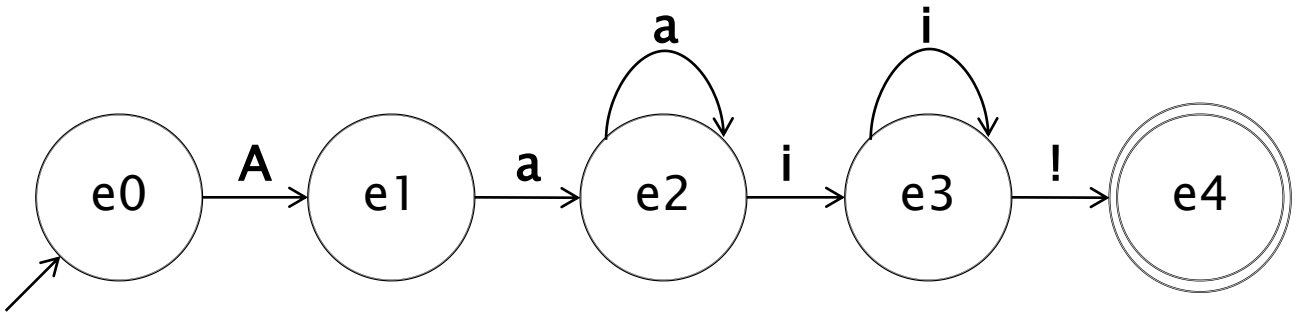


ou



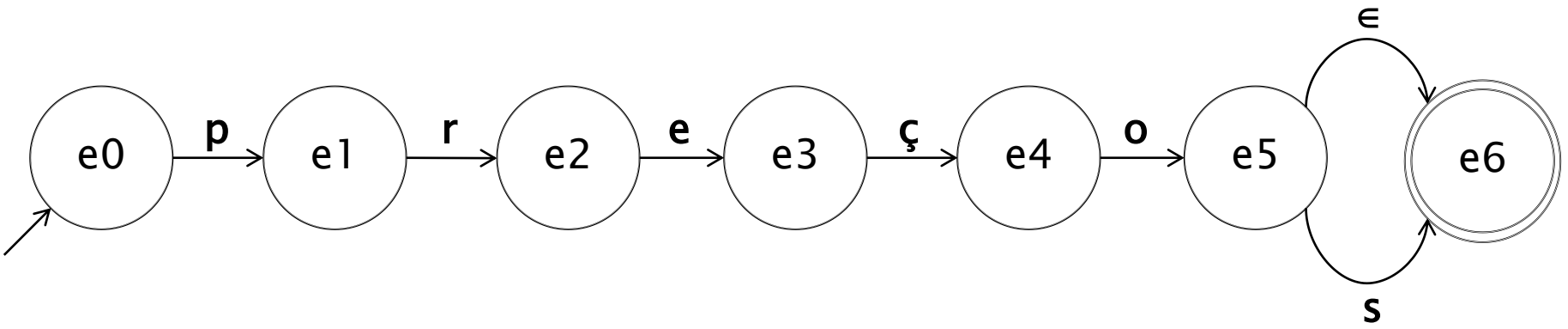
Exemplo

- Aa+i+!

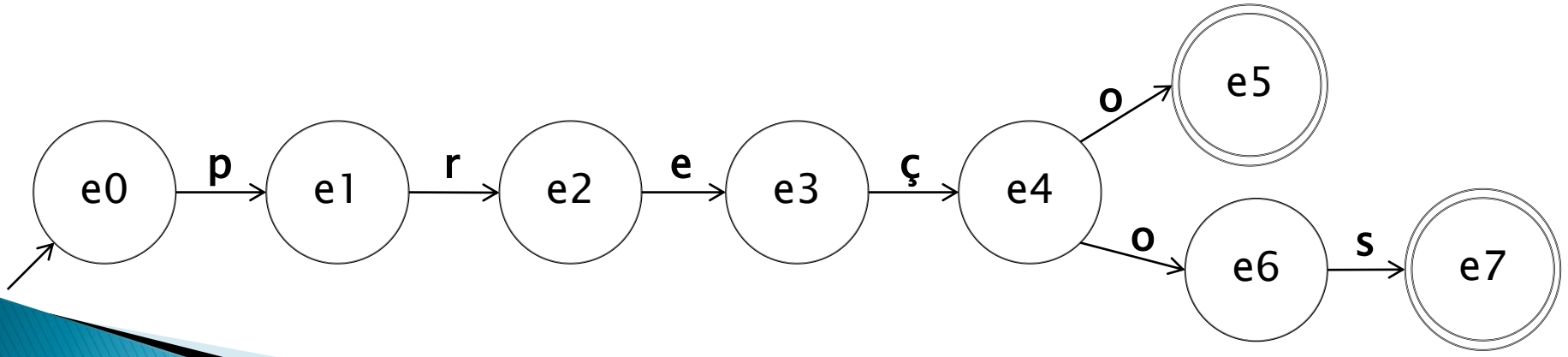


Exemplo

- preços?



ou



Exercício

- Criar autômato para reconhecer valores monetários
  - US\$ [0-9]+,[0-9]+ [mb]ilhões

Autômatos

- Variações
  - Transdutores
    - Além de reconhecerem a entrada, geram saída
      - Usados em análise morfológica
  - Modelos de Markov
  - Redes de transição

Análise morfológica

- *Parsing* morfológico
  - Analisar uma palavra e identificar seus componentes
    - Morfemas
    - Possibilidades
      - meninos → lema (menino), masculino (o), plural (+s), subst
      - meninos → radical (menin), masculino (+o), plural (+s), subst
      - meninas → lema (menino), feminino (–o +a), plural (+s), subst

Análise morfológica

- Relevância da tarefa
  - ???

Análise morfológica

- Relevância da tarefa
  - Reconhecer palavras e suas variações
    - *salvamento*, *salvamentos*, *salvação*, etc.
      - Revisão ortográfica, busca na web, sumarização, extração de informação
      - Stemming, lematização
  - ... e também produzir a forma adequada das palavras, derivar palavras novas, lidar com neologismos
    - *processamento*, mas não *processação*
      - Geração textual, tradução automática
        - "Máquina morfológica"
  - Caracterização léxica da língua, no geral

Terminologia básica

- Morfemas: unidade mínima de significado
  - Raiz/radical
    - Alguns diferenciam esses termos, outros não
  - Afixos

- Afixos
  - Prefixo: desamor, infeliz
  - Sufixo: lealdade, facilmente, quebrado, comia
  - Infixo: rabiscar
    - Raro, alguns dizem que não existe para o português
  - Circunfixo: anoitecer, descampado

Terminologia básica

- Morfe
  - Realização de um morfema
    - Morfema é abstrato, enquanto morfe é concreto
    - Exemplo: morfema de negação pode ser expresso pelos morfes in (de infeliz) ou i (de imutável)

- Alomorfes
  - Morfes que expressam um mesmo morfema
    - In e i para negação
    - Ante, pré e pró para anterioridade

Terminologia básica

- Processos principais de formação de palavras
  - Flexional: variações de uma mesma palavra
    - Flexão nominal: número, gênero
    - Flexão verbal: modo-tempo, número-pessoa
      - Adição de morfemas gramaticais
  - Derivacional: palavras novas
    - Podem mudar classe e sentido
      - "modelo" → "modelagem"
        - Adição de morfemas lexicais

Análise morfológica

- Para construir um parser morfológico, são necessários
  - Léxico
    - Radicais e afixos e suas possíveis classificações (substantivos, verbos, etc.)
  - Conhecimento de morfotática
    - Como os morfemas se ordenam para que as palavras se formem
      - Exemplo: em português, o morfema de plural aparece após o substantivo, e não antes
      - "Sintaxe da morfologia"
  - Regras ortográficas
    - Modelam mudanças que ocorrem nas palavras quando morfemas se combinam
      - Exemplo: casa+PL=casaS, mas flor+PL=flor<u>E</u>S

Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Exaustiva: léxico de formas analisadas (também chamadas flexionadas ou plenas)
      - Palavras com todas as suas variações
        - Pouca economia, redundância, compactação de arquivos

Exemplo do UNITEX-PB

DELAF_PB.dic - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

```
iguarias,iguaria.N:fp
iguaria,iguaria.N:fs
ilaqueações,ilaqueação.N:fp
ilaqueação,ilaqueação.N:fs
ilações,ilação.N:fp
ilação,ilação.N:fs
ilegalidades,ilegalidade.N:fp
ilegalidade,ilegalidade.N:fs
ilegitimidades,ilegitimidade.N:fp
ilegitimidade,ilegitimidade.N:fs
ilhéus,ilha.N:Dmp
ilhéu,ilha.N:Dms
ilhotas,ilha.N:Dfp
ilhota,ilha.N:Dfs
ilhas,ilha.N:fp
ilha,ilha.N:fs
ilhais,ilhal.N:mp
ilhal,ilhal.N:ms
```

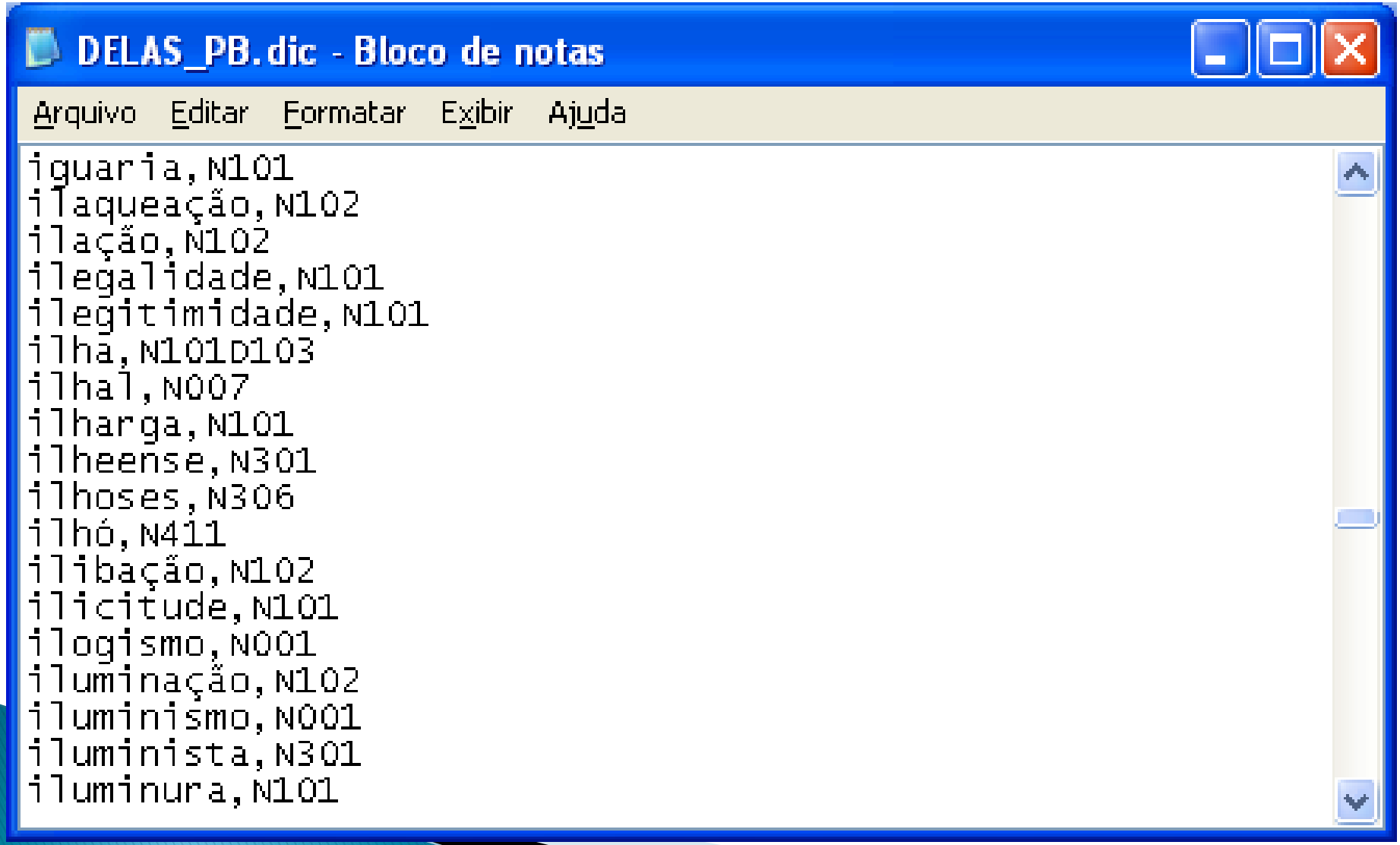
Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Econômica: léxico de raízes (ou de morfemas)
      - Listagem de raízes + regras de formação das palavras (morfotática e regras ortográficas)
        - Mais economia, processo mais caro

Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Meio termo
      - Léxico de lemas (ou formas canônicas) associados as suas variações
      - Palavras irregulares em formas plenas no léxico + léxico de raízes para palavras regulares
      - Etc.

Exemplo do UNITEX-PB



Análise morfológica

- Alternativa 1
  - Listagem de palavras
    - Problemas para lidar com
      - Novas palavras e variações: novos verbos (denominais, inclusive; por exemplo, “perfume” → “perfumar”), nomes próprios, etc.
      - Línguas morfologicamente complexas
        - Turco, por exemplo

Turco – exemplo

(Jurafsky e Martin, 2008)

uygarlaştıramadıklarımızdanmışsınızcasına

uygar +laş +tır +ama +dık +lar +ımız +dan +mış +sınız +casına

civilized +BEC +CAUS +NABL +PART +PL +P1PL +ABL +PAST +2PL +AsIf

"(behaving) as if you are among those whom we could not civilize"

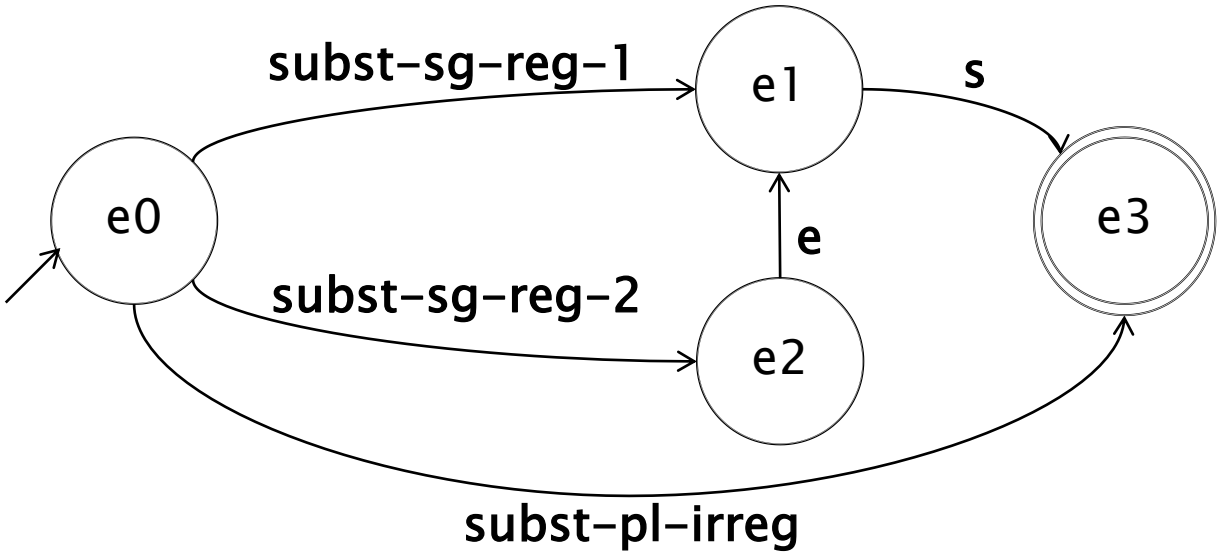
+BEC	"become"
+CAUS	the causative verb marker ('cause to X')
+NABL	"not able"
+PART	past participle form
+P1PL	1st person pl possessive agreement
+2PL	2nd person pl
+ABL	ablative (from/among) case marker
+AsIf	derivationally forms an adverb from a finite verb

Análise morfológica

- Alternativa 2
  - Codificação em forma de autômatos: maior eficiência computacional
    - De forma complementar com o léxico
      - Formas básicas/raízes no léxico e regras de formação de palavras (morfotática e regras ortográficas) mapeadas em autômatos
    - De forma isolada
      - Todo o léxico da língua mapeado em autômatos

Exemplo simples

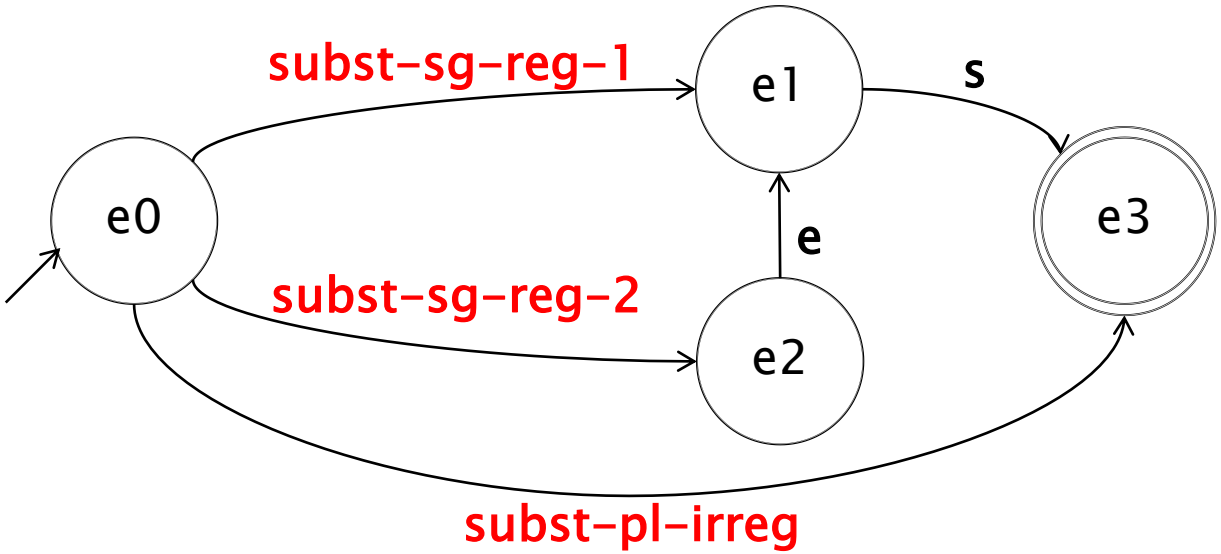
- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato



subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato



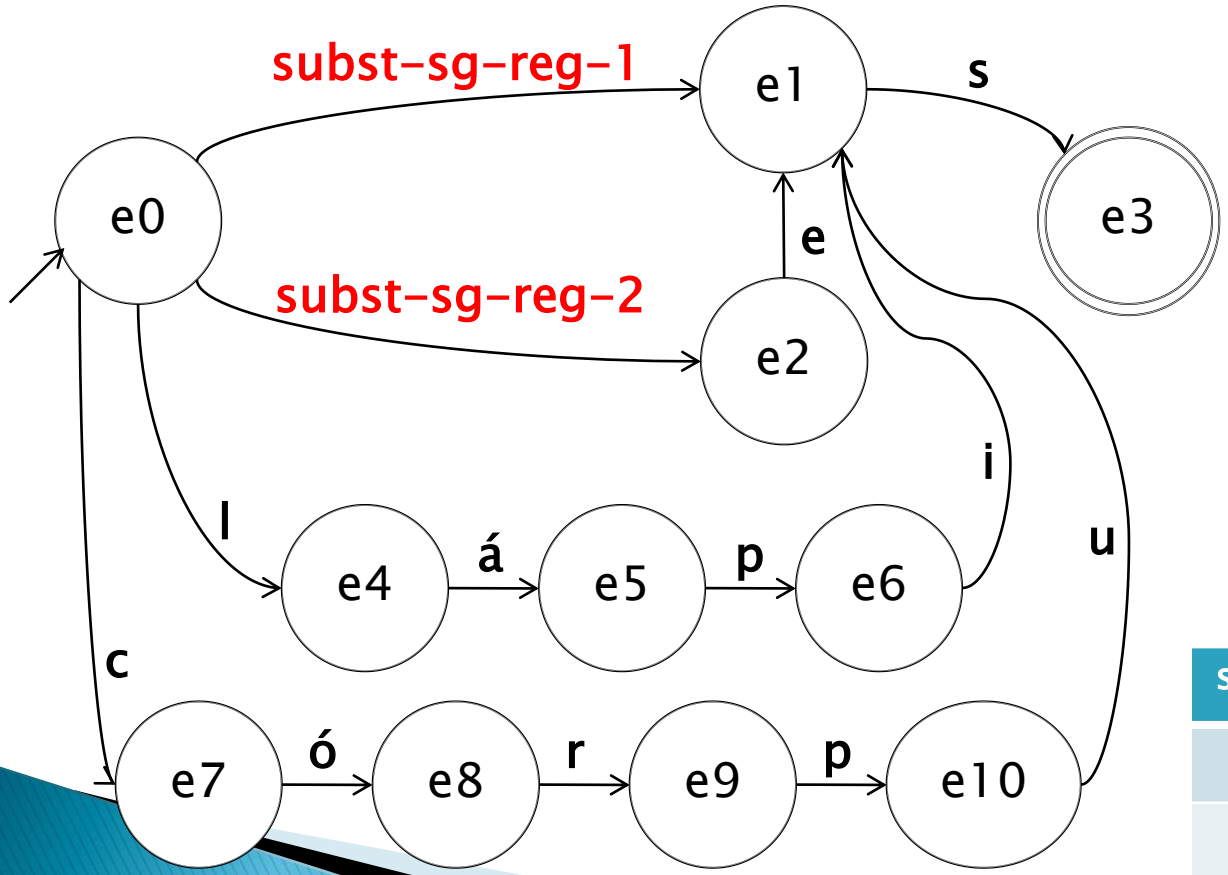
Podem ser substituídos pelos autômatos correspondentes!

Como?

subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Reconhecimento/geração de alguns substantivos no plural
  - Léxico de lemas + autômato

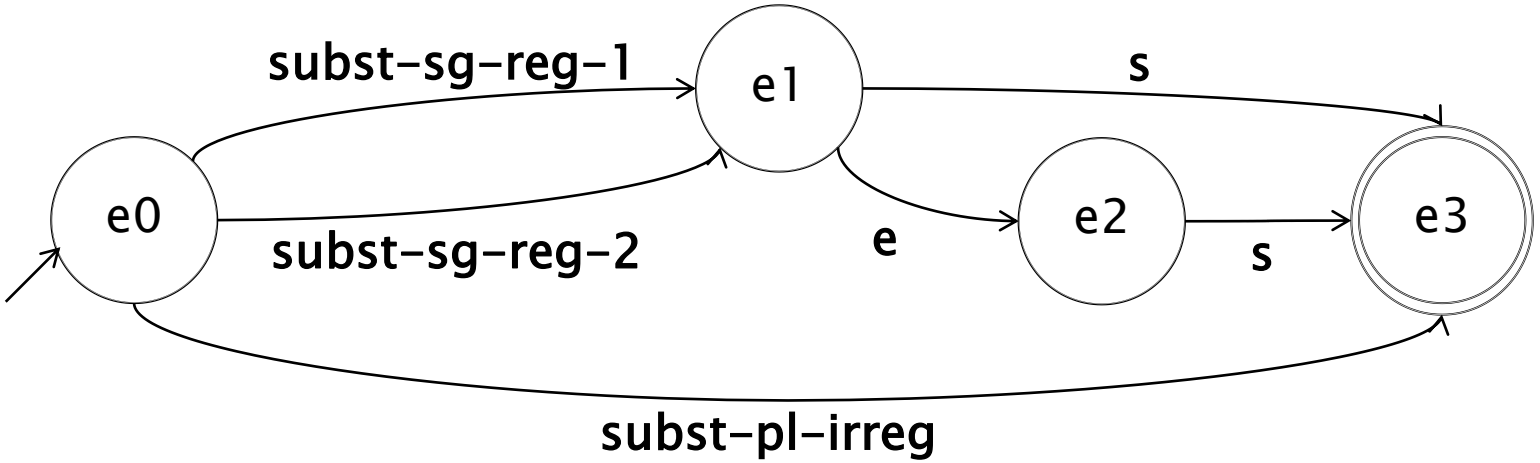


Podem ser substituídos pelos autômatos correspondentes!

subst-sg-reg-1	subst-sg-reg-2
casa	flor
porta	lar
...	...

Exemplo simples

- Cuidado com *overgeneration* !
  - O que acontece no caso abaixo?

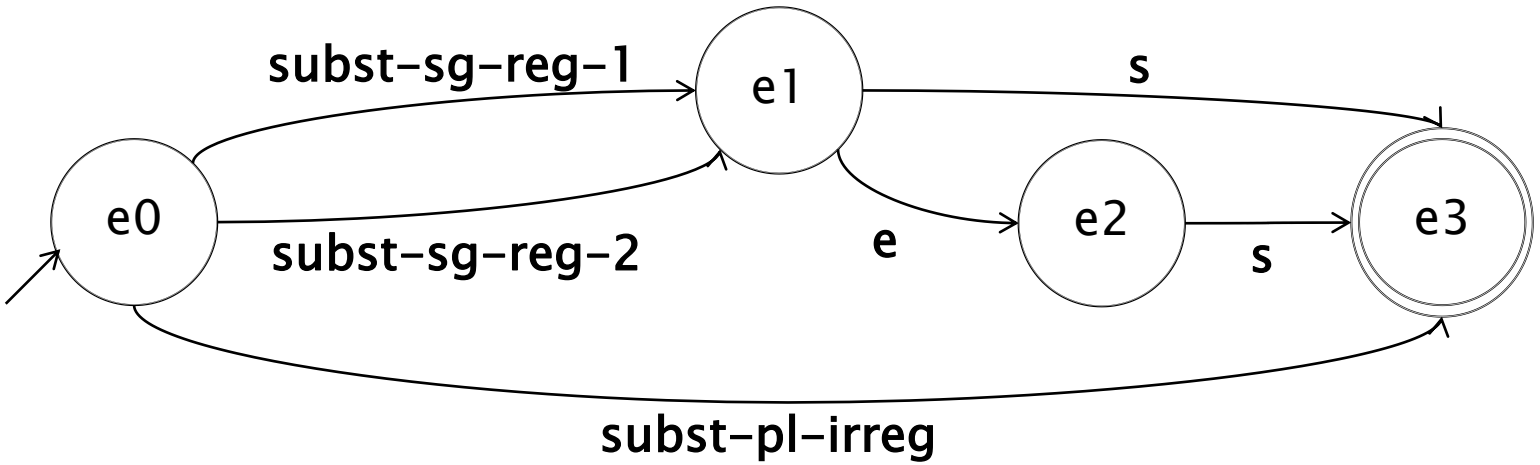


subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Exemplo simples

- Cuidado com *overgeneration* !
  - O que acontece no caso abaixo?

casas
*casaes
*flors
flores
...



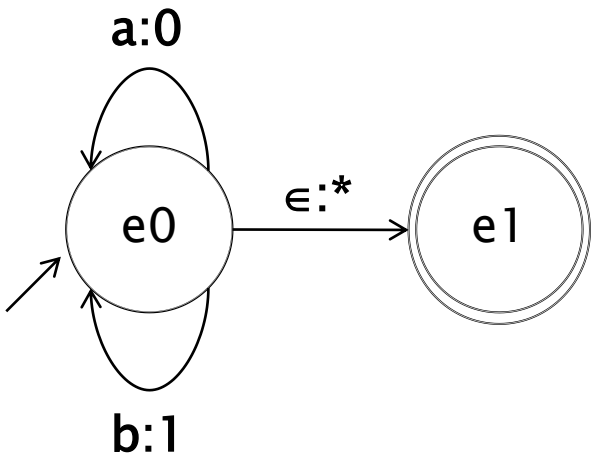
subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
porta	lar	córpus
...	...	...

Análise morfológica

- Para nossa tarefa, precisamos de mais poder
  - Além de se reconhecer/gerar as palavras, é necessário identificar os componentes
    - gatos → gato + SUBST + MASC + PL
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind
  - Transdutores
    - Reconhecem a entrada e, em paralelo, geram saída

Transdutores

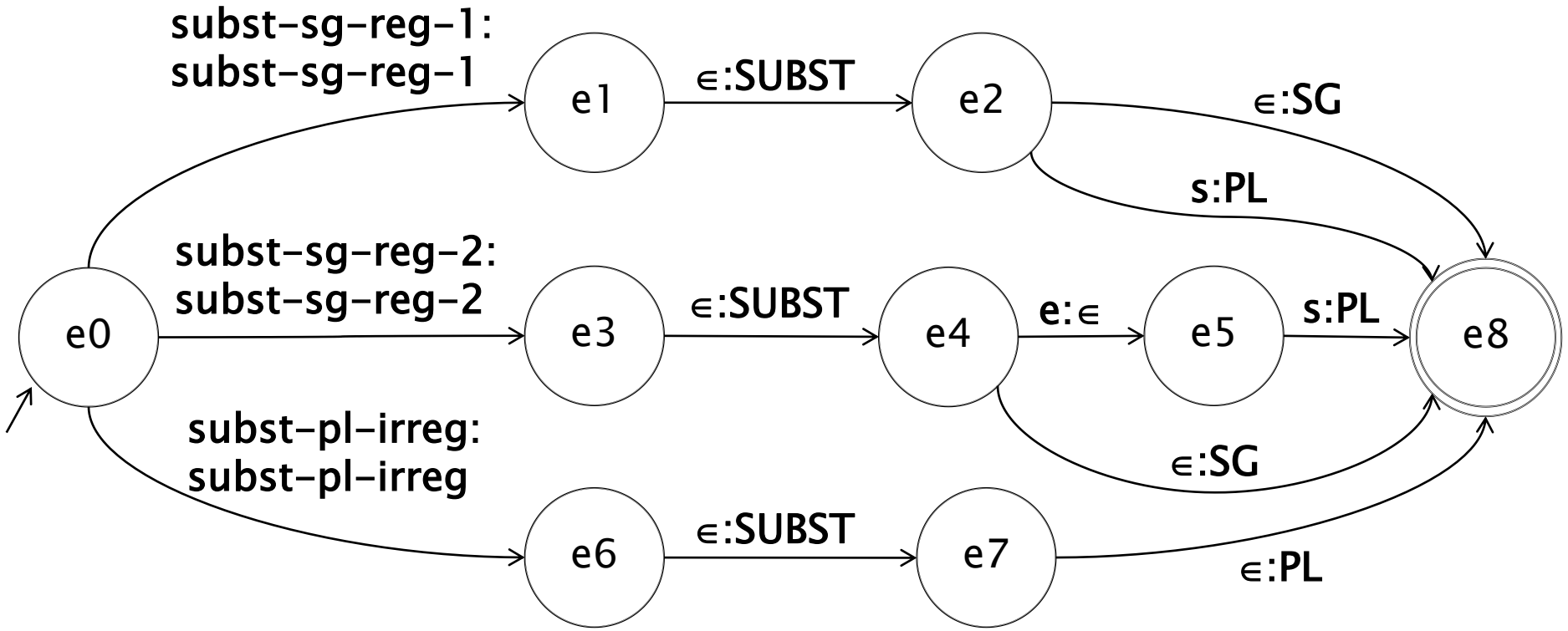
- Lendo a_s e b_s e gerando 0_s e 1_s, respectivamente, terminando com *



Análise de abba

Transdutores: exemplo

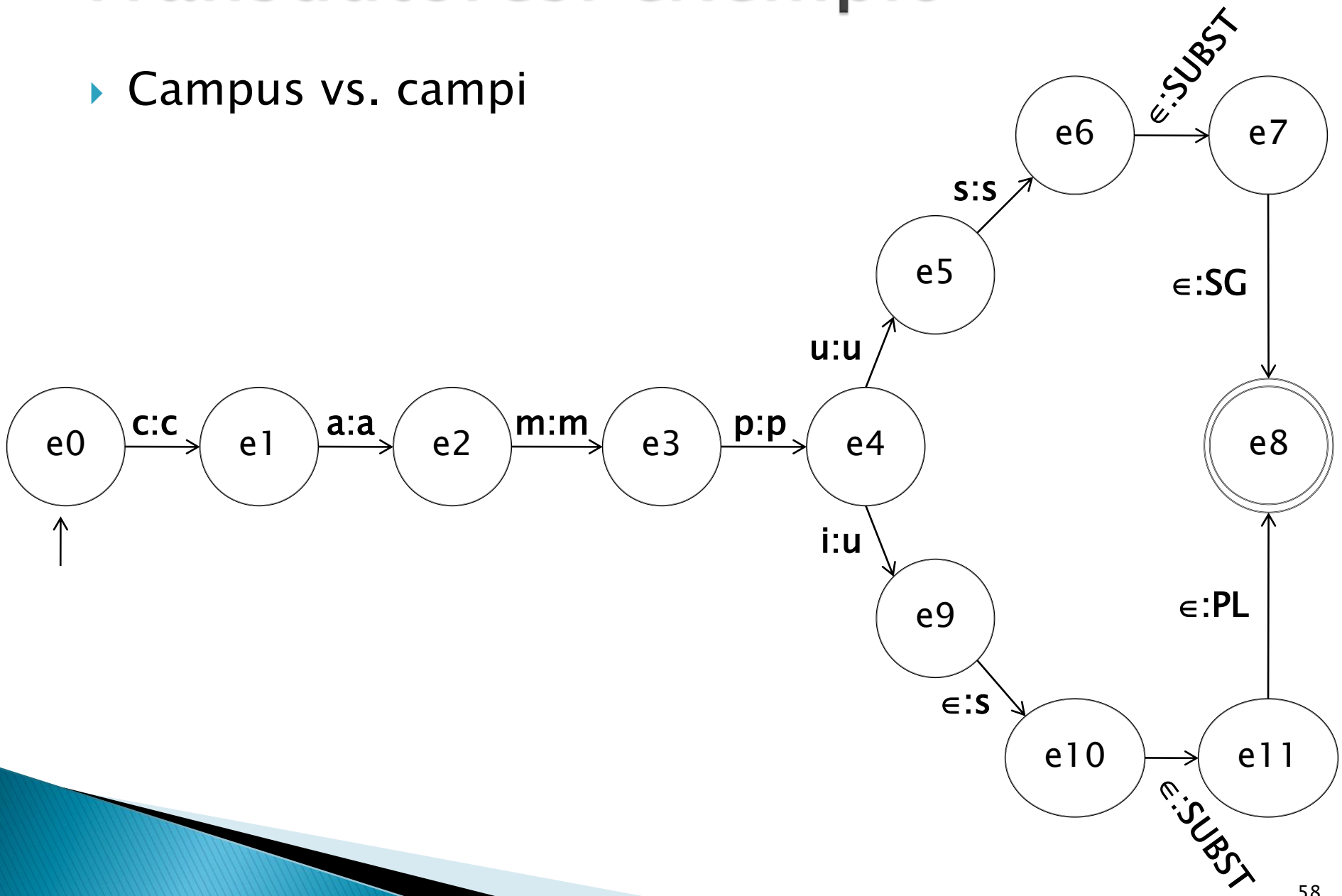
- Releitura do autômato de substantivos



subst-sg-reg-1	subst-sg-reg-2	subst-pl-irreg
casa	flor	lápis
...	...	...

Transdutores: exemplo

- Campus vs. campi



Exercício

- Menino, menina, meninos, meninas: exercício
  - Reconhecer número, gênero, raiz e etiqueta morfossintática

Transdutores

- E casos como o de “canto”?
  - Como identificar que “canto” pode ser um verbo ou um substantivo, gerando-se os atributos correspondentes para cada caso?
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind

Transdutores

- E casos como o de “canto”?
  - Como identificar que “canto” pode ser um verbo ou um substantivo, gerando-se os atributos correspondentes para cada caso?
    - canto → canto + SUBST + MASC + SG
    - canto → cantar + V + 1P + SG + Pind
  - **A palavra seria reconhecida por mais de um transdutor!**
    - Análise morfossintática para desambiguar

Origens da Morfossintaxe

- Dionísio Trácio, 100 AC
  - Esboço da gramática do grego
  - Cunhou o vocabulário atual
    - Sintaxe, ditongo, etc.
    - 8 etiquetas morfossintáticas: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio, artigo
      - Vocabulário usado até hoje!
- Morfossintaxe
  - Morfologia: tipos de afixos possíveis variam com a classe
  - Sintaxe: palavras com comportamentos/funções similares em seus contextos são de uma mesma classe
  - Algo mais?

Origens da Morfossintaxe

- Dionísio Trácio, 100 AC
  - Esboço da gramática do grego
  - Cunhou o vocabulário atual
    - Sintaxe, ditongo, clítico, etc.
    - 8 etiquetas morfossintáticas: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio, artigo
      - Vocabulário usado até hoje!
- Morfossintaxe
  - Morfologia: tipos de afixos possíveis variam com a classe
  - Sintaxe: palavras com comportamentos/funções similares em seus contextos são de uma mesma classe
  - Semântica: substantivos têm uma preferência por objetos, lugares e coisas, adjetivos por propriedades, etc.
  - Pragmática

Conjuntos de etiquetas

- Variam muito
  - Penn Treebank (Marcus et al., 1993): 45
  - Brown Corpus (Francis, 1979): 87
  - CLAWS 7 (Garside et al. 1997): 146
  - Palavras (Bick, 2000): 14
  - Mac-Morpho/Lácio-Web (Aluísio et al., 2003): 31
  - Universal Dependencies (Nivre et al., 2016): 17

Exemplo: Penn Treebank

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun

RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Exemplo: Mac-Morpho

Tag	Definition
ADJ	open-class noun modifier
ADV-KS-REL	relative subordinating Adverb
ADV-KS	Non-relative subordinating Adverb
ADV	Non-subordinating adverb
ART	Article
KC	coordinating conjunction
KS	coordinating conjunction
IN	interjection
N	open-class noun phrase nucleus
NPROP	proper noun
NUM	numeral as a noun modifier
PCP	past participle or adjective
PDEN	emphasis/focus
PREP	preposition
PROPESS	personal pronoun
PRO-KS-REL	relative subordinating pronoun
PRO-KS	Non-relative subordinating pronoun
PROSUB	non-subordinating pronoun as a noun phrase nucleus
PROADJ	Non-subordinating pronoun as a modifier
VAUX	Auxiliary verb
V	Non-auxiliary verb
CUR	Currency symbol

Compl. Tag	Definition
|EST	foreign
|AP	apposition
|+	contraction/ enclitic
|!	mesoclitic
|[|... |]	beginning, middle part, and end of discontinuous compound (further discussed in Section 3)
|TEL	phone number
|DAT	date
|HOR	time
|DAD	formatted data not falling into above categories

Exemplo: Universal Dependencies

Universal POS tags

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Terminologia

- Palavras de classes fechadas, palavras funcionais/gramaticais
  - Conjunto de palavras varia pouco
    - Preposições, conjunções, artigos
- Palavras de classes abertas, palavras lexicais
  - O conjunto varia bastante, surgindo novas palavras
    - Substantivos, verbos

Terminologia

- Substantivos/nomes
  - Comuns, próprios
  - Contáveis (abelha, casa), incontáveis (ar, açúcar)
- Verbos
  - Principais, auxiliares
- Advérbios
  - Tempo, local, modo, direção, etc.
- Conjunções
  - Coordenativas e subordinativas
- Pronomes
  - Pessoais, possessivos, interrogativos, relativos, etc.

Etiquetas morfossintáticas

- Nem sempre a distinção é simples
  - Advérbios vs. preposições
    - *Near*, *around*
  - Adjetivos vs. particípios
    - Eles estão casados.
  - Advérbios: tudo que não cabe nas outras classes

Etiquetação morfossintática

- Tagging, ou parsing morfossintático
  - Associação de etiquetas às palavras de uma sentença
    - Faz-se necessário, portanto, tokenização e segmentação sentencial
  - Tarefa de desambiguação: dentre as etiquetas (tags) possíveis previstas (pelo léxico, por exemplo), determinar a mais apropriada
    - Contexto desambigua!

Tagging

- Útil para um infinidade de tarefas de PLN
  - *Stemming*, lematização
  - Tradução, sumarização, auxílio à escrita
  - Identificação de autoria, extração de informação
  - Pesquisas linguísticas variadas: neologismos, comportamento de palavras, etc.
  - Etc.

Tagging

- Algumas abordagens principais
  - Regras
    - Por exemplo, uma palavra antecedida por um artigo é um substantivo
  - Probabilidades (também pode ser AM)
    - Classe mais provável de uma palavra em função das palavras vizinhas, com aprendizado a partir de córpus
  - Aprendizado de Máquina (AM)
    - A máquina aprende a identificar automaticamente as classes gramaticais, aprendendo padrões
- Hibridismo também é possível
  - Por exemplo, aprendizado de regras a partir de córpus

Tagging: regras

- Primeiras abordagens (década de 60)
  - 2 passos tradicionais
    - Léxico fornece possíveis classes para cada palavra
    - Regras criadas manualmente são utilizadas para desambiguar
- Mais recentemente
  - Dicionários maiores e muito mais regras!

Exemplo: EngCG tagger (Voutilainen, 1999)

→ Análise morfológica da sentença (tag correta em negrito)

Pavlov had shown that salivation.

Pavlov:	**PAVLOV N NOM SG PROPER**
had:	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown:	**SHOW PCP2 SVOO SVO SV**
that:	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation:	**N NOM SG**
.:	**PUNC DOT**

Exemplo: EngCG tagger (Voutilainen, 1999)

→ Aplicação de regras para determinar as melhores tags

Exemplo de regra

WORD: that
IF
next word is adj, adverb, or quantifier AND
after this word there is the sentence boundary AND
the previous word is not a verb that allows adjs as complements
THEN eliminate non-adv tags
ELSE eliminate adv tag

Exemplo: ReGra (Martins et al., 1998)

Exemplo de entrada (com erros) para o revisor gramatical

OS	Definite article (the): masculine, plural
	Personal pronoun (them): masculine, plural
MENINO	Noun (boy): masculine, singular
PREFERE	Verb (to prefer): 3rd person, singular, present tense, indicative, transitive
BRINCAR	Verb (to play): infinitive
DO	Contraction: preposition (of) + definite article (the): masculine, singular
QUE	Relative Pronoun (which)
	Adverb (what)
	Conjunction (than)
ESTUDAR	Verb (to study): infinitive

Regras de desambiguação utilizadas

OS	Definite article (the): the following word is a masculine noun
QUE	Conjunction (than): the previous word is a contraction (preposition + article)

Tagging: regras

- Zellig Harris (1962) e o primeiro tagger (provavelmente)
  - 14 regras de desambiguação

- UNITEX-PB (Muniz, 2004)
  - 80 regras de desambiguação no formalismo ELAG

ELAG
16 regras para Adjetivos
30 regras para Advérbios
22 regras para Artigos
12 regras para Substantivos

Tagging: probabilidades

- Abordagem antiga, desde a década de 60
- Modelo de Markov Oculto (HMM – *Hidden Markov Model*), um dos mais utilizados
  - Um tipo de inferência bayesiana
  - Tarefa de classificação: dadas algumas observações, quais as classes mais prováveis
    - Tagging: dada uma sequência de palavras, qual a sequência de tags mais provável

Tagging: probabilidades

- Tagging: dada uma sequência de palavras, qual a sequência de tags mais provável
  - Queremos a sequência de tags SeqTags que maximize a probabilidade P(SeqTags|SeqPalavras)

$$\hat{SeqTags} = \underset{SeqTags}{\operatorname{argmax}}\ P(SeqTags \mid SeqPalavras)$$

  - Exemplo: *O menino prefere brincar do que estudar*
    - SeqTags 1: art subst verbo verbo contração pro verbo
    - SeqTags 2: art subst verbo verbo contração adv verbo
    - SeqTags 3: art subst verbo verbo contração conj verbo
    - SeqTags 4: pro subst verbo verbo contração pro verbo
    - SeqTags 5: pro subst verbo verbo contração adv verbo
    - SeqTags 6: pro subst verbo verbo contração conj verbo
      - Qual a melhor sequência de tags, ou seja, qual destas sequências maximiza a probabilidade P(SeqTags|SeqPalavras)?

Tagging: probabilidades

- Qual a maior probabilidade?

$$\hat{SeqTags} = \underset{SeqTags}{\operatorname{argmax}}\ P(SeqTags \mid SeqPalavras)$$

  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)
  - P(art subst verbo verbo contração adv verbo | O menino prefere brincar do que estudar)
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração adv verbo | O menino prefere brincar do que estudar)
  - P(pro subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar)

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$ → constante

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$

$P(SeqPalavras)$ → constante

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$

$P(SeqPalavras)$ → constante

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

Tagging: probabilidades

- Como calcular essas probabilidades?

$$P(SeqTags \mid SeqPalavras)$$

$$P(SeqTags \mid SeqPalavras) = \frac{P(SeqPalavras \mid SeqTags) \times P(SeqTags)}{P(SeqPalavras)}$$ → constante

$$P(SeqTags \mid SeqPalavras) = P(SeqPalavras \mid SeqTags) \times P(SeqTags)$$

- Simplificações para facilitar o cálculo
  - Cada palavra depende apenas de sua tag
  - Uma tag depende apenas da tag anterior na sentença

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

Como calcular essas 2 probabilidades?

Tagging: probabilidades

- Exemplo
  - Supondo que se usa o Brown Corpus

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

$P(palavra_i|tag_i) = P(\text{“is”}|VBZ)$

tag=VBZ → 21.627 ocorrências no córpus
palavra=“is” com tag=VBZ → 10.073 ocorrências no córpus

P(“is”|VBZ) = número de vezes de “is” com VBZ / número de VBZ
P(“is”|VBZ) = 10.073/21.627 = 0.47 ou 47%

Tagging: probabilidades

- Exemplo
  - Supondo que se usa o Brown Corpus

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

$P(tag_i|tag_{i-1}) = P(NN|DT)$

tag_{i-1}=DT → 116.454 ocorrências no córpus
tag_i=NN com tag_{i-1}=DT → 56.509 ocorrências no córpus

P(NN|DT) = número de vezes de NN precedido por DT/ número de DT
P(NN|DT) = 56.509/116.454 = 0.49 ou 49%

Tagging: probabilidades

- Exemplo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar)

$$P(SeqTags \mid SeqPalavras) = \prod_{i=1}^{\text{número de palavras}} P(palavra_i \mid tag_i) \times P(tag_i \mid tag_{i-1})$$

  - = P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|pro) x P(pro|contração) x
    P(estudar|verbo) x P(verbo|pro)

- Faz-se isso para todas as possíveis sequências de tags (com probabilidades aprendidas de córpus)
  - A sequência com maior probabilidade é escolhida

Tagging: probabilidades

- Considerando a probabilidade abaixo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|**pro**) x P(**pro**|contração) x
    P(estudar|verbo) x P(verbo|**pro**)
- Por que ela é provavelmente menor do que a abaixo – que é a interpretação correta?
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|**conj**) x P(**conj**|contração) x
    P(estudar|verbo) x P(verbo|**conj**)

Tagging: probabilidades

- Considerando a probabilidade abaixo
  - P(art subst verbo verbo contração pro verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|pro) x P(pro|contração) x
    P(estudar|verbo) x P(verbo|pro)
- Por que ela é provavelmente menor do que a abaixo – que é a interpretação correta?
  - P(art subst verbo verbo contração conj verbo | O menino prefere brincar do que estudar) =
    P(O|art) x P(art|<início da sentença>) x
    P(menino|subst) x P(subst|art) x
    P(prefere|verbo) x P(verbo|subst) x
    P(brincar|verbo) x P(verbo|verbo) x
    P(do|contração) x P(contração|verbo) x
    P(que|conj) x P(conj|contração) x
    P(estudar|verbo) x P(verbo|conj)

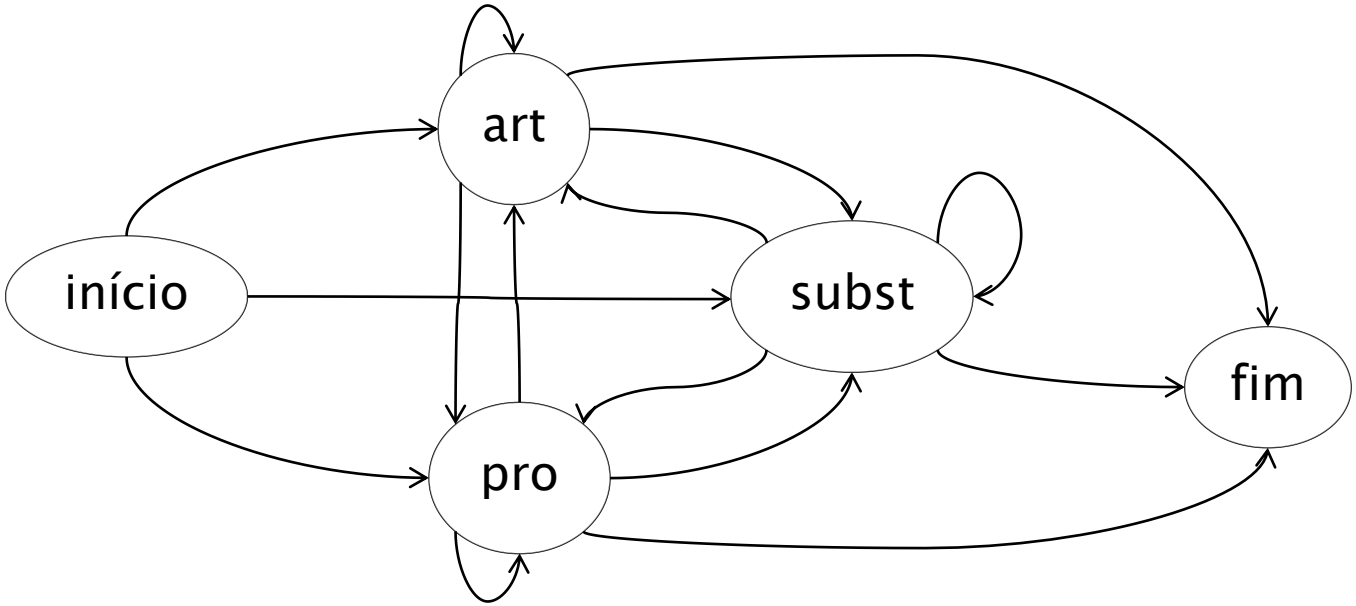
Esses termos devem ser mais prováveis do que os correspondentes na intepretação errada

Tagging: probabilidades

- Modelo de Markov oculto
  - Modelam-se eventos observados (palavras) e eventos não observados, ou seja, ocultos (tags)
  - Como dito anteriormente, tipo especial de autômato
    - Probabilidades nos arcos (transições): $P(tag_i|tag_{i-1})$
    - Probabilidades nos nós: $P(palavra_i|tag_i)$

Tagging: probabilidades

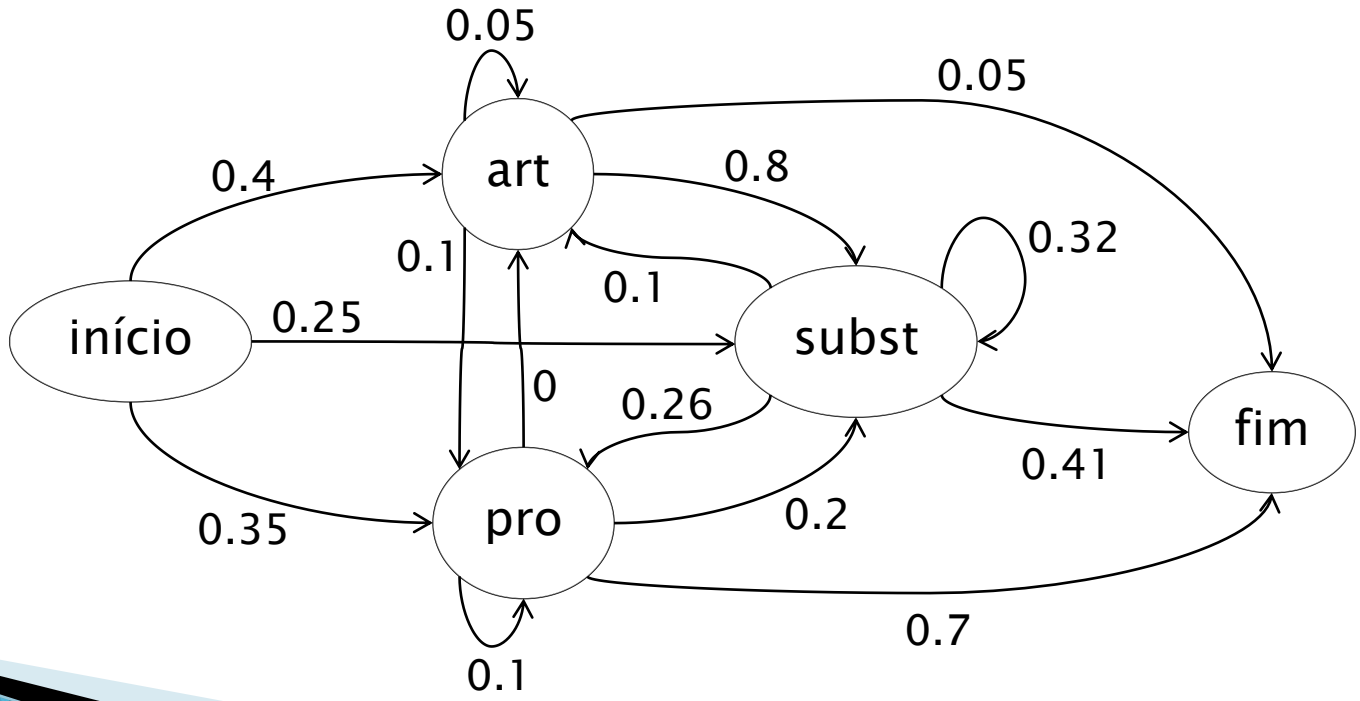
- Modelo de Markov oculto
  - Exemplo hipotético



Tagging: probabilidades

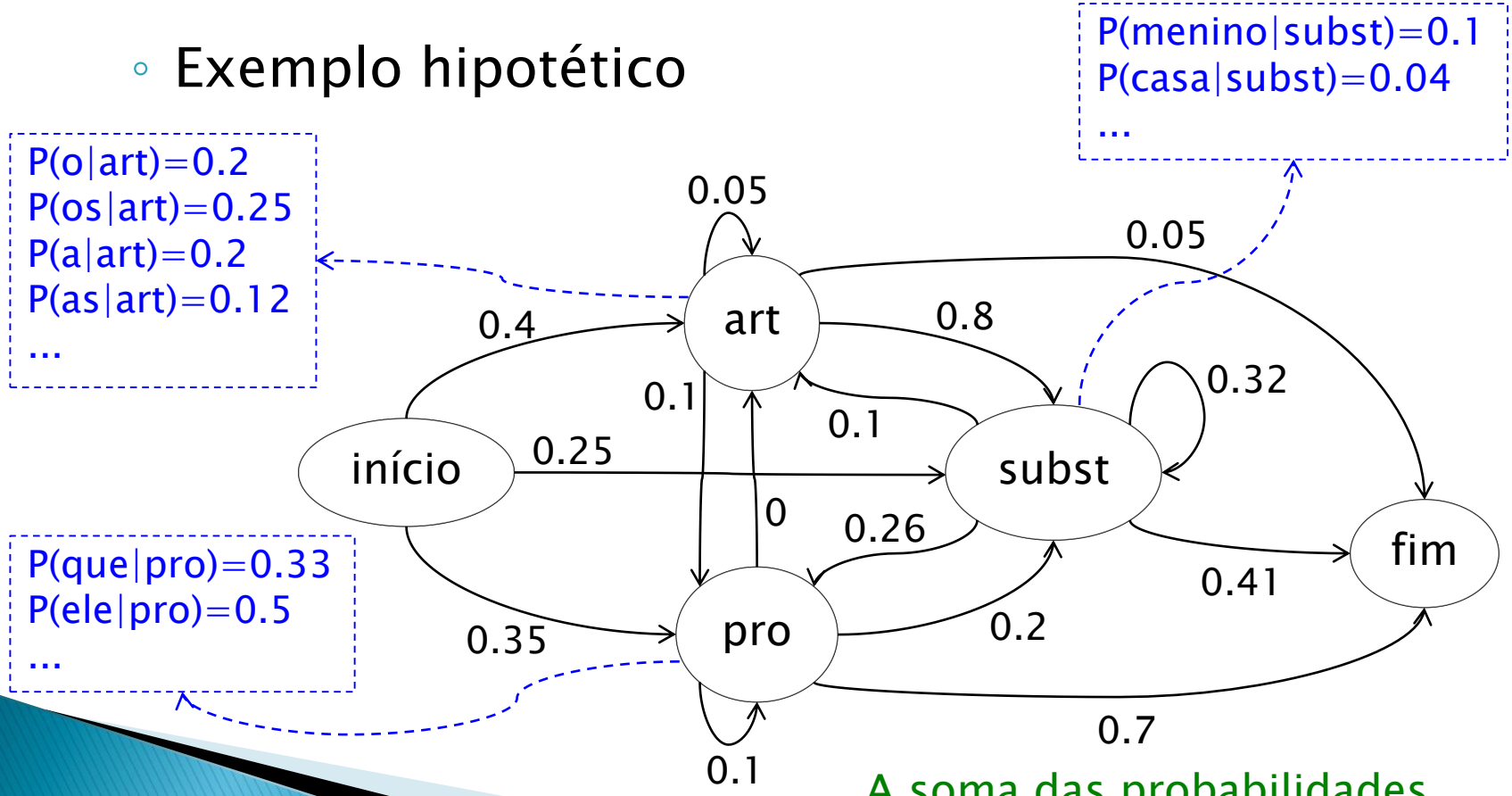
- Modelo de Markov oculto
  - Exemplo hipotético

A soma das probabilidades dos arcos que saem de cada nó devem somar 1



Tagging: probabilidades

- Modelo de Markov oculto
  - Exemplo hipotético



A soma das probabilidades associadas a cada nó devem somar 1

Tagging: probabilidades

- Taggers mais comuns usam normalmente 2 tags anteriores como contexto
  - $P(tag_i|tag_{i-1},tag_{i-2})$
- É preciso otimizar a busca por sequências de tags mais prováveis, senão pode haver explosão combinatória
  - **Programação dinâmica** (estudaremos logo)
- É preciso lidar com probabilidades muito baixas ou próximas de zero
  - Por exemplo, situações em que $P(tag_i|tag_{i-1}, tag_{i-2})\approx 0$
  - Solução usual: combinar probabilidades ponderadas
    - $P(tag_i|tag_{i-1}, tag_{i-2}) = p_1*P(tag_i|tag_{i-1},tag_{i-2})+p_2*P(tag_i|tag_{i-1})+p_3*P(tag_i)$, com $p_1+p_2+p_3=1$

EXERCÍCIO

- Construir o modelo de Markov oculto (grafo e probabilidades) a partir da receita de churros abaixo
  - *Compre doce de leite. Faça a massa com farinha, água e sal. Frite porções da massa. Recheie com o doce.*
    - Se desejarem, usem o LX-Tagger como suporte

Tagging: aprendizado de máquina

- *Transformation-Based tagging*
  - Regras são aprendidas e aprimoradas automaticamente
    - Várias iterações
    - A cada iteração, o processo melhora
    - Ao estabilizar, fim do processo de aprendizado

Tagging: aprendizado de máquina

- *Transformation-Based tagging*
  - Processo básico com base em um córpus anotado manualmente
    - Inicialmente, anota-se um córpus automaticamente, assumindo-se que a tag de uma palavra é a sua tag mais frequente (segundo um córpus/léxico)
    - Verificam-se os erros cometidos (comparando-se com a anotação humana correspondente) e, dentre todas as possibilidades de correção, monta-se uma regra de correção com maior precisão
    - Aplica-se essa regra nova em todo o córpus
    - Verificam-se novamente os erros cometidos e monta-se uma segunda regra de correção com maior precisão
    - E assim por diante, até não se obter mais melhora de performance

Tagging: aprendizado de máquina

- Exemplo

1. Inicialmente, anotação com base em frequência

   ... is/VBZ expected/VBN to/TO race/NN tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/VB to/TO...

2. Verificando-se os erros, aprende-se uma nova regra

   Troque NN para VB quando a tag anterior é TO

3. Corrige-se a etiquetação anterior

   ... is/VBZ expected/VBN to/TO race/VB tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/VB to/TO...

Tagging: aprendizado de máquina

- Exemplo

4. Aprende-se uma nova regra com base nos erros existentes

   Troque VB para NN quando a tag anterior é DT e a posterior é TO

5. Corrige-se a etiquetação anterior

   ... is/VBZ expected/VBN to/TO race/VB tomorrow/NN
   ... the/DT race/NN for/IN outer/JJ space/NN
   Book/VB the/DT flight/NN to/TO...

6. E assim por diante

Tagging: aprendizado de máquina

- Exemplo
  - **Resultado** do processo

    Regra 1: etiquete as palavras com suas tags mais frequentes

    Regra 2: troque NN para VB quando a tag anterior é TO

    Regra 3: troque VB para NN quando a tag anterior é DT e a posterior é TO

    ...

Tagging: aprendizado de máquina

- *Transformation-Based tagging*
  - Ao final do processo, há um conjunto de regras ordenadas que devem ser aplicadas sequencialmente para etiquetar um novo texto
  - Como há muitas regras possíveis de serem aprendidas, costuma-se limitar as estruturas aceitas de regras
    - Troque a tag da palavra corrente de A para B se a palavra anterior tem a tag X
    - Troque a tag da palavra corrente de A para B se a palavra anterior tem a tag X e a posterior tem a tag Y
    - Etc.

Caso contrário, o que acontece?

Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o português (Fonseca et al., 2015)
    - Entrada: *word embeddings* (em vez das palavras em si)
    - Saída: escore para cada tag

Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o português (Fonseca et al., 2015)
    - Entrada: *word embeddings* (em vez das palavras em si)
    - Saída: escore para cada tag

Palavras são mapeadas em seus vetores
→ Vantagens disso?

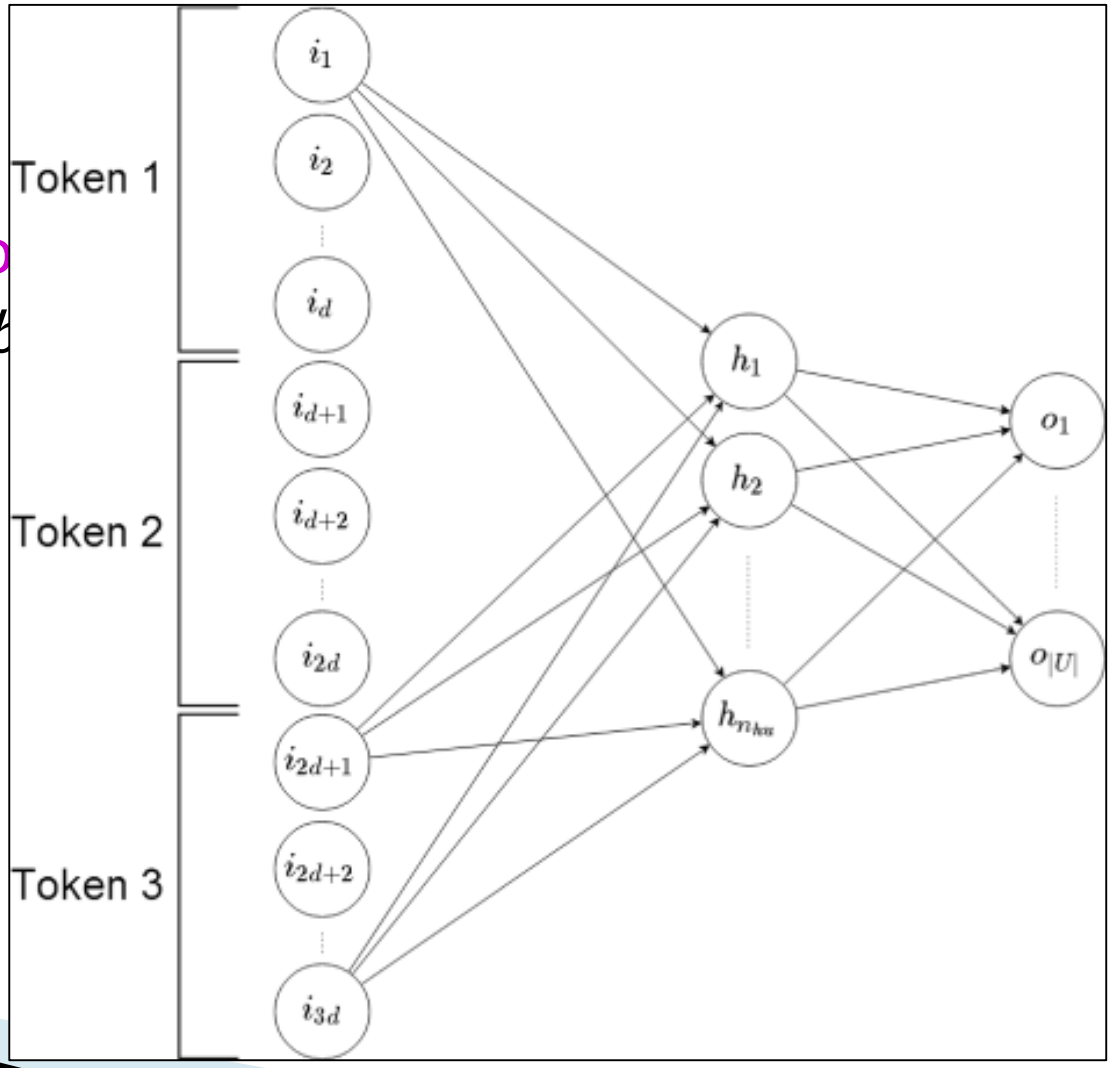
Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o português (Fonseca et al., 2015)
    - Entrada: *word embeddings* (em vez das palavras em si)
    - Saída: escore para cada tag

Palavras são mapeadas em seus vetores
→ maior generalidade, agrupamento categorial, facilidade em lidar com palavras desconhecidas

Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o po
    - Entrada: *word emb*
    - Saída: escore para



Rede faz a classificação

Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o português (Fonseca et al., 2015)
    - Atributos
      - *Word embeddings* originais
      - Capitalização das palavras (tudo em minúscula, primeira letra maiúscula, outras combinações, etc.)
      - Prefixos e sufixos das palavras (de 1 a 5 letras)
    - Janela de 5 palavras

Tagging: aprendizado de máquina

- *Redes neurais*
  - Muitas variações
  - Exemplo para o português (Fonseca et al., 2015)
    - Resultados do estado da arte sobre o córpus MAC-MORPHO
      - Melhor resultado: 97,57% de acurácia

Humanos e máquinas

- Palavras/morfologia e léxico mental
  - Nem a listagem exaustiva, nem todas as regras de flexão/derivação
  - Há indícios de que humanos armazenam em seu léxico mental os lemas das palavras e também algumas formas plenas
    - Stanners et al. (1979): são mantidas separadamente as palavras *happy* e *happiness*, mas somente o verbo *to pour* (derramar), sem suas flexões
- Morfossintaxe
  - Experimentos mostram que humanos discordam em 3-4% das tags
    - Melhores taggers com 97%, normalmente com problemas justamente onde os humanos discordam
  - Voutilainen (1995) mostra que humanos atingem 100% se se permite que discutam as tags com problemas

Atenção

- Estamos falando de língua geral até agora, mas há recursos e processos específicos para tarefas e gêneros e domínios textuais de interesse
  - Há de tudo!

Quem lembra dessas vendas que tinham de tudo? ☺

VOC

- Vocabulário Ortográfico Comum da Língua Portuguesa
  - Trabalho envolvendo vários países

Léxico de sentimentos: LIWC

%	
1	funct
2	pronoun
3	ppron
4	i
5	we
6	you
7	shehe
8	they
9	ipron
10	article
11	verb
12	auxverb
13	past
14	present
15	future
16	adverb
17	preps
18	conj
19	negate
20	quant
21	number
22	swear
121	social
122	family
123	friend
124	humans
125	affect
126	posemo
127	negemo
128	anx
129	anger
130	sad

abafar	125	127	129
abafara	125	127	129
abafaram	125	127	129
abafaras	125	127	129
abafardes	125	127	129
abafarei	125	127	129
abafareis	125	127	129
abafarem	125	127	129
abafaremos	125	127	129
abafares	125	127	129
abafaria	125	127	129
abafariam	125	127	129
abafarias	125	127	129
abafarmos	125	127	129
abafará	125	127	129
abafarás	125	127	129
abafarão	125	127	129
abafaríamos	125	127	129
abafaríeis	125	127	129
abafas	125	127	129
abafasse	125	127	129
abafassem	125	127	129
abafasses	125	127	129
abafaste	125	127	129
abafastes	125	127	129
abafava	125	127	129
abafavam	125	127	129
abafavas	125	127	129

Propriedades psicolinguísticas

PortLEX

Login | Register

Articles

Psycholinguistic Properties of Brazilian Portuguese

Abstract

Psycholinguistic properties of words have been used in various approaches to Natural Language Processing tasks, such as text simplification and readability assessment. Most of these properties are subjective, involving costly and time-consuming surveys to be gathered. Recent approaches use the limited datasets of psycholinguistic properties to extend them automatically to large lexicons. However, some of the resources used by such approaches are not available to most languages. This study presents a method to infer psycholinguistic properties for Brazilian Portuguese (BP) using regressors built with a light set of features usually available for less resourced languages: word length, frequency lists, lexical databases composed of school dictionaries and word embedding models. The correlations between the properties inferred are close to those obtained by related works. The resulting resource contains 26,874 words in BP annotated with concreteness, age of acquisition, imageability and subjective frequency.

Related Publications

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido Junior, A., Paetzold, G. H., Aluísio, S. M. (2017). A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. In Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017. Springer.

Contact

Leandro Borges dos Santos, ICMC-NILC, University of São Paulo, e-mail: leandrobs@usp.br or sborgesleandro@gmail.com;

Sandra Aluisio, ICMC-NILC, University of São Paulo, Brazil, e-mail: sandra@icmc.usp.br.

Download

Lexical database

Source code

Recursos e ferramentas “bio”

- Tagging e parsing para textos biomédicos
  - https://stanfordnlp.github.io/stanza/biomed.html
- Ontologia SMASH
  - https://bioportal.bioontology.org/ontologies/SMASH

Dicionários e córpus históricos

- http://www.nilc.icmc.usp.br/nilc/projects/hpc/

Historical Portuguese Corpora

About Historical Portuguese Corpora (HPC)

HPC is a sub-project of the Historical Dictionary of Brazilian Portuguese project, which is funded by CNPq, Brazil. In the HPC project tools and resources for manipulation of historical corpora and management of historical dictionaries are developed. The tools and resources were released under public domain.

About Historical Dictionary of Brazilian Portuguese (HDBP)

The Historical Dictionary of Brazilian Portuguese (HDBP), the first of its kind, is based on a corpus of Brazilian Portuguese texts from the sixteenth through the eighteenth centuries (including some texts from the beginning of the nineteenth century). The HDBP is a five-year project, which started in 2006. This project has participants from various regions of Brazil and Portugal, including linguists and computer scientists from 11 universities. Private resources from HDPB project can be found here (access restricted to HDBP members).

Redes sociais

- Desafios para léxico, morfologia e morfossintaxe
  - Marcas de informalidade
  - Erros, norma culta da língua ignorada
  - Jargões e expressões próprios
  - Especificidades do meio
  - Às vezes, quase um “dialeto” próprio
  - Mensagens ininteligíveis, em certos casos

- Tweets, por exemplo

Tarefas da semana

- Leitura da semana
  - Gonçalves, M.; Coheur, L.; Baptista, J.; Mineiro, A. (2020). Avaliação de recursos computacionais para o português. LinguaMÁTICA, Vol. 12, N. 2, pp. 51-68.
    - No e-Disciplinas
- Provinha 11