Metodologia da pesquisa em psicolinguística: desenvolvimento de uma ferramenta para a geração automática de pseudoverbos

Research methods in psycholinguistics: developing a tool for the automatic generation of pseudo-verbs

Mailce Mota
Universidade Federal de Santa Catarina/CNPq – Florianópolis – Santa Catarina – Brasil
Natália Resende
Universidade Federal de Santa Catarina – Florianópolis – Santa Catarina – Brasil

Resumo: O uso de pseudopalavras tem beneficiado estudos psicolinguísticos dedicados à aquisição e processamento da linguagem em seus diversos aspectos, incluindo o processamento da leitura e escrita. No entanto, em uma grande parte dos estudos que fazem uso de pseudopalavras, a combinação de fonemas em uma sequência geralmente é realizada com base na intuição dos pesquisadores, sem levar em conta as ocorrências estruturais mais frequentes na língua. Nesse sentido, o presente estudo visa contribuir para o desenvolvimento de procedimentos que possam auxiliar pesquisas que utilizam pseudopalavras em seus designs. Neste artigo, apresentamos uma ferramenta computacional desenvolvida com o objetivo específico de gerar pseudoverbos. O processo de criação da ferramenta envolveu a análise dos padrões estruturais dos 500 verbos mais frequentes do português do Brasil, empregando técnicas do PLN e mineração de dados para coletar e extrair padrões. O presente estudo contou com uma abordagem não supervisionada de aprendizado de máquina, uma vez que utilizou um algorítmo de agrupamento de dados (*cluster)* para extração dos padrões estruturais dos verbos.

Palavras-chave: Pseudopalavras; Pseudoverbos; Mineração de dados; Ferramenta computacional

Abstract: The use of pseudo-words has benefitted psycholinguistic studies dedicated to the different aspects of the acquisition and processing of language, including reading and writing. However, in most of the studies that make use of pseudowords, the combination of graphemes in a string is often based on the intuition of researchers, who generally do not take the frequency of structural occurrences of the language into consideration. The present study aims at contributing to the development of procedures to aid research which makes use of pseudowords. In this paper, we present a computational tool developed for the generation of pseudoverbs. The process of development of this tool involved the analysis of the structural patterns of the 500 most frequent verbs in Brazilian Portuguese. This analysis was carried out with the use of NLP techniques and data mining in order to collect and extract patterns. The present study adopted an unsupervised machine learning approach since it used an algorithm of clustering for the automatic generation of pseudo-verbs.

Keywords: Pseudo-words; Pseudo-verbs; Data mining; Computational tool

Introdução

Uma das questões mais controversas da ciência cognitiva diz respeito à representação e ao processamento da linguagem (STILLINGS et al., 1987). No âmbito do paradigma simbólico, os modelos de via dual (CHOMSKY, 1995; PINKER, 1999, ULLMAN, 1997; ULLMAN, 2001; ULLMAN, 2007) postulam a existência de duas capacidades para a representação e processamento da linguagem: um léxico mental e uma gramática mental. Por outro lado, modelos de via única, em consonância com o paradigma conexionista, postulam que o processamento da linguagem se realiza através de um único mecanismo associativo (BATES e WULFECK, 1989; BATES e WULFECK, 1991; BATES e MacWHINNEY, 1989; MacWHINNEY, 1997; MacWHINNEY, 2002, 2001,

2005; CHATER e MANNING, 2006). Ao longo das últimas décadas, estudos ancorados nos modelos de via dual têm implementado métodos de pesquisa que buscam entender a natureza da dualidade no processamento da linguagem, tanto na língua materna quanto na segunda língua (por exemplo, PARADIS, 2009; PINKER, 1991; PINKER e PRINCE, 1991; PINKER, 1994; ULLMAN, 2001; ULLMAN, 2005, para citar apenas alguns). Entre os métodos adotados, destaca-se o uso de pseudopalavras (pseudoverbos, pseudosubstantivos, pseudoadjetivos). Pseudopalavras podem ser definidas como uma sequência de grafemas que apresentam as regras fonotáticas das palavras naturais de uma língua, mas sem qualquer significado correspondente (GATHERCOLE et al., 1999; VITEVITCH e LUCE, 1999).

As pseudopalavras têm sido utilizadas em estudos psicolínguísticos e neurocognitivos dedicados à avaliação do processamento sintático, fonológico e semântico bem como das habilidades de compreensão leitora e produção escrita. Devido à falta do componente semântico associado, o emprego de pseudopalavras em estudos sobre os processos cognitivos relacionados à linguagem permite avaliar a habilidade de falantes típicos e atípicos em tarefas que exigem automaticidade na operação de regras de natureza computacional. Nesse sentido, pseudopalavras podem fornecer pistas para o entendimento de questões sobre a representação e processamento da linguagem, tais como a separação entre léxico e gramática proposta por modelos simbólicos de via dual (por exemplo, PRADO e ULLMAN, 2009). Estudos sobre a neuroanatomia da linguagem também utilizam tarefas com pseudopalavras para verificar se há diferentes padrões de ativação cerebral na execução destas tarefas em comparação com ativações resultantes do desempenho em tarefas com palavras reais (GROSS et al., 1997, PENKE et al., 1997). Estudos sobre a aquisição fonológica também utilizam pseudopalavras na avaliação do processamento da memória procedural em pacientes com déficit fonológico (TYLER et al., 2002). Na área da compreensão leitora, KAST et al., (2011) investigaram os substratos neurais envolvidos no processamento audiovisual de palavras e pseudopalavras em pacientes disléxicos e não disléxicos. Do mesmo modo, pseudopalavras vêm sendo frequentemente utilizadas em estudos com pacientes afásicos para avaliação de suas habilidades cognitivas em leitura (ROCH e JARROLD, 2012; DE BREE et al., 2007).

Uma característica relevante das pseudopalavras é a sua semelhança estrutural e fonotática com as palavras naturais de uma língua. Entretanto, geralmente, esses itens são construídos com base na intuição que os pesquisadores têm sobre a língua. Por esse motivo, a construção intuitiva de pseudopalavras é um procedimento que, via de regra, não leva em conta a sensibilidade dos falantes em relação à frequência das estruturas das palavras. Ellis (2002) argumenta que falantes de uma língua são sensíveis aos efeitos de frequência de estruturas em todos os níveis de análise (do fonológico ao discursivo). Partimos dessa idéia para sugerir que o conhecimento das ocorrências estruturais mais frequentes das palavras naturais de uma língua seja utilizado na construção de sequências que possam ser prontamente reconhecidas pelos falantes como estruturas pronunciáveis.

À luz dessas considerações, o presente estudo visa investigar as ocorrências estruturais fonológicas mais frequentes em verbos regulares e irregulares do Português do Brasil (PB) para que os padrões encontrados possam ser empregados na construção de uma ferramenta computacional capaz de gerar pseudoverbos automaticamente. Por ocorrências estruturais mais frequentes entendem-se os padrões silábicos que mais aparecem na estrutura dos verbos no PB de primeira, segunda e terceira conjugações. Além dos padrões silábicos, avaliamos também os fonemas que mais aparecem na estrutura silábica em posição de onset, núcleo e coda, segundo definido por Selkirk (1982). Para tanto, utilizamos o arcabouço teórico e metodológico da Linguística de *Corpus* e da Inteligência Artificial através do emprego de técnicas de PLN e mineração de dados na coleta e análise dos dados.

O presente artigo está organizado da seguinte forma: A primeira seção apresenta o molde silábico do Português do Brasil; a seção 2 descreve a metodologia empregada e a seção 3 discute os resultados obtidos. Na seção 4, apresentamos a arquitetura da ferramenta. A seção 5 traz as considerações finais.

1 O molde silábico no português do Brasil

O presente estudo adota a estrutura da sílaba como apresentada por Selkirk (1982). De acordo com a autora, a estrutura silábica divide-se em duas partes: um ataque (ou onset) e uma rima. A rima compreende um núcleo e uma coda. O onset é a posição consonatal pré-vocálica e pode ser simples (contendo apenas uma consoante) ou complexo (contendo duas consoantes). O núcleo é composto por vogais e pode conter, também, semivogais.[1] A coda representa a posição consonantal pós-vocálica e pode ser simples (apenas uma consoante) ou complexa (duas consoantes). No PB, é possível ter onset complexo (***pr**ato*) e coda complexa (*per**s**picácia*). O núcleo pode conter uma vogal oral (***a**-mo**r***) ou uma vogal e uma semivogal (***doi**s*). Na estrutura silábica, apenas o núcleo é obrigatório, pois onset e coda são posições de preenchimento opcional.

[1] Há autores que consideram a semivogal na coda silábica.

No PB, qualquer consoante pode ocupar a posição de onset, exceto (/ɾ/, /ʎ/, /ɲ/).[2] Em onset complexo, a segunda consoante será sempre preenchida por uma consoante líquida (/ɾ/, /l/). Na coda, temos sempre os arquifonemas /N/, /S/, /R/. Em relação ao arquifonema /N/, alguns autores argumentam a favor das vogais nasais. Assim, palavras como *anterior* não teriam o fonema /N/ em coda, mas uma vogal nasal /ã/ uma vez que essas são responsáveis por distinção de significado e, por isso, são consideradadas fonemas (como, por exemplo, nos pares mínimos *cata* /'Kata/ e *canta* /'Kãta/).

A Figura 1 é a representação arbórea da sílaba do português, fundamentada na proposta de Selkirk (1982).

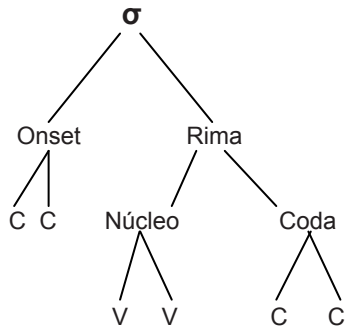

Figura 1 – Representação arbórea da sílaba do português, com base em Selkirk (1982)

CCVVCC constitui uma sílaba de projeção máxima (SPM), uma vez que mostra todas as posições passíveis de preenchimento dentro da estrutura silábica. O presente estudo baseou-se nessa representação para decompor os verbos mais frequentes do PB a fim de responder às seguintes perguntas:

a) Quais padrões silábicos são mais comuns em primeira, segunda e terceira conjugações nas sílabas iniciais, mediais e finais?
b) Quais padrões silábicos co-ocorrem com maior frequência nos verbos da primeira, segunda e terceira conjugações?
c) Quais são os fonemas mais frequentes em posição de onset, núcleo e coda nas sílabas iniciais, mediais e finais dos verbos da primeira, segunda e terceira conjugações?

A próxima seção apresenta a metodologia empregada na anotação manual dos verbos bem como as técnicas usadas para extrair os padrões estruturais do *corpus*.

2 Metodologia

A metodologia empregada no presente estudo utilizou técnicas provenientes da Linguística de *Corpus* (LC) e da Inteligência Artificial (IA). Para tanto, a estrutura fonológica dos verbos da primeira, segunda e terceira conjugações foi analisada a partir dos 500 verbos mais frequentes do português brasileiro, os quais foram extraídos do *corpus* CETEMFOLHA compilado pelo NILC – Núcleo interestadual de Linguística Computacional – disponível no site da Linguateca.[3] O CETEMFOLHA é um *corpus* de 24 milhões de palavras contendo textos jornalísticos do ano de 1994. Devido a sua extensão, entendemos ser este um *corpus* representativo do português do Brasil, apesar de conter apenas o gênero jornalístico.

Os verbos foram selecionados segundo sua frequência de lema no *corpus*. Esse procedimento foi importante para a seleção de verbos com estrutura fonológica de alta frequência no PB. Após a extração dos dados, a segunda etapa do processo envolveu a anotação manual dos verbos em relação às propriedades fonológicas, as quais foram transcritas em um arquivo de treino[4] para futura mineração dos dados.

No arquivo de treino, cada verbo corresponde a uma instância no arquivo. Cada instância inclui os 23 atributos listados na Tabela 1, a seguir, e seus respectivos valores.

Conforme Tabela 1, de C1 a C17 há três sequências de símbolos CCVVCC. Cada sequência equivale a uma sílaba de projeção máxima (SPM). Assim, de 6 a 12 temos a SPM inicial, de 13 a 18 SPM medial e de 19 a 22 a SPM final. Cada uma das posições dentro da SPM foi preenchida com os valores (fonemas) das sílabas iniciais, mediais e finais dos verbos mais frequentes do PB. Essas informações foram importantes para a visualização dos fonemas mais frequentes em cada posição (onset, núcleo e coda) da estrutura das sílabas iniciais, mediais e finais dos verbos.

Quatro critérios principais foram estabelecidos para a anotação no arquivo de treino:

a) Na SPM, cada atributo C foi preenchido por uma consoante e cada V por uma vogal ou semivogal. As semivogais foram anotadas como parte do núcleo da estrutura silábica[5];
b) Os verbos monossílabos como (por exemplo, *ser*, *ter*) foram encaixados na última SPM (de 19 a 22) para que sua marca de infinitivo fosse alinhada com a marca do infinitivo dos verbos dissílabos e trissílabos;
c) Para cada verbo, uma nova linha foi escrita no arquivo de treino com os valores dos 22 atributos separados por vírgula, usando o formato padrão CSV (*comma separated value*). Caso o verbo não

[2] Exceto nas palavras *lhama* e *nhame.*
[3] <www.linguateca.pt>.
[4] Arquivo de treino é o arquivo em formato ARFF, contendo todas as instâncias de dados, seus atributos e valores.
[5] A anotação dos verbos no arquivo de treino foi realizada com base na transcrição fonológica dos verbos no infinitivo.

Tabela 1 – Propriedades dos verbos anotadas no arquivo de treino, definição e valores.

		Atributos	Definição	Valor
	1	conj	conjugação verbal: primeira, segunda, terceira ou quarta	P,S,T
	2	numsilabas	numero de sílabas	U,D,T
	3	padrãoposicaoum	padrão da sílaba inicial do verbo	V,VV,VC,CV,VCC, CVC,CVV,CVCC, CCVV,CCV,CCVC, CCVCC,CCVVC
	4	padrãoposicaodois	padrão da sílaba medial do verbo	
	5	padrãoposicaotres	padrão da sílaba final do verbo	
Sílaba inicial	6	C1	primeira consoante da sílaba inicial (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	7	C2	segunda consoante da sílaba inicial (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	8	V3	primeira vogal da sílaba inicial	a,A,ã,e,E,i,I,o,õ,u,U,L
	9	V4	segunda vogal da sílaba inicial (caso haja)	a,A,ã,e,E,i,I,o,õ,u,U,L
	10	C5	primeira consoante em posição de coda (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	11	C6	segunda consoante em posição de coda (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
Sílaba medial	12	C7	primeira consoante da sílaba medial (caso haja Onset)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	13	C8	segunda consoante da sílaba medial (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	14	V9	primeira vogal da sílaba inicial (núcleo da sílaba)	a,A,ã,e,E,i,I,o,õ,u,U,L
	15	V10	segunda vogal da sílaba inicial (caso haja)	a,A,ã,e,E,i,I,o,õ,u,U,L
	16	C11	primeira consoante em posição de coda medial (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	17	C12	segunda consoante em posição de coda medial (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
Sílaba final	18	C13	primeira consoante da sílaba final (caso haja Onset)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	19	C14	segunda consoante da sílaba final (caso haja)	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~
	20	V15	primeira vogal da sílaba final (núcleo da sílaba)	a,A,ã,e,E,i,I,o,õ,u,U,L
	21	V16	segunda vogal da sílaba final (caso haja)	a,A,ã,e,E,i,I,o,õ,u,U,L
	22	C17	consoante em posição de coda final	A,b,B,c,ç,d,D,f,g,G,H, j,l,L,m,n,p,P,q,Q,r,R,s,S, t,v,x,X,z,Z,~

apresentasse um determinado valor (por exemplo, verbos sem a posição de coda da sílaba inicial preenchida), o espaço a ser preenchido pelo valor do atributo foi ocupado pelo símbolo "?";

d) Os fonemas representados por mais de um grafema, como *CH, SS, NH, LH, RR, HA,HE*, e casos de sílabas contendo vogais nasalizadas pelo arquifonema /N/ receberam um símbolo diferenciado. Esse procedimento foi adotado como uma forma de evitar distorções na análise dos dados, pois um verbo como *chamar* não poderia ser anotado como tendo na sílaba inicial o padrão silábico CCV, uma vez que os grafemas *CH* representam apenas 1 fonema (/ʃ/). Do mesmo modo, verbos contendo vogais nasalizadas por uma consoante nasal /N/ em coda foram consideradas vogais nasais. Os demais fonemas foram anotados com o grafema correspondente em letras minúsculas. Os símbolos criados são apresentados na Tabela 2.

Tabela 2 – Códigos de anotação de fonemas

Grafemas	Fonemas	Códigos
NH	/ɲ/	~
LH	/ʎ/	H (maiúsculo)
AN	/ã/	ã
EN	/ẽ/	E (maiúsculo)
IN	/ĩ/	I (maiúsculo)
ON	/õ/	õ
UN	/ũ/	U (maiúsculo)
CH	/ʃ/	X (maiúsculo)
S,SS,X,Ç,C	/S/	S (maiúsculo)
RR	/ʁ/	R (maiúsculo)
GU	/g/	G (maiúsculo)
QU	/K/	Q (maiúsculo)
S,X	/z/	Z (maiúsculo)
J,G	/ʒ/	j
L em coda	/w/	L (maiúsculo)

A anotação dos dados provenientes dos 500 verbos mais frequentes do PB permitiu determinar i) os fonemas mais frequentes em posição de onset, núcleo e coda nas sílabas iniciais, mediais e finais dos verbos das três conjugações; ii) os padrãos silábicos que co-ocorrem com mais freqüência nos verbos de primeira, segunda e terceira conjugações e iii) a quantidade de sílabas mais frequentes nas três conjugações. Portanto, as propriedades acima descritas oferecem várias informações estruturais e fonológicas dos verbos naturais do PB, as quais foram coletadas a fim de serem reproduzidas ao gerar pseudoverbos automaticamente.

2.1 *Mineração dos dados com WEKA*

WEKA[6] é um *workbench* que contém uma coleção dos mais recentes algorítmos de aprendizado de máquina para mineração de dados. Através do WEKA, é possível programar algoritmos de aprendizado em banco de dados. Além disso, é possível determinar o algoritmo mais adequado para o padrão de banco de dados que se tem em mãos. Esse *workbench* oferece ainda métodos estatísticos para avaliação de qualquer problema de mineração de dados, tais como medida F, Precisão e Recall.

Há várias maneiras de se usar o WEKA. O presente estudo adotou uma abordagem não supervionada de classificação ao escolher a técnica de *clustering*. O algoritmo utilizado foi o *simplekmeans*, que possibilita avaliar de modo rápido e objetivo um grande conjunto de dados ao agrupar elementos que compartilham características similares em *clusters*. Assim, ao empregar a técnica de *clustering*, o conjunto de dados é dividido em grupos de objetos similares. Cada grupo, denominado *cluster*, é formado por elementos que são semelhantes entre si e diferentes de elementos de outros grupos. No entanto, representar elementos em poucos *clusters* significa perder informações sobre o banco de dados, pois quanto menor o número de *clusters* maior a chance de o algoritmo descartar certos detalhes importantes para a avaliação, pois estará simplificando e reduzindo informações para encaixá-las dentro de poucos agrupamentos. No presente estudo, o conjunto de dados foi divido em 5 *clusters* para ampliar a quantidade de informação relativamente às três conjugações verbais do PB. Os padrões encontrados pelo algoritmo são apresentados nas Tabelas 3, 4 e 5 da seção de resultados (seção 3).

3 Resultados

Os resultados apresentados nessa seção foram obtidos através da mineração dos dados. Os padrões silábicos apresentados na Tabela 3 equivalem aos que aparecem com uma frequência superior a 5% da frequência dos verbos de cada conjugação contendo aquela posição de sílaba dentro da palavra. Por exemplo, os padrões das sílabas iniciais considerados frequentes na primeira conjugação (-ar) são os padrões que apareceram com uma ocorrência superior a 5% nos 290 verbos dessa conjugação. O mesmo procedimento foi aplicado a todos os padrões silábicos de cada posição dentro da palavra para cada conjugação verbal.

Tabela 3 – Padrões silábicos vs. fonemas em relação aos verbos de primeira, segunda e terceira conjugações

Primeira conjugação		
Posição das sílabas na estrutura do verbo	**Padrão de sílaba**	**Grafemas dos fonemas mais frequentes em cada posição**
Sílaba inicial	C	C,P,T,D,V
	V	A,E,I,O,On
		ou
	V	A,E,O,I
Sílaba medial	C	C,P,T,V,M
	V	E,I,O,U
		ou
	C	C,P,T,V,M
	V	E,I,O,U
	C	R,S
Sílaba Final	C	T,R,C
	V	A
	C	R

continua

[6] <http://www.cs.waikato.ac.nz/~ml/weka/>.

Tabela 3 (continuação)

Segunda conjugação		
Posição da sílaba dentro do verbo	**Padrão silábico**	**Grafemas dos fonemas mais frequentes em cada posição**
Sílaba inicial	C	C,P,D,F,R,V
	V	A,E,O
Sílaba medial	C	C,D,F,M,P,T,V
	V	A, I, E,O,En
		ou
	V	E,O,En
	C	R,S
Sílaba final	C	C,D,V,T
	V	E
	C	R
Terceira conjugação		
Posição da silaba	**Padrão de silaba**	**Grafemas dos fonemas mais frequentes em cada posição**
Sílaba inicial	V	Ôn,In,A,E,O
		ou
	C	C,D,P,S
	V	A,E,O
Sílaba medial	C	C,M,F,T,V
	V	E,I,U
	C	R,S
		ou
	C	C,M,F,T,V
	V	A,E,I,O,U
Sílaba final	C	B,T,D,R,J,V
	V	I
	C	R

Da mesma forma, os fonemas considerados frequentes em cada posição dentro do padrão silábico são aqueles que apresentaram uma frequência de ocorrência superior a 5% da frequência dos verbos de cada conjugação. Por exemplo, no núcleo da sílaba inicial da primeira conjugação, os fonemas /a/, /e/ e /o/ aparecem acima de 14 vezes (5% de 290), por isso foram considerados fonemas frequentes naquela posição dentro da estrutura da sílaba. Assim, todos os fonemas e padrões silábicos com ocorrência inferior a 5% nos verbos de primeira, segunda e terceira conjugações foram descartados para que o sistema possa lidar apenas com padrões silábicos e fonemas que apresentam frequência significativa.

3.1 *Discussão dos resultados*

Os padrões estruturais encontrados na análise dos 500 verbos que constituem o *corpus* revelam que os verbos trissílabos são mais frequentes que os dissílabos e monossílabos, como pode ser visto na Tabela 4.

Tabela 4 – Quantidade de silabas nos verbos

Nº de sílabas	Quantidade	Percentagem
Monossílabos	9	2%
Dissílabos	208	42%
Trissílabos	283	56%

Os verbos de primeira conjugação são mais frequentes que os verbos de segunda conjugação e os verbos de segunda são mais frequentes que os verbos de terceira conjugação, como pode ser visto na Tabela 5.

Tabela 5 – Quantidade de verbos de primeira, segunda e terceira conjugações coletados do *corpus*

Conjugação Verbal	Quantidade	Percentagem no corpus
1ª	290	58%
2ª	120	24%
3ª	90	18%

A análise mostra também que os padrões silábicos mais frequentes nos verbos são CV ou CVC. Em relação à estrutura fonológica, os resultados indicam diferenças de sequência de fonemas dentro da estrutura silábica dos verbos de primeira, segunda e terceira conjugações. No entanto, em posição de onset, as consoantes plosivas e fricativas são as mais frequentes nas sílabas iniciais, mediais e finais em todas as conjugações. Em posição de coda, o arquifonema /R/ é mais frequente que /l/ ou /S/ em qualquer sílaba dentro da estrutura dos verbos de qualquer conjugação. A análise mostra tambem que sílabas contendo coda não são comuns em início de palavra. De fato, sílabas contendo posição de coda em sua estrutura são as mediais e, principalmente, as finais, uma vez que os verbos foram extraídos em sua forma no infinitivo (contendo a marca de infinitivo /R/ em coda final). Os dígrafos representados pelos fonemas /ʎ/, /ɲ/ e /ʃ/ não são frequentes em posição de onset, por isso foram descartados. Sílabas contendo semivogal não são frequentes, por isso não foram consideradas como padrão silábico da estrutura fonológica dos verbos.

Surpreendentemente, nos verbos de primeira e terceira conjugações, sílabas contendo apenas o núcleo (com ou sem vogal nasalizada) são mais frequentes na sílaba inicial do que nos verbos de segunda conjugação. Observa-se que na posição de núcleo, certas vogais são mais frequentes que outras conforme o posicionamento da sílaba na estrutura do verbo.

4 Arquitetura da ferramenta

Após a extração de padrões, a informação fornecida foi aplicada na contrução de uma ferramenta computacional que gera pseudoverbos automaticamente.

A ferramenta foi construída usando a linguagem de programação *Python* e é capaz de gerar pseudoverbos de primeira, segunda e terceira conjugações nos moldes de um verbo natural do PB em sua forma infinitiva. Ao gerá-los, a ferramenta combina os fonemas representados por seus grafemas de forma a agrupá-los em padrões silábicos mais frequentes e em situação de co-ocorrência conforme padrões extraídos dos 500 verbos naturais para cada conjugação verbal através da mineração dos dados.

O usuário tem a opção de obter verbos dissílabos e trissílabos de primeira, segunda e terceira conjugações. Ao ser acionado, o sistema pede que o usuário entre com a conjugação verbal desejada. O programa lista 3 verbos dissílabos e 3 verbos trissílabos da conjugação verbal escolhida, oferecendo opções ao usuário. Nos casos dos verbos dissílabos, para cada conjugação verbal, o sistema combina apenas os padrões silábicos das silabas iniciais e finais, seguindo os padrões silábicos encontrados em situação de co-ocorrência. Em verbos trissílabos, o sistema gera peseudoverbos combinando as sílabas iniciais, mediais e finais, conforme a situação de co-ocorrência de padrões silábicos para cada conjugação verbal. A Figura 2 mostra um exemplo do output da ferramenta.

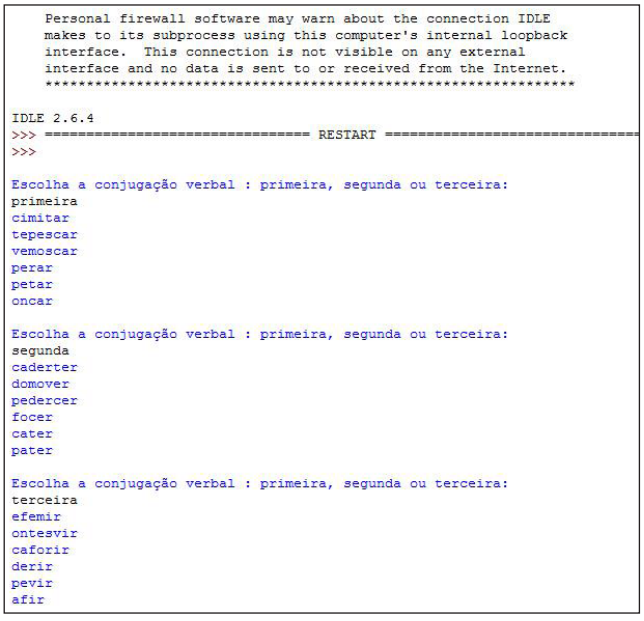

Figura 2 – interface do sistema

5 Considerações finais

O presente trabalho teve como objetivo encontrar padrões estruturais e fonológicos dos 500 verbos mais frequentes no PB a fim de empregar o conhecimento extraído na construção de uma ferramenta computacional para geração automática de pseudoverbos. Para tanto, a abordagem metodológica utilizada exigiu que os verbos fossem manualmente anotados para codificação das informações com relação a sua estrutura fonológica. Um conjunto de 23 atributos foi criado para anotação das 500 instâncias verbais as quais foram mineiradas usando um algoritmo de agrupamento de dados (*cluster*) disponível na ferramenta WEKA. A metodologia adotada no presente estudo mostrou-se eficaz para a construção de uma ferramenta computacional de geração automática de pseudoverbos que se baseia na frequência de co-ocorrência de padrões silábicos no PB. A ferramenta pode auxiliar estudos que utilizam pseudopalavras na investigação dos mecanismos cognitivos envolvidos na aquisição e processamento da liguagem.

Referências

BATES, Elisabeth; MACWHINNEY, Brian. Functionalism and the competition model. In: MACWHINNEY, Brian; BATES, Elisabeth. (Eds.). *The crosslinguistic study of sentence processing* Cambridge, UK: Cambridge University Press, 1989. p. 3-73.

BATES, Elisabeth; WULFECK, Bevely. Crosslinguistic studies of aphasia. *Brain and Language*, v. 41, p. 123-148, 1991.

CHATER, Nick; MANNING, Christopher. Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, v. 10, fasc. 7, p. 335-344, 2006.

CHOMSKY, Noam. *The Minimalist Program*. Cambridge, Mass.: MIT Press, 1995.

DE BREE, Elise; JANSE, Esther; VAN DE ZANDE, Anne Marie. Stress assignment in aphasia: Word and non-word reading and non-word repetition. *Brain and Language,* v. 103, fasc. 3, p. 264-275, dez. 2007.

ELLIS, Nick. Frequency effects in language acquisition. Edição especial. *Studies in Second Language Acquisition*, v. 24, p. 143-188, 2002.

GATHERCOLE, S.E.; FRANKISH, C.F.; PICKERING, S; PEAKER, S. Phonotactic Influences on Short-Term Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, v. 25, n. 1, p. 84-95, 1999.

GROSS, Matthias; SAYB, Tessa; KLEINGERSA, Michael; HARALD, Clahsen; MUNTE, Thomas. Human brain potentials to violations in morphologically complex Italian words. *Neuroscience Letter*, v. 241, p. 83-86, 1998.

KAST, Monika; BEZZOLA, Ladina; JÄNCKE, Lutz; MEYER, Martin. Multi- and unisensory decoding of words and nonwords result in differential brain responses in dyslexic and nondyslexic adults. *Brain and Language*, v. 119, p. 136-148, 2011.

MACWHINNEY, Brian. The competition model: The input, the context, and the brain. In: ROBINSON, P. (Ed.). *Cognition and Second Language Instruction.* Cambridge: Cambridge University Press, 2001. p. 69-90.

MACWHINNEY, Brian. New directions in the competition model. In: TOMASELLO, M.; SLOBIN, D. I. (Eds.). *Beyond Nature-Nurture: Essays in honor of Elizabeth Bates*. Mahwah, NJ: Lawrence Erlbaum, 2005. p. 81-110.

NEWMAN, S.; TWIEG, D. Differences in auditory processing of words and pseudowords: An fMRI study. *Human Brain Mapping*, v. 14, p. 39-47, 2001.

PENKE, Martina; WEYERTS, Helga; GROSS, MAtthias; ZANDER, Elke; MUNTE, Thomas; CLAHSEN, Harald. How the brain processes complex words: an event-related potential study of German verb inflections. *Cognitive Brain Research*, v. 6, p. 37-52, 1997.

PINKER, S. Rules of language. *Science*, v. 253, p. 530-535, 1991.

PINKER, S.; PRINCE, A. Regular and irregular morphology and the psychological status of rules of grammar. In: LIMA, Susan; CORRIGAN, Roberta; IVERSON, Gregory. *The reality of linguistic rules*. Amsterdam/Philadelphia: John Benjamins, 1991. p. 321-351.

PINKER, Steven. *Words and rules:* The ingredients of language. New York: Science Masters, 1999.

PRADO, E.; ULLMAN, M. Can imageability help us draw the line between storage and composition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, v. 35, p. 849-866, 2009.

ROCH, Maja; JARROLD, Christopher. A follow-up study on word and non-word reading skills in Down syndrome. *Journal of Communication Disorders*, v. 45, p. 121-128, 2012.

SELKIRK, Elisabeth. The syllable. In: HULST, Harry; SMITH, Van der. *The structure of phonological representations* (part II). Foris: Dordrecht, 1982. p. 337-383.

STILINGS, N.; FEINSTEIN, M.; GARFIELD. J.; RISSLAND, E.; ROSENBAUM, D.; WEISLER, S.; BAKER-WARD, L. *Cognitive science:* an introduction. Cambridge, MA: The MIT Press, 1987.

TAROYAN, Naira; NICOLSON, Roderick. Reading words and pseudowords in dyslexia: ERP and behavioural tests in English-speaking adolescents. *International Journal of Psychophysiology*, v. 74, p. 199-208, 2009.

TYLER, Lorraine; RANDALL, Billi; MARSLEN-WILSON, William. Phonology and neuropsychology of the English past tense. *Neuropsychologia*, v. 40, n. 8, p. 1154-1166, 2002.

ULLMAN, Michael; CORKIN, Suzanne; COPPOLA, Marie; HICKOK, Gregory; GROWDON, John, KOROSHETZ, Walter; PINKER, Steven. A neural dissociation within language: Evidence that the mental dictiorary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience,* v. 9, p. 266-276, 1997.

ULLMAN, Michael. The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research*, v. 30, p. 37-69, 2001.

ULLMAN, Michael. The biocognition of the mental lexicon. In: GASKELL, M. G. (Ed.). *The Oxford Handbook of Psycholinguistics.* Oxford, UK: Oxford University Press, 2007. p. 267-286.

Recebido: 08 de setembro de 2012
Aprovado: 12 de novembro de 2012
Contato: mailce@cce.ufsc.br; natcarol2000@yahoo.com.br