7th. Eicefala

A generic representation for orthographic structure in texts written by children

Adelaide H.P. Silva (UFPR/DELLIN, LIAMF/CNPq);
Fabiano Silva; Wanderlan Carvalho (UFPR/DInf/LIAMF);
Lourenço Chacon (UNESP/Marília)

7th. Eicefala

- **Motivation**

  - spelling is taken as a parameter to evaluate whether an individual is literate or not;

  - children are submitted to an exam that evaluates their performance in spelling twice in Elementary School (3rd and 5th grade);

  - last edition of the National Literary Exam (ANA, 2016): 34% of the children evaluated did not achieve the expected scores for them to be considered literate students.

7th. Eicefala

- **Objectives**

- design a system that generates orthographic forms such as the ones found in texts written by 3rd and 5th grade children, departing from “patterns of errors”;

- offer subsidies for teachers to understand the hypotheses underlying the forms that deviate from standard orthographic ones;

- provide teachers resources to evaluate whether or not the "errors" fit a given grade.

7th. Eicefala

- **First steps**

  - analyse data to classify "errors" and to detect patterns that group them following criteria such as type of "error" and school grade;

  - elaborate a generic representation for orthographic structure in texts written by children:

    - the representation is machine readable and corresponds to an abstraction departing from “real” data

7th. Eicefala

- **Dataset**

  - 168 texts, containing 45561 words;

  - written productions by children from 3rd and 5th grades (Chacon, 2018);

  - texts rewrite narratives for children that the teacher of each class read to the students, with the specific objective of collecting data for a written language database (CHACON, 2018).

7th. Eicefala

- **Theoretical background**

  - based on Chacon, Pezarini (2018)

  - authors claim that literacy process involves transparent correspondence between graphemes and sounds and also opaque correspondence;

  - opaque correspondences are set by conventions that may or may not consider the context of occurrence of the sound;

  - conventions set spelling rules, independent from phonological variation

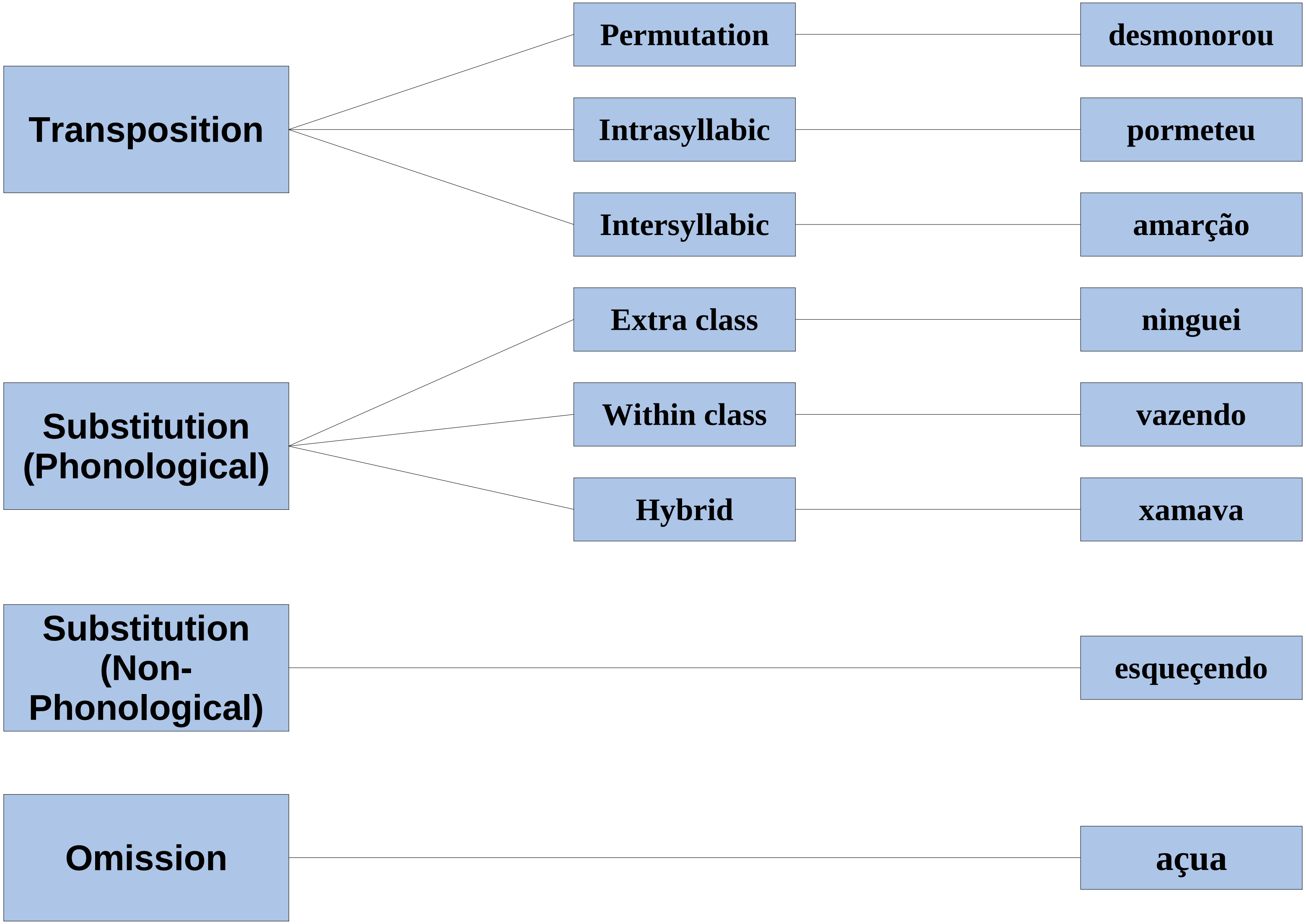
7th. Eicefala

- **Theoretical background**

  - Chacon, Pezarini (2018): there’s a gradiency in the relationship between phonic aspects and orthographic system in Brazilian Portuguese;

  - they conceive that gradiency lies in the distinction between different types of errors, e.g., phonological substitutions can be different, involving or not the same classes of sounds.

“Errors”



Transposition

Permutation

desmonorou

Intrasyllabic

pormeteu

Intersyllabic

amarção

Substitution (Phonological)

Extra class

ninguei

Within class

vazendo

Hybrid

xamava

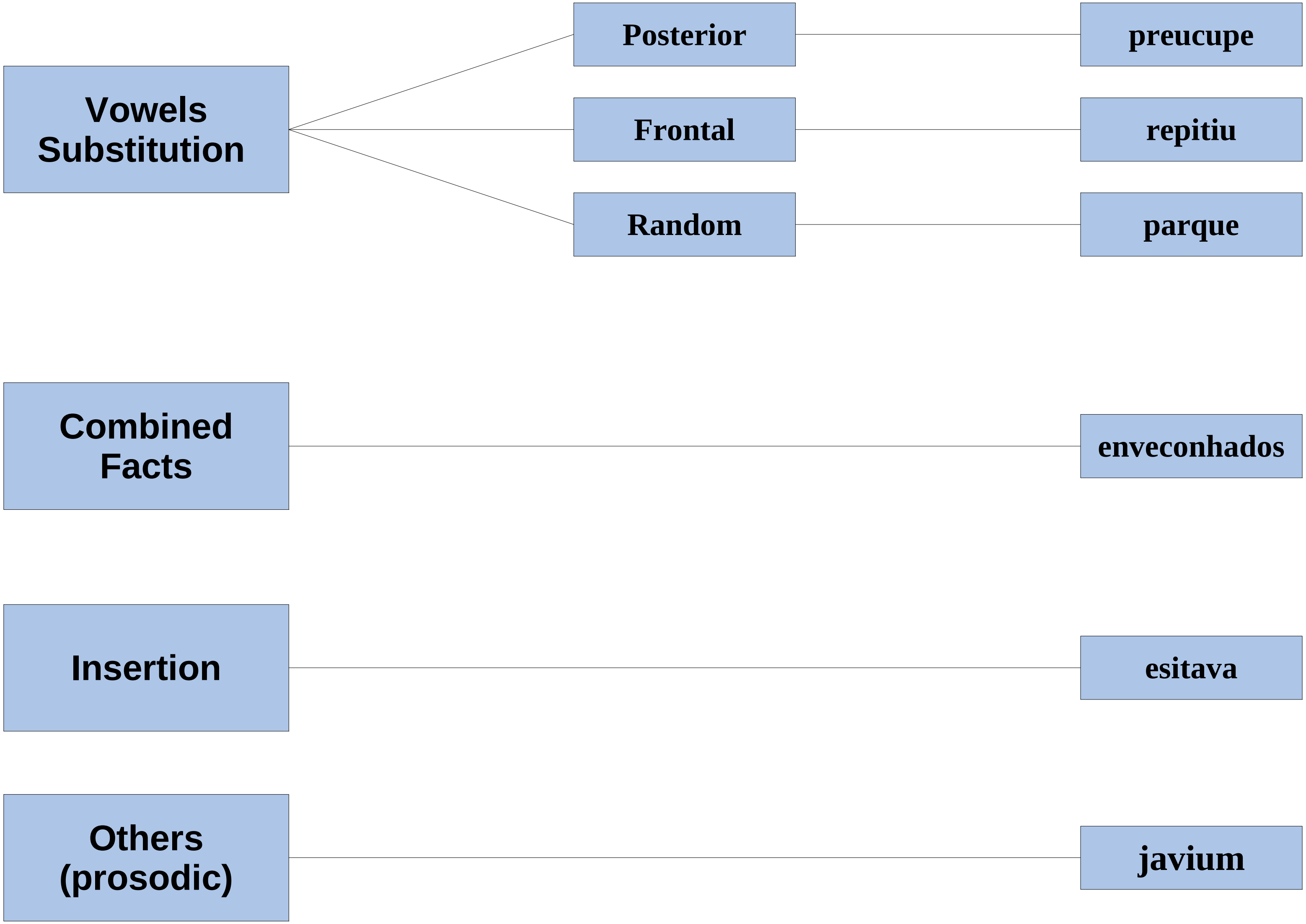
Substitution (Non-Phonological)

esqueçendo

Omission

açua

“Errors” (additional)



Vowels Substitution

Posterior – preucupe

Frontal – repitiu

Random – parque

Combined Facts – enveconhados

Insertion – esitava

Others (prosodic) – javium

7th. Eicefala

- **Data analysis**

  - allowed us to classify the “errors” as exposed;

  - allowed us to propose a set of labels and to manually deal with some data in order to verify whether or not the labels would perform adequatly.

7th. Eicefala

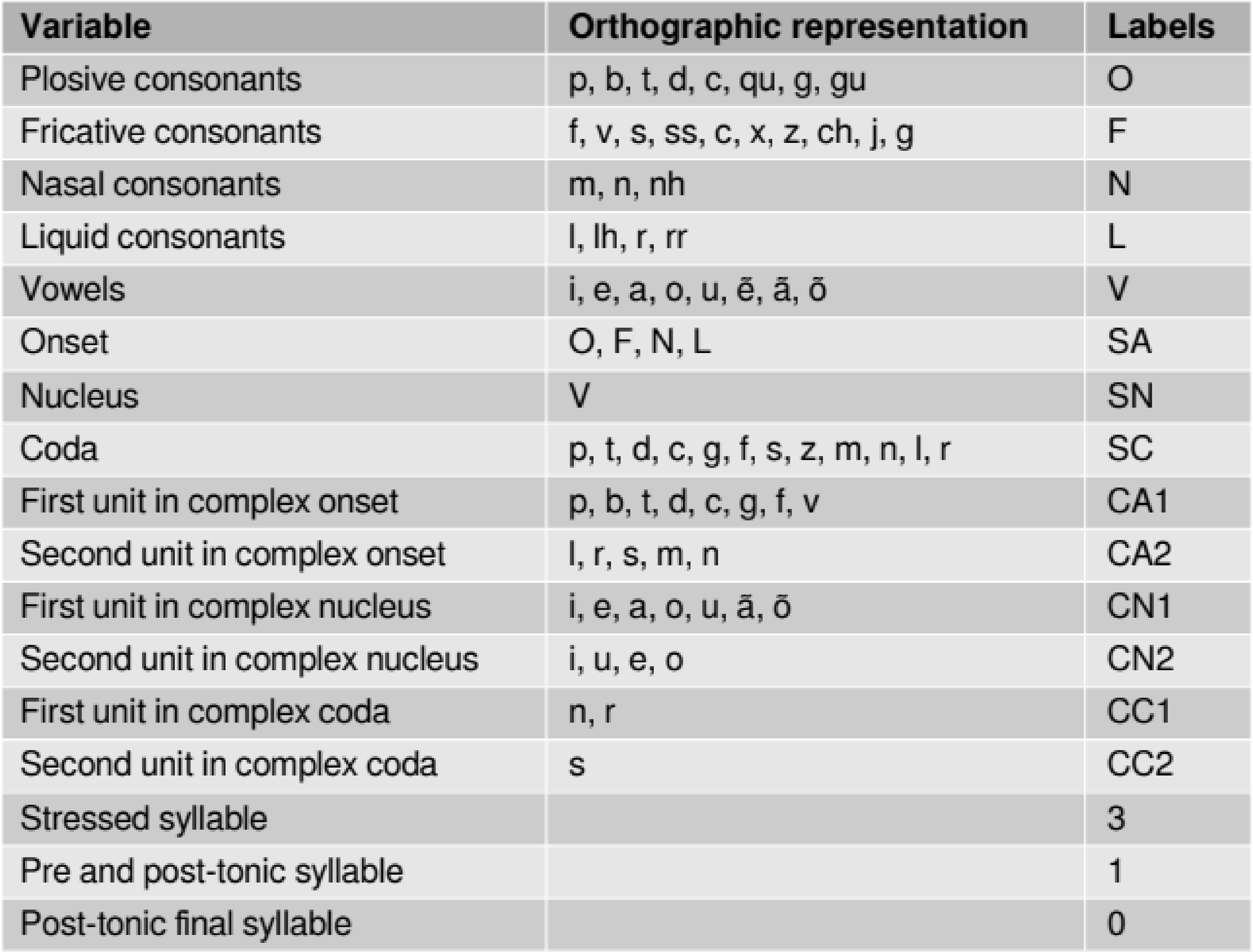
- **Labels**

- crucial for elaborating predictions on “errors”;

  - predictions:

    - related to the graphemes involved in the “errors”;
    - take into account information such as syllable internal structure; syllabic boudaries; primary stress placement; stress degree; consonant class.

Labels for variables



Variable	Orthographic representation	Labels
Plosive consonants	p, b, t, d, c, qu, g, gu	O
Fricative consonants	f, v, s, ss, c, x, z, ch, j, g	F
Nasal consonants	m, n, nh	N
Liquid consonants	l, lh, r, rr	L
Vowels	i, e, a, o, u, ẽ, ã, õ	V
Onset	O, F, N, L	SA
Nucleus	V	SN
Coda	p, t, d, c, g, f, s, z, m, n, l, r	SC
First unit in complex onset	p, b, t, d, c, g, f, v	CA1
Second unit in complex onset	l, r, s, m, n	CA2
First unit in complex nucleus	i, e, a, o, u, ã, õ	CN1
Second unit in complex nucleus	i, u, e, o	CN2
First unit in complex coda	n, r	CC1
Second unit in complex coda	s	CC2
Stressed syllable		3
Pre and post-tonic syllable		1
Post-tonic final syllable		0

(elaborated by the authors)

7th. Eicefala

- **Variables comprise**

  - classes of segments (vowels and consonants) and subsets of consonats based on manner of articulation, as well as subsets of oral and nasal vowels;

7th. Eicefala

- **Variables comprise**

  - the position of each unit within the syllable, considering

    1) vowels to be the only possibile units in syllable nucleus;

    2) subset of consonants occurring in coda is smaller than that in onset;

    3) which units occur in second position of complex syllabic constituents (numerical index consonant to indicating its placement in a complex constituent).

7th. Eicefala

- **Variables comprise**

  - prosodic structure of the word, by assigning stress levels to the syllables (Camara Jr., 1970): 3 for primary stress;1 for pretonic and postonic syllables; 0 for postonic syllables in word-final position; 2 for secondary stress, as in

  (ca)1(fe)2(zi)3(nho)1

7th. Eicefala

- **Variables comprise**

  - labels for different syllable constituents, such as N (nucleus), A(onset) and C (coda);

  - labels for signalizing whether the syllable constituent is a simple (S) or a complex (C) one.

7th. Eicefala

- **Labels allow**

  - capturing and understanding how units relate to each other

  - predicting possible sequences, as well as sequences of units that violate constraints of well-formedness, e.g sequences of graphemes that write sound sequences that do not obey the sonority scale, and also sequences of graphemes that annotate randomized sequences of consonants.

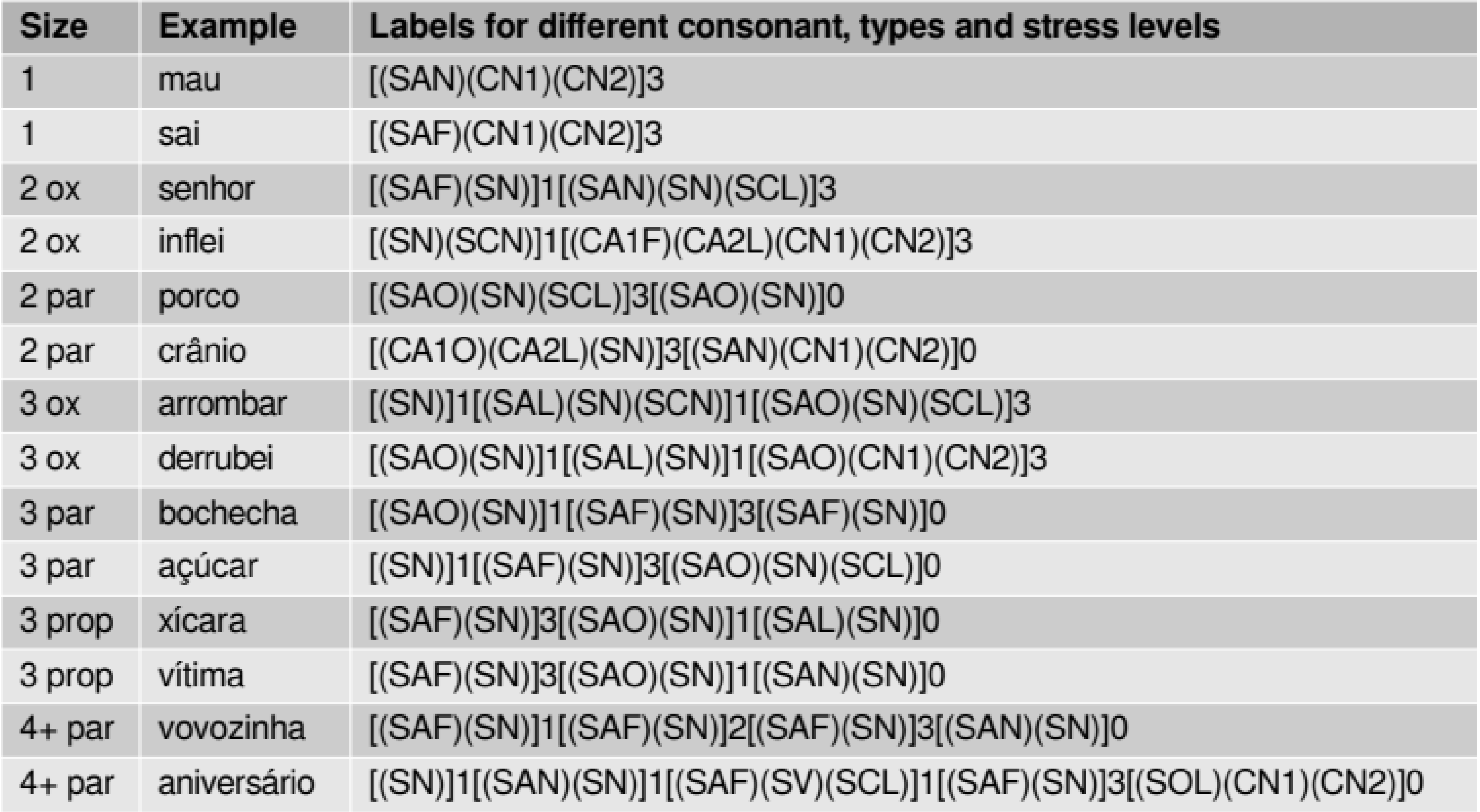
7th. Eicefala

- **Labels**

  - indicate segment boundaries with parentheses;

  - indicate syllable boundaries with square brackets.

Labeling the words in the dataset



Size	Example	Labels for different consonant, types and stress levels
1	mau	[(SAN)(CN1)(CN2)]3
1	sai	[(SAF)(CN1)(CN2)]3
2 ox	senhor	[(SAF)(SN)]1[(SAN)(SN)(SCL)]3
2 ox	inflei	[(SN)(SCN)]1[(CA1F)(CA2L)(CN1)(CN2)]3
2 par	porco	[(SAO)(SN)(SCL)]3[(SAO)(SN)]0
2 par	crânio	[(CA1O)(CA2L)(SN)]3[(SAN)(CN1)(CN2)]0
3 ox	arrombar	[(SN)]1[(SAL)(SN)(SCN)]1[(SAO)(SN)(SCL)]3
3 ox	derrubei	[(SAO)(SN)]1[(SAL)(SN)]1[(SAO)(CN1)(CN2)]3
3 par	bochecha	[(SAO)(SN)]1[(SAF)(SN)]3[(SAF)(SN)]0
3 par	açúcar	[(SN)]1[(SAF)(SN)]3[(SAO)(SN)(SCL)]0
3 prop	xícara	[(SAF)(SN)]3[(SAO)(SN)]1[(SAL)(SN)]0
3 prop	vítima	[(SAF)(SN)]3[(SAO)(SN)]1[(SAN)(SN)]0
4+ par	vovozinha	[(SAF)(SN)]1[(SAF)(SN)]2[(SAF)(SN)]3[(SAN)(SN)]0
4+ par	aniversário	[(SN)]1[(SAN)(SN)]1[(SAF)(SV)(SCL)]1[(SAF)(SN)]3[(SOL)(CN1)(CN2)]0

(elaborated by the authors)

7th. Eicefala

- **How to “read” the labels**

([(SAF)(SN)]1[(SAF)(SN)]2[(SAF)(SN)]3[(SAN)(SN)]0

- from left to right: first syllable has a fricative within a simple onset, followed by a vowel; second syllable also has a fricative within a simple onset, followed by a vowel; third syllable has a fricative within a simple onset, followed by a vowel and carries primary stress; fourth syllable has a nasal within a simple onset, followed by a vowel.

7th. Eicefala

- **Final remarks**

  - labels are machine readable and correspond to some abstraction departing from real data;

  - labels can provide teachers a way to understand the hypotheses children formulate when they make spelling "errors";

  - labels help machine learning and simulating the “errors”.

7th. Eicefala

- **Final remarks**

  - labels do not specify which vowel occurs in syllable nucleus;

  - some "errors" involve vowel quality, such as "rechunchuda" (for "rechonchuda", chubby) or "ispludiu" (for "explodiu", it exploded);

  - new labels accomodating vowel aperture (1,2,3,4) and place of articulation (ft, ct, pt).

7th. Eicefala

- **Next step**

  - implementing the labels in the system;

  - verifying how the system deals with the labels for the words of the dataset and also for additional words;

  - improving the labels, if necessary.

7th. Eicefala

- References

Camara Jr., J.M. Estrutura da língua portuguesa. Petrópolis: Editora Vozes, 1970.

Chacon L, Pezarini I.O. Gradiência na correspondência fonema/grafema: uma proposta de caracterização do desempenho ortográfico infantil. In: César ABP, Seno MP, Capellini SA. Tópicos em Transtornos de Aprendizagem: Parte VI. Ribeirão Preto: Chacon, L.; Pezarini, I. O. Gradiência na correspondência fonema/grafema: uma proposta de caracterização do desempenho ortográfico infantil. In: César ABP, Seno MP, Capellini SA. Tópicos em Transtornos de Aprendizagem: Parte VI. Ribeirão Preto: Booktoy Livraria e Editora, 2018.

Collischonn, G. A sílaba em português. In: Leda Bisol (org.) Introdução a Estudos de Fonologia do Português Brasileiro. Porto Alegre: EDIPUCRS, 1996.