The Encoder-Decoder Model Applied to Brazilian-Portuguese Verbal Irregularities

Sample Extracted and Translated from the Original Text "O modelo Encoder-Decoder aplicado em irregularidades verbais do Português Brasileiro"

Sample of Master Thesis

General Semiotics and Linguistics

University of Sao Paulo

Beatriz Albiero
Advisor: Marcelo Barra Ferreira (Phd)
Linguistics Department FFLCH-USP

São Paulo
2019

Abstract

Albiero, B. **The Encoder-Decoder Model Applied to Brazilian-Portuguese Verbal Irregularities**. 2019. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências Humanas, Universidade de São Paulo, São Paulo, 2019.

Inspired by the controversial debate about the acquisition of irregular verbs in English language (Chomsky, N. & Halle (1968/1991), Pinker & Prince (1988), Albright, A. & Hayes (2003), Kirov & Cotterell (2018)), this research aims to study the inflection process of irregular verbs in Portuguese through the perspective of the computational model *Encoder-Decoder*. To do this, we proposed the task of predicting an inflected verbal form given a primary form (*Stem + Thematic Vowel*). The scope of the research was restricted to the study of the singular first-person paradigm in the indicative mood and present tense. The model, in turn, is an associative model that belongs to the group of Artificial Neural Networks models. Also, it was necessary to construct a linguistic *corpus* composed by the chosen paradigm and then transcribe it into a specific phonetic notation to enable the usage of the chosen model. The resulting *corpus* consists of 423 verbs that were marked as belonging to either regular (51%) or irregular (49%) verb families. Moreover, within the scope of irregular verbs, it was possible to identify 15 subgroups through the identification of inflection patterns. Through the phonetic notation provided, verbs could be associated with new representations that included information related to the phonetic features. Thus, the proposed model attempts to predict inflected forms by identifying the involved phonetic relationships during the inflection process. The model was submitted to multiple trainings and tests and presented an average accuracy of 13.55%, but it got to 17% in one of the experiments. Considering the segmentation between regular and irregular verbs, the model performed better among the regular class. However, considering all 16 classes individually (15 irregular + 1 regular), it was observed that the first two classes in which the model performed best were irregular classes, leaving the regular class with the third place.

Key-words: verbal morphology, machine learning, connectionism.

Introduction

Inspired by the great controversy regarding the acquisition of irregular verbs in the English language (Pinker (1999), Chomsky, N. & Halle (1968/1991), Pinker & Prince (1988), Rumelhart & McClelland (1986)), this research aims to study the inflection process of Brazilian Portuguese irregular verbs making use of a computational model known as *Encoder-Decoder* (Bahdanau, D. (2014)).

The debate surrounding the process of learning irregular verbs started in the 1960s, having been explored by researchers from different fields (linguists, psychologists, neuroscientists, computer scientists, among others). In Chap. 1, we present the origins and motivations of the discussion, as well as the main researchers involved and raised hypotheses. In addition, a category of associative computational models known as the Artificial Neural Network will be introduced, of which the *Encoder-Decoder* model is part. Next, in the first part of the motivation section (Sec. 1.2) we point out some grammatical specificities of the Portuguese language that could hinder the learning process of irregular inflection. In the second part of the section, we will present a historical review regarding the computational contributions that followed in response to the discussion.

After the exhibition of the different conducted studies, it will become evident why this theme caught the attention of so many different areas. Moreover, we will see that the problem of inflecting irregular verbs has become a computational challenge beyond the linguistic or cognitive issues under debate, in such a way that many subsequent researches have ended up distancing themselves from the linguistic matters and focused on the mathematical aspects that would make artificial learning feasible. This research will also continue this computational aspect of the problem and will not aim to take any side regarding the cognitive aspects of the debate. In this way, we will evaluate the use of the *Encoder-Decoder* model from a practical point of view according to its performance in the proposed task and in comparison to other computational models already presented in other research.

To achieve the proposed objective, the first stage of this research was the development of a specific linguistic corpus for the task. The resulting corpus is presented in full in the Appendix (B), but will be discussed in detail in Chap. 2, in Sec. 2.2. Still in this chapter, we will see the importance of applying appropriate preprocessing treatments in the units of the corpus to enable the intended learning. For that, we will revisit a preprocessing algorithm proposed in a previous work, carried out by Rumelhart & McClelland (1986). Then, we will present the preprocessing algorithm developed in this research.

Chapter 3 introduces basic concepts regarding the subject of *Machine Learning* and then present an introduction to Artificial Neural Network models. In addition, we will address the concepts of *Language Models* and architectures of *Recurring Neural Networks* - essential issues for understanding the highlighted model of this research, the *Encoder-Decoder*. After making the necessary introductions, we will be ready to present the *Encoder-Decoder* model in Chapter 4.

Chapter 5, in turn, will present and discuss the results obtained. In it we will see the settings chosen for the definition of the model and the metrics used for its evaluation. We will also analyze the different errors observed and look for possible explanations for the results obtained.

To conclude, Chapter 6 will display a summary of the subjects covered in this research and also point out suggestions for future research on the subject.

Chapter 1

Learning Irregular Verbs

The first part of this chapter will be dedicated to a presentation regarding the origin of the discussion about the inflectional learning process of irregular verbs. Thus, we will build an overview of the problem and present the main hypotheses and research involved. Then, we will discuss the motivations that led to the production of this research. In this sense, we will see that, on the one hand, there is a linguistic interest, motivated by a greater complexity of the Portuguese language compared to English. On the other hand, we will also see that there is a computational interest, motivated by the emergence of new modeling resources. Finally, we will see a section aimed at delimiting the scope of the research.

1.1 Context

In the field of Language Acquisition, the problem regarding the inflection process of irregular verbs in the English language is certainly among one of the most controversial topics of debate in Linguistics (Chomsky, N. & Halle (1968/1991), Pinker (1999), Pinker & Prince (1988), Albright, A. & Hayes (2003), Kirov & Cotterell (2018)). The heart of the debate is the exact characterization of the mechanisms that enable a speaker to be able to relate a verb in its non-inflected form (*walk*, for example) to its inflected form in *Simple Past* (*walked*).

Past tense verbs in English can be subdivided into a variety of families. A first group would be the set of *regular verbs*, whose form corresponds to the application of the rule *root + ed* (as in the example seen from the verb *to walk*). Among the irregular verbs, these can be considered suppletive, that is, they have a unique inflection process and no apparent rule, such as *go* → *went*, or they can conglomerate following phonetic patterns (Nelson (2010)):

1. blow – blew, grow – grew, know – knew, throw – threw

2. bear – bore, swear – swore, tear – tore, wear – wore

3. drink – drank, shrink – shrank, sink – sank, stink – stank

One could think that the learning of such patterns would depend on a case by case memorization. However, Bybee & Moder (1983) shows a psycholinguistic study in which individuals are presented with several invented verbs (hypothetically in a non-inflected form). The research revealed that, instead of systematically applying the regular rule (verb + *ed*), individuals showed tendencies to allocate some verbs in some irregular subgroups. For example, for the invented verb "*spling*", most individuals chose the form "*splang*" or "*splung*". This example contradicts the idea that speakers could be simply reproducing memorized forms and suggests that they are actively identifying patterns. In addition, they have a natural intuition about the appropriateness of allocating a verb to one group of verbs or another.

From the given example, it is reasonable to think that the motivation behind such trends occurs from the similarities between the phonetic units of the invented verbs and the real verbs that already have an irregular character inflection. However, the circumstances that lead to the acquisition of this linguistic *intuition* are undetermined. On the one hand, it makes sense to say that for a human being to be able to be introduced into the world of speakers, it would require some inherited capabilities, otherwise the learning process would not be possible. In contrast, studies show that children deprived of contact with a speaking society are permanently unable to fully master the grammar of a language (Pinker (1994/2007)), which leads us to conclude that children's experience with society, as well as their own genetic pre-dispositions are both partly responsible for the language development process. The difficulty, therefore, is in the attempt to quantify, delimit and point out the knowledge acquired from cultural contact, as well as the so-called *inate* linguistic knowledge. It is, therefore, around this issue that the debate about the learning of irregular verbs in the English language begins.

On one side of the debate, there is the theory of Generative Phonology by Chomsky and Halle (1968/1991). In this theory, individuals would be carriers of a language acquisition device (*LAD*) responsible for *formulating* and *manipulating* abstract phonological structures in a system of rules. In a simplified way, the theory proposes that the speaker is able to intuitively identify and formulate rules to account for the learning of irregular forms of the language. An example of this is the family of verbs ending in "-ind".

bind – bound, find – found, grind – ground, wind – wound

We can see that, in a simplified way, a rule can be proposed as:

aɪ→ aʊ/ **X** ___nd]+past

The proposed rule suggests that the segment [a ɪ] becomes [a ʊ] when followed by [nd] and inflected to the past. The symbol _ _ represents the place where such a transformation occurs and **X** represents an arbitrary phonological unit.

In other words, it can be said that the knowledge said *innate* defended by Chomsky and Halle refers to a certain cognitive capacity for formulating rules from the identification of some fundamental elements (such as the elements pointed out in the proposed rule).

Such a structure would allow speakers to build generalizations and, eventually, abstract the phonological rules of their language.

On the other side of the debate, the researchers Rumelhart & McClelland (1986) confront the previous theory by arguing that behaviors of a regulated character can be reproduced by mechanisms that do not depend on any symbolic manipulation. Instead, the researchers suggest that the mechanisms involved in the verbal inflection process can be constructed in such a way that its performance can be described through rules, but that the rules themselves are not explicitly represented in any part of the process. To support this idea, Rumelhart & McClelland (1986) present a computational model based on associative patterns that do not use constructions with rules of this type. Subsequently, the model built was fundamental for the emergence of a new school within the cognitive sciences: connectionism.

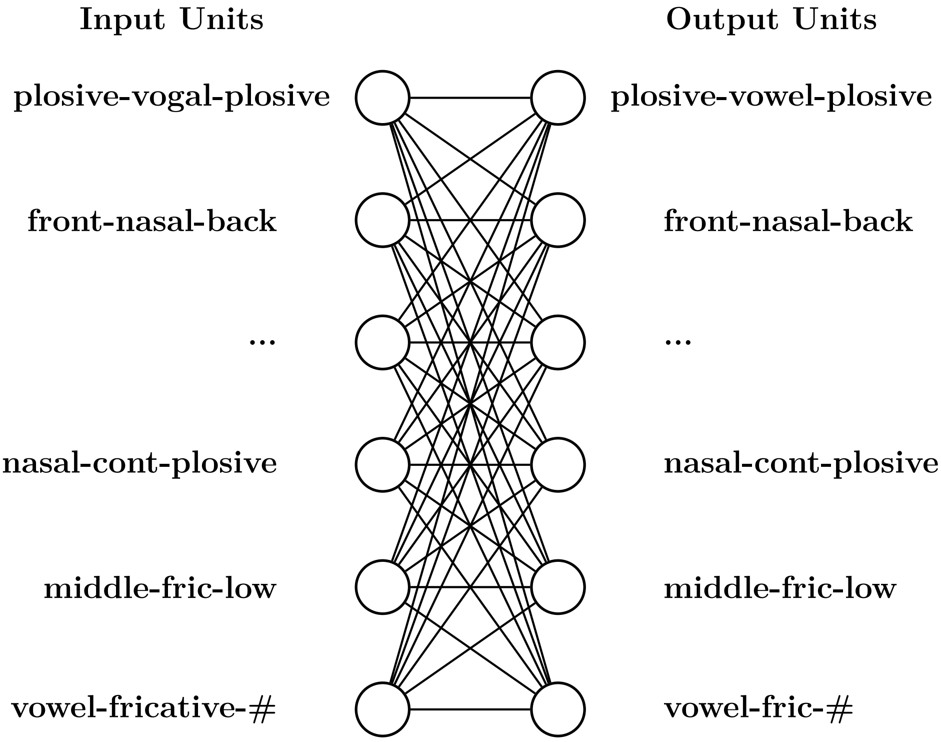

Figure 1.1: *Model Scheme proposed by researchers Rumelhart and McClelland*

The model was created by analogy to the structure in which neurons in the brain relate. It is basically composed of an artificial network of nodes interconnected in parallel (Fig. 1.1).

The first layer of nodes in the model is responsible for receiving the input data, which are the data regarding the distinctive phonetic features that characterize the sounds of a verb in its root form. Phonetic features can be characterized as distinctive properties of phonic units (Seara, I. C. (2015)). Such properties can be based on acoustic, articulatory or perceptual criteria. In the figure, each node is shown next to a sequence of three features. The first node, for example, refers to the sequence **plosive-vowel-plosive**. In this case, *plosive* indicates a common property among some consonants, related to the interruption of the air flow (as in the phone [k], for example). In the figure we still have **fric** for fricatives, **vowel** for vowels, **nasal** for nasality, places of execution (**front** and **back**), tongue features (**middle**, **low**), among others. Still on the *input* layer, it is possible to observe that each node is represented by a triad of phonetic features. This was the solution provided by the authors to perform the mapping between the features of the verbs in their root form to their *Past Simple*. The *inputs* are structured in this way to overcome the difficulty of inserting data of a sequential nature and of variable size (as is the case of a verb - composed of a sequence of sounds). Each triad is an association of three features, each referring to a phone. For example, for the verb *came* (transcribed in phonetic form by the authors as */kAm/*), each phone has multiple features. The phone [*k*], for example, is an plosive, voiceless and "back" consonant. Subsequent phones are also composed by their respective phonetic features, in such a way that we can represent the verb as a sequence of lists of phonetic features (Table 1.1). The topic regarding the phonetic features used in this research, as well as the entire preprocessing scheme used by the authors Rumelhart & McClelland (1986) will be discussed in greater depth in Chap. 2.

k	A	m
voiceless	long	nasal
interrupted	vowel	interrupted
back	low	front
consonant	front	consonant

Table 1.1: *Trigrams of Phonetic Features Used as Inputs to Rumelhart and McClelland's model*

Returning to Fig.1.1, we see a network of connections after the *input* layer. Each connection, in turn, has a *weight*. These weights will act as a kind of *input* filter, making weights with higher values pass the information on with more effect (or strength), and smaller weights with less. The second layer of nodules in Fig. 1.1 is a response layer (known as a *output* layer) that aims to return data regarding the features that characterize the sounds of the same verb provided in *input*, but in *Simple Past*.

During the learning process of the model, the output data from the *output* layer must then be compared to the correct form of the verb in the past tense, through a kind of template, known as *target* (Fig. 1.2). Having made this comparison, it is possible to change the network of connections between *input* and *output* layers in order to reinforce (or weaken) their weights to achieve the proposed learning. Before the first comparison, the network is initialized with random weights. As the number of comparisons increases, the tendency is for the weights to be gradually calibrated in order for the model to reach its objective, which in this case, is to learn the inflection patterns of the verbs.

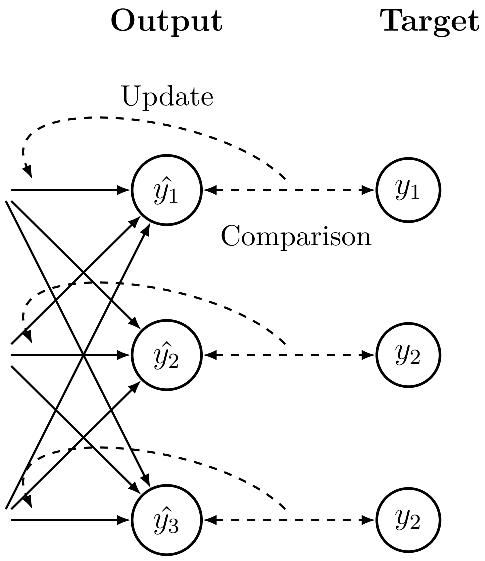

Figure 1.2: *Comparison between Output and Target*

To conduct the training, Rumelhart & McClelland (1986) fed 420 verbs into the model repeatedly (200 times each, 84,000 insertions in total). After training, the model was able to correctly reconstruct all of their irregular forms. In addition, in a new group with 86 unknown verbs, it hit about 2/3 of the set. Among the new irregular verbs, it made some interesting errors of *overregularization* (such as *catched* (instead of *caught*) and *digged* (instead of *dug*)). These errors were observed in 11 of the 14 irregular verbs tested (Prasada, S. & Pinker (1993)).

In addition to these results, Rumelhart & McClelland (1986) report that the model's learning process presents an interesting phenomenon, with a performance similar to behaviors observed in children during an acquisition phase: the U-Shaped Development Curve, Marcus, G. (1992)). The U-Shaped Development curve basically refers to a learning process that takes place in three stages:

(i) first, children can correctly reproduce the expected irregular inflection of verbs (*come*→*came*);

(ii) next, they go through a process of *overgeneralization* (in which they produce shapes like *comed*) ;

(iii) finally, they start to correctly reproduce both regular and irregular verbs.

Rumelhart & McClelland (1986) describe how it was possible to observe such behavior in their model. In the initial phase of the training process, the model was fed with a small amount of verbs, such as: *come*, *get*, *give*, *look*, *take* , *go*, *have*, *live* and *feel*. The model's performance was compatible with the first stage of the curve, that is, for these verbs it was able to correctly identify the corresponding form in *Simple Past*. In a second step, the model was fed with a much larger amount of verbs. At this stage, it is possible to verify that the model is undergoing a process of systematic regularization of verbs. It produced results like: *breaked*, *comed*, *gived*; and also combinations between regular and irregular patterns (e.g. *gaved*). After a series of many repeated insertions, the model was finally able to respond correctly to a larger number of verbs, as in the last stage of the natural learning process.

The results presented by Rumelhart & McClelland (1986) had a considerable impact on the scientific community at the time. Many researchers saw the new model as a complete paradigm shift, not only in linguistics, but also as a new way of studying learning in general (Schneider (1987)).

Despite this, Pinker & Prince (1988) continue the debate by pointing out a number of pertinent issues that the model has failed to explain. First, as the model receives only a phonetic representation of the verb as *input*, it is unable to generate two different responses for verbs with identical sound (for example *break* → *broke* and *brake* → *braked*). To make these predictions correctly, the model would require an additional module to distinguish between the two words, but then it could not be considered as a purely associative model anymore. Second, the model is extremely dependent on the patterns observed during training, having a low capacity for generalizations. Pinker (1999) comments that the model displayed no response when fed with the verbs *jump*, *pump*, *warm, trail* and *glare* (which have a unusual sound). In addition, the model presented some completely distorted results, such as: *squat - squakt, tour - toureder* and *mail - membled*; unacceptable associations for any native speaker.

Regarding the observed learning pattern (the U-Shaped Development Curve), Pinker (1999) explains that this behavior was caused according to the way in which the verbs were inserted into the model during training: Rumelhart & McClelland performed the training in parts and controlling the amount of repetitions of each batch of verbs. In the first part of the training, they selected some high frequency verbs in the English language (many of them irregular), reproducing stage (i) of the curve. Then, they trained the model with these verbs, reintroducing them multiple times until the model managed to achieve reasonable performance on those verbs. Next, they introduced a larger number of verbs, these being less frequent than the previous ones, but mostly regular. Thus, the model began to adjust to apply the regular rule and thus it was possible to observe stages (ii) and (iii) of the curve. Still, according to Plunkett, K. & Marchman (1991), the stages of development (i), (ii) and (iii) can be considered part of a behavior *macro U-shape*, but in the process of natural learning, it is still possible to observe the occurrence of a behavior *micro U-shape.* Plunkett, K. & Marchman (1991) point out that the reproduction of irregular verbs in spontaneous speech by apprentice children varies considerably between correct inflections and *over-regularized.* They also note that these oscillations occur in different proportions for each verb and that children rarely "*irregularize*" regular verbs (like *ping* → *pang*), and rarely mix an irregular shape with a regular (a fact that occurred while learning the model with *gaved*).

To conclude, Pinker & Prince (1988) presents the formulation of a new hypothesis for such a question: a hybrid theory in which generative phonology would apply to the regular inflection process and an associative mechanism would apply to the irregular inflection process. The researchers propose that regular forms are computed from a mechanism that should abstract the stem of the verb and combine it with the suffix –ed. Such a mechanism can be applied to any word, in a process independent of memory. Irregular forms, in turn, go through a different process: irregular verbs must go through a memorization process, with not only the association between one verbal form and another, but also between properties (phonetic features, rhyme, radical, nucleus, etc.), similar to what was proposed by Rumelhart and McClelland.

1.2 Motivation

Once a contextualization about the problem has already been presented, we will now focus on the questions that motivated the development of this research. Thus, this section will be divided into two parts: (i) motivation in the field of General Linguistics, and (ii) motivation in the field of Computer Science.

1.2.1 Motivation in the field of General Linguistics

The verbal morphology of the English language is quite simple if compared to Portuguese. First, Portuguese verbs are divided into three conjugation classes, each of which is defined from a *thematic vowel* (*/a/*, */e/* and */i/*). Given a verb in its infinitive form, for example *Amar - ("to love")*, the thematic vowel (TV) is the vowel that is found between the verb lexical morpheme (the root) and the infinitive ending *r*.

$$\text{Am} + \text{a} + \text{r}$$
$$\text{Root} + \text{TV} + \text{r}$$

With this, the three possible types of conjugation are: 1st - ar (amar (to love), brigar (to fight)), 2nd - er (beber (to drink), comer (to eat) and 3rd - ir (rir (to laugh), descobrir (to discover)). In the English language, this distinction does not exist.

A child in the process of language acquisition in the Brazilian Portuguese system goes through many challenges. Part of the process is precisely to understand the relationship between the thematic vowel and the possible regular conjugations. In this process, it is not uncommon to observe the appearance of conjugation exchanges. Wuerges (2014) presents linguistic data produced by children with several of examples. Some of these can be seen in Tab. 1.2.

Verb	Translation	Observed	Correct	Exchange
botar	*to put*	“boti”	botei	1^{a} with 2^{a} or 3^{a}
comer	*to eat*	“comei”	comi	2^{a} with 1^{a}
jantar	*to have dinner*	“janti”	jantei	1^{a} with 2^{a} or 3^{a}

Table 1.2: *Examples of Exchanges in Verbal Conjugations During Verbal Acquisition*

Irregular verbal forms are an additional difficulty in this process for children who speak Portuguese. Wuerges (2014) also points out observed examples of *regularization* of irregular verbs: “eu *consego*” (regularization of the verb “conseguir” (to succeed)) and “eu *podo*” (regularization of the verb “poder” (can)).

A verb is said to be irregular if it presents changes in the stem (in relation to the stem of the infinitive form) and/or in the flexional suffix (in relation to the regular pattern imposed by each conjugation) (Wuerges (2014)). Flexional suffixes (FS) are the segments added after the verb stem. They can be divided into two types: (i) mode-time suffix (MTS) and (ii) personal-number suffix (PNS). For the verb “gostávamos” (we liked), for example, the segment */gost/* is considered as the stem of the verb, */av/* as the suffix mode-time, which in this case simultaneously marks the indicative mode and past imperfect tense; and *amos* as the personal-number suffix indicating first person plural (we).

Following the proposed definition, it is necessary to reinforce that the interest of this study is in capturing irregularities in the phonetic level, therefore verbs such as: “gosto” → /gɔstʊ/ (i like)), “boto” → /bɔtʊ/ (i put) and “coloco” → /kolɔkʊ/ (i place), whose spelling presents the regular pattern, will be classified as irregular. This will happen due to the phonetic transformations that occur in speech and that are not captured in writing. In the case of “gosto”, for example, we see that the first vowel “/o/” actually has a / ɔ/ sound. Thus, Table 1.3 displays some examples of the applied categorizations.

Infinitive	**Conjugated**	**Applied Category**	**Phonetic Transc.**	**Translation**
Falar	Falo	Regular	falʊ	(I) speak
Gostar	Gosto	Irregular	gɔstʊ	(I) like
Testar	Testo	Irregular	tɛstʊ	(I) test
Ansiar	Anseio	Irregular	ãseʒʊ	(I) wish
Pedir	Peço	Irregular	pɛsʊ	(I) ask
Mentir	Minto	Irregular	mintʊ	(I) lie
Por	Ponho	Irregular	poñʊ	(I) put

Table 1.3: *Examples of Verb Categories Regarding the Presence of Irregularities*

An observation about the disposition of the irregularities present in Brazilian Portuguese (taking into account only first-person singular (present tense and indicative mode) allows us to observe some regularities (patterns) among the irregular verbs:

Bobear – Bobeio, Bloquear – Bloqueio, Chatear – Chateio, Clarear – Clareio, Golpear – Golpeio

Agredir – Agrido, Conseguir – Consigo, Inserir – Insiro, Perseguir – Persigo, Preferir – Prefiro, Proferir – Profiro, Repetir – Repito, Servir – Sirvo, Vestir – Visto

Cobrir – Cubro, Dormir – Durmo, Engolir – Engulo

Al[e]gar – Al[ɛ]go, C[e]gar – C[ɛ]go, Compl[e]tar – Compl[ɛ]to, Col[e]tar – Col[ɛ]to, Entr[e]gar – Entr[ɛ]go, Pr[e]gar – Pr[ɛ]g,

Ad[o]rar – Ad[ɔ]ro, Ad[o]tar – Ad[ɔ]to, B[o]tar – B[ɔ]to, C[o]lar – C[ɔ]lo, F[o]car – F[ɔ]co, M[o]rar – M[ɔ]ro, S[o]ltar – S[ɔ]lto, S[o]lar – S[ɔ]lo, T[o]car – T[ɔ]co, M[o]strar – M[ɔ]stro

Mentir - Minto, Sentir - Sinto

The patterns observed from the exposure of some irregular classes, allow, as in English, the proposition of formulas, or phonetic rules, that explain the inflections performed in each class. It is possible to notice, for example, that a verb from the same family of *conseguir* follows the rule:

$$e \rightarrow i/\_C]ir$$

The proposed rule indicates that /e/ becomes /i/ when in a third conjugation context (ir). In this case, C indicates any consonant.

The patterns found suggest not only the possibility of elaborating rules, but also the possibility of developing networks capable of capturing such dependencies.

1.2.2 Motivation in the field of Computational Science

Since the presentation of the research by Rumelhart & McClelland (1986), the associative model used by the authors has gone through several advances. In reality, this type of modeling today is called *Artificial Neural Network* and has come to be used in a variety of computational tasks, such as image classification, text classification, automatic translation, conversational agents, among others.

In recent years, computing power has increased a lot. The development of *hardwares* makes it possible to perform many more computations today and with much more speed than in the 1980s. With this, more complex architectures could be explored and developed. As an example, it is possible to add intermediate layers between the layers of *input* and *output*. Networks built in this way make it possible for input information to be distributed over more connections and thereby enhance learning. This is because the type of architecture without intermediate layers can approximate only one type of function, the linear functions. By increasing the number of intermediate layers, it is possible to expand the universe of solutions for solving more complex problems (see Goodfellow, I. & Bengio (2016) for more detailed explanations of the intermediate layers). This type of modeling that follows a one-way flow (from *input* to *output*) is called *Feedforward* (FFD). However, there are many types of architectures whose flows do not follow this configuration. In this context, a type of architecture that has become very well known is the *convolutional* (*Convolutional Neural Networks (CNN's)*) (widely used in the area of computer vision (Krizhevsky *et al.* (2012), for example). Regarding linguistic tasks, *Recurrent Neural Networks (RNN's)* are widely used, since the architecture allows the insertion of sequential data (Liu, P. & Qiu (2016) , for example) The theme of Neural Network models, especially RNR's architecture, will be discussed in more detail in Chapter 3.

With regard to the question of learning irregular English verbs, a series of new experiments followed on from the criticisms of (Pinker & Prince (1988)). (Plunkett, K. & Marchman (1991), (1993)) simplify the issue by considering only fixed-length verbs (3 syllables) and address the problem using an architecture with the addition of intermediate layers (*Multi-Layered Perceptron - MLP*). Other works transformed the issue into a problem of *classification*, so the model would no longer have the objective of finding a flexed form. Instead, they would have a finite and predetermined set of possible ways. Nakisa & Hahn (1996), for example, classify plurals of German nouns. Plunkett, K. & Nakisa (1997) attack the same problem, but for the Arabic language.

Westermann, G. & Goebel (1997) present a model built to map non-inflected verbs from the German language to the participle form. The model presented is capable of handling data with sequences of varying sizes and uses an architecture based on RNN's. However, the model was built from a double route mechanism, so that irregular verbs passed through a specific memorization route and regular verbs through another route, with the application of the rule. Despite this, the performance of the model left something to be desired. Some different experiments were carried out, one of which consisted of training groups of verbs in isolation. For this, the regular verbs were in one group and the irregular verbs were divided into two others. In this case, the percentage of correctness of the model reached almost 100% for the groups trained in isolation. However, when mixing the verbs in a single training, the percentage of correct responses stabilized around 60 to 70%.

As part of the construction of non-associative models, that is, based on rules, Albright, A. & Hayes present the *Minimal Generalization Learner (MGL)* model, whose implementation is very close to the theory proposed by Pinker & Prince (1988). The MGL model is based on discovering and assigning weights to pre-established rules for irregular transformations. We will talk more about the results of this model in the results discussion section (Section 5.3).

1.2.3 State-of-the-Art Architecture

Bahdanau, D. (2014) e Sutskever (2014) present a new type of Neural Network architecture built to map two sequences of varying sizes. The new architecture, known as *Encoder-Decoder*, or also *Seq2Seq*, is recognized especially for its good performance in linguistic tasks, especially in the field of machine translation (Wu (2016)). This architecture consists of the concatenation of two RNN's. The first network, *Encoder*, reads each symbol from an input string (for example, an English word) and generates an abstract representation of the word read in response. The second network, *Decoder*, receives as input the representation returned by *Encoder* and aims to produce another corresponding target sequence (in this case, a translation task, the word translated into another language). The architecture of the *Encoder-Decoder* model will be covered in more detail in Chapter 5.

In the task of learning irregular verbs, Faruqui (2015) elaborates the question at the level of *characters* (letters), that is, it does not use phonetic features such as *inputs* in the *Encoder-Decoder*. Kann, K. & Schütze (2016) use morphological labels (participle markings, gerund, etc.) as *inputs* in the same model. Cotterell *et al.* (2016) achieve *state-of-the-art* performance in a problem of morphological flexion, postulated in a shared task (*SIGMORPHON Shared Task* (http://www.sigmorphon.org/)). In this problem, morphological data sets for 10 languages (Spanish, German, Finnish, Russian, Turkish, Georgian, Navajo, Arabic and Hungarian) with different typological characteristics were introduced. For the task of generating motto inflections, the Cotterell *et al.* (2016) system obtained an average accuracy of 95.56% in all languages, varying from Maltese (88.99%) to Hungarian (99.30%).

Thus, we have as motivation for this research the use of the *Encoder-Decoder* model for learning verbal inflection. The *Encoder-Decoder* model, despite presenting promising results, has not yet been applied at an exclusively phonetic level (as in (Rumelhart & McClelland (1986))), since previous research has shown results at the level of characters and with morphological labels. In addition, we also see that the model has never been tested in the Portuguese language.

1.2.4 Scope

In comparison to English, it is possible to argue that the Portuguese language is more complex. In English, considering the present tense, we see the distinction of only two forms: (i) the form for *I, We, They* (as in *walk*) and (ii) *he/she/it*(*walks*). An exception to this rule is the verb *to be*, which has three possible forms: (i) I *am*, (ii) he/she/it *is* and (iii) they *are*. In the case of the *Simple Past*, it presents a greater number of forms also only for the group *to be*: (i) I/he/she/it *was*, and (ii) They/We *were*; the others do not have person identification (Nelson (2010)). In Portuguese, the traditional norm distinguishes six people: 1st: Eu (I), 2nd: Tu (You), 3rd: Ele/Ela (He/She), 4th: Nós (We), 5th: Vós (You - plural), 6th: Eles (They). However, today we see that the forms associated with 2nd e 5th persons are falling into disuse. The 2nd person, "Tu", is still used in some regions of Brazil, but it has usually been reproduced as the form of 3rd person (Você (You)/Ele (he) *gosta* (like)). Likewise, the 5th form has been replaced by the 6th person (Voc ês (You - plural)/Eles (They) *gostam* (like)) (Câmara Jr. (1999)).[1]

Regarding irregularities, in English irregular verbs are found only in the *Simple Past* and *Past Participle*, while the verbal system of Portuguese is full of irregularities in all times, modes and persons. Take the verb "*dizer*" (to say), for example. At the present time and indicative way, we note the presence of irregularity in the first and third persons of singular ("*digo*" e *"diga"*). In the Perfect Past tense (pretérito perfeito), the paradigm is totally irregular, with: "*disse*", "*disseste*", "*disse*", "*dissemos*", "*dissestes*" and "*disseram*". In the Simple Future, we have: "*direi*", "*dirás*", "*dirá*", "*diremos*", "*direis*" e "*dirão*".[2]

Despite the complexity of Portuguese being relevant to the challenge of learning to inflect irregular verbs, we opted for a more restricted study. The reason for this limitation stems mainly from the absence of a corpus prepared to carry out the task. Thus, by limiting irregularities to a single paradigm, we put the level of difficulty at a level similar to that of the research by Rumelhart & McClelland (1986), with the main difference (and possibly extra difficulty) being the three possible combinations.

The scope was therefore restricted to 1st Person of the Singular in the Present tense and Indicative mode (with examples already explored in Section **??**). In this way, verbs that show irregularity in another tense, mode or person other than the 1st singular person in the present tense and indicative mode, will be treated as belonging to the regulars class. As an example, consider the verb *correr* (*to run*). This verb has regular inflection for the 1st Person (*corro* (/ko rʊ/))), but it is irregular for the 3 rd Person (*corre* (/kɔ ri/)). Thus, although *correr* is an irregular verb, as it has a regular inflection in the chosen paradigm, it will be treated as regular for the purposes of this research.

[1] It is also interesting to note that lately the 4th (N ós - We) has been alternated with the use of "A gente", whose verbal form also corresponds to the 3rd person (A gente *gosta*).

[2] It is also possible to observe irregularities in the subjunctive and imperative modes in all paradigms.

Chapter 2

Preprocesing Verbs for Neural Networks

The objective of this chapter is to introduce one of the most important steps in Neural Network modelling process: data preprocessing. In the field of Machine Learning, there is frequently a lot of excitement around the development of new modelling techniques and architectures. However, it is fair to say that the step of preprocessing - which is the step we usually spend most of the time - does not receive this same attention.

In this work, the data of interest are *verbs*, and for a computer, a verb is just a sequence of characters. In addition, a Neural Network model is a computational model. Hence, for the model to work, it is necessary to prepare the data so that they can be used in a format suitable for the calculations to be performed. In this way, even without going into the theoretical part of the model's functioning, it is worth discussing how the verbs will be inserted into the it. Simply put, neural network models are fed with *numeric vectors* (or *arrays*). These vectors are essentially a list of numbers (e.g. [0, 1, 2, 3]). Thus, preprocessing the verbs to feed the model means finding a *vector representation* for each verb. Furthermore, there will be another constraint for this vector representation, which is that all verbs will be represented by vectors of equal length, and that it will be this same length that will define the dimension (number of nodes) of the first layer of the network (the layer of *input*). In addition, it is also at this point that we will define the level of abstraction of the study. In other words, this means that we can cut out a verb in several ways: we can consider that verbs are a sequence of letters, a sequence of phonemes, a sequence of morphemes, etc. However, when we make this cut, we will automatically be skewing the final result of the study. For example, if we describe it by representing a verb using its spelling, the model will fail to find the most subtle relationships between the phonetic elements of that verb.

Hence, it is evident the importance of this step. The first section of this chapter will be dedicated to presenting the preprocessing algorithm adopted by Rumelhart & McClelland (1986). Next, the corpus used in this research will be introduced. Finally, the last sections of the chapter will be dedicated for the presentation of the preprocessing algorithm used in this work and that will be applied to feed the developed model. In terms of terminology, a disclaimer must be made: sometimes the term *coding* will be used as an option for preprocessing. Similarly, *decodification* and *post-processing* will be used to describe the reverse process.

2.1 Preprocessing presented by Rumelhart & McClelland

The architecture of the model used by Rumelhart & McClelland (1986) was somewhat limited to the problem of learning the inflection of verbs. The limitation resulted from the fact that both the *inputs* and the *targets* of the model had variable sizes. The *input*, for example, could be "*like*" or "*overtake*", and the *targets*, "*liked*" and "*overtook* ". However, the architecture used (already presented in Fig. 1.1) is composed of a fixed number of nodes in each layer. We could simply assume that each node represented each phone in the phonetic alphabet. Thus, the hypothetical *input* would be the set of phones of the verb. However, in doing so, the network would completely lose track of the phones. Given the limitation, Rumelhart & McClelland (1986) ended up developing a coding system composed of several steps. We will use the verb "*came*" (the past tense of "*to come*") as an example to detail each of the coding stages used by the researchers.

The first step consisted of transcribing the verbs using an alphabet compatible with the ASCII code (Mackenzie (1980)). The ASCII code is a code used to represent text on computers. It encodes letters of the Latin alphabet, punctuation marks and mathematical signs. The option to use the ASCII code is necessary, as the phonetic code is not interpretable by the programming languages. According to the researchers' transcription key (Appendix A), the verb "*came*" was transcribed to "*#kAm#*". The symbol " " is used to mark the beginning and end of the verb.

However, such an encoding would not be enough. Rumelhart & McClelland (1986) needed to find a way to present the phones sequentially so that the model could capture the linear order of phonetic information. The problem is that the verb would have to be inserted all at once, so it would not be enough to find representations for each phone individually, it would be necessary to find a complete representation for all phones in the verb and that somehow still preserved some notion of sequence. Thus, Rumelhart & McClelland (1986) had the idea of reusing a concept created by Wickelgren (1969), the concept of *Wickelphone*. With that, the transcripts were restructured into *trigrams*. Doing this means analyzing the verb sequence in three segments, that is, in this step, the sequence " *kAm* " started to be treated as " *kA*" + "*kAm* "+" *Am* ", each of which is considered a *Wickelphone*.

Despite this, treating verbs with their respective *Wickelphones* would still not be enough for the model of Rumelhart & McClelland (1986). The point is that the authors would like the model to be able to identify more subtle relations between verbal forms, that is, to identify relations closer to the sound execution of words than at the level of the phones. In this way, the authors present a new structure inspired by the representation of *Wickelphones* by Wickelgren (1969), entitled *Wickelfeatures.* The *Wickelfeatures* are precisely the triads of phonetic features introduced in Cap. 1. As seen in Table 1.1, each of the phones can be described by some characteristics related to the execution of their respective sounds. However, the characteristics of the formed trigrams can be combined in many ways, more precisely, 4^3 possibilities, since each phone can be described by 4 dashes (see Appendix A)).

Thus, for the construction of the model's nodes, Rumelhart & McClelland (1986) computed all possible *Wickelfeatures* between the verbs and excluded some redundant ones to simplify the computation. In the end, the verbs *Wickelfeatures* were mapped as keys in a fixed-length dictionary (a vector) in which the values could assume only 0 or 1; 1 if that *Wickelfeature* was present in the verb and 0 otherwise. Figure 2.1 illustrates the coding scheme for the verb "came".

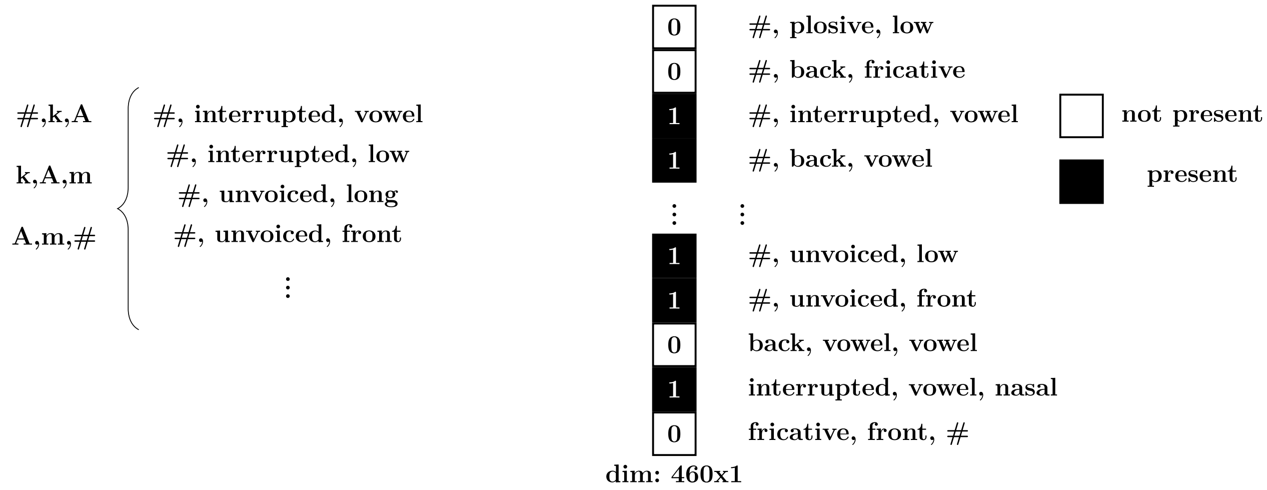

Figure 2.1: *Coding Scheme by Rumelhart and McClelland*

As mentioned in Chap. 1, the model of Rumelhart & McClelland (1986) achieved a reasonable performance, hitting 2/3 of the test set. Despite this, we will see that the *Feedforward* architecture is not the most suitable for this type of problem. First, the proposed coding scheme is quite limited, since the network's *input* vectors can only mark the presence or absence of *Wickelfeatures.* Pinker (1999) comments on this problem using as an example the word "*algalgal*" (a word from the *Oykangand* language), in which trigrams are repeated. We can observe, in this case, that it would be impossible to mark such repetition by using the chosen representation. Furthermore, we can see that the problem occurs not only in case of repetition of trigrams, but also in cases of repetition of *Wickelfeatures.* As an example of the problem, we can analyze the verb "*understand*". Table 2.3 displays a comparison between the *Wickelfeatures* of the trigrams *der* and *tan*, both present in the example verb. In the table, we can see that many *Wickelfeatures* are repeated despite the fact that the two trigrams are completely different.

d	e	r
consonant	vowel	consonant
voiced	frontal	nasal
plosive	short	middle
middle	anterior	voiced

t	a	n
consonant	vowel	consonant
unvoiced	front	long
plosive	short	middle
middle	low	voiced

Table 2.3: *Wickelfeatures de der e tan*

In addition, just as verbs have to be processed to enter models as vectors, the model output also needs to be decoded for the reconstruction of the verb - the *decoding* process of the *output* of the model. The decoding process involved a scheme that consisted of several steps. In a simplified way, Rumelhart & McClelland (1986) listed the candidate trigrams (*Wickelphones*) for each verb and these “competed” for the most relevant *Wickelfeatures* vectors in the model output. As the output from the FFD network is also a vector that stores only the presence or absence of the dashes, and not how many times they appear, the decoding process presents problems to correct verbs with repetitions of *Wickelfeatures*. The problems observed both in coding and in the decoding of the data reflected the need for a more adequate architecture for the problem in question.

2.2 Introducing the Corpus

The corpus used in this research was built from the list of verbs contained in the address www.conjugacao.com.br.

First, the verbs and their respective inflected forms were extracted from this page to a *.csv* file through a technique called *webscraping*, a technique that extracts information contained in the web pages (Mitchell (2015)). Then, the irregular verbs were selected manually for different families of verbs, that is, groups that contained the same inflection pattern. Some of the irregular verbs listed in the reference source were not irregular in the process of inflection of interest, and therefore were reallocated to the group of regular verbs (see example of the verb *"correr"* in Section 1.2.4). In the sequence, the verbs collected in infinitive form had the phone /*r*/ extracted so that only the stem + thematic vowel of the verbs entered in the model *input*. The reason for this was only a matter of convenience since this markup was redundant because it is present in all verbs.

An experiment was carried out in an attempt to use the automatic phonetic transcriber provided by Guide (2016) to make the transcription process faster. However, the transcriber failed to transcribe verbs whose written forms coincide with nouns, such as: "apoio" (support), "peso" (weight), "toco" (touch/stump), "posto" (post/gas station), "jogo" (play/game); ignoring the phonetic characteristics of these words as verbs.

Thus, the verbs collected were transcribed manually using the transcription key presented in Table 2.9. In total, 423 verbs were obtained, 83 less than in the experiment carried out by (Rumelhart & McClelland (1986)). Of the 423 verbs extracted, 20 were considered verbs with no possible grouping (such as "ir" (to go), "trazer" (to bring) and "saber" (to know)), totaling a base of 214 regular and 209 irregular verbs (50.6 % and 49.4 % respectively). The study by Rumelhart & McClelland (1986), in turn, had a portion of irregular verbs of only 20%.

Table 2.4 associates examples of verbs of the classes obtained to their respective counts and proportions in the corpus.

	Examples	Count	Portion
1	ansiar, anseio	9	2.13%
2	botar, boto	30	7.09%
3	cobrir, cubro	7	1.65%
4	dizer, digo	7	1.65%
5	fazer, faço	15	3.55%
6	crer, creio	5	1.18%
7	sentir, sinto	8	1.89%
8	pedir, peço	7	1.65%
9	pôr, ponho	27	6.38%
10	seguir, sigo	27	6.38%
11	ter, tenho	10	2.36%
12	testar, testo	20	4.73%
13	ver, vejo	6	1.42%
14	vir, venho	10	2.60%
15	saber, sei	20	4.73%
16	falar, falo	214	50.59%

Table 2.4: *Corpus composition*

Although the volume of irregular verbs is considerably greater than the volume used in the study of Rumelhart & McClelland (1986), it can be said that some of the families collected show some unproductivity, in the sense that many verbs are reused by others through use of prefixes. For example, the family of the verb “fazer” (to do) is composed only of verbs derived from it: “desfazer” (undo), “refazer” (redo), etc. This can be seen in other classes, such as the verb “cobrir” (to cover), “pedir” (to ask), “dizer” (to say), “ver” (to see), among others.

2.2.1 Types x Tokens

Another important issue regarding the corpus used, refers to the frequency in which the verbs are present in it. For this, we introduced the concepts of *word type* and *word token*Peirce, S. & Santiago (1906)). The term *type*, refers to the number of **unique** words present in a Corpus. The phrase "This phrase is an example phrase." Therefore has 6 *types*. The term *token*, in contrast, refers to the total number of terms present, including repetitions. In this case, the same example sentence has 7 *tokens*. Thus, treating verbs as *word tokens* would mean having a corpus that respects the frequency of use of verbs. Treating them as *word types* would mean just treating them as a list of verbs, without repetition.

Discussing the use of *word types* or *word tokens* in the constitution of the corpus is relevant, since the training of Neural Networks takes placecycles and we have already seen in Chap. 1 that the network learning consists of adjusting the connection weights. Thus, if the model is fed with verbs at different frequencies, it will tend to prioritize the adjustment of these connections (and therefore of learning as a whole) to the most frequent verbs.

For this research, we chose to use *word types*, i.e., all the verbs were treated equally and introduced the same number of times in model. [1]

2.3 Preprocessing for the Encoder-Decoder

In the case of the *Encoder-Decoder* architecture, unlike the architecture used by Rumelhart & McClelland, there is no reason to be concerned with the data sequence, neither during the insertion process nor in the prediction process. model. The reason for this will become clearer after the chapters 3 and 4. However, what can already be said is that there is no need to treat verbs in trigrams, neither in *Wickelfeatures*. In this type of architecture, verbs can usually be treated as a sequence of phones. However, even with this simplification, there are multiple ways to approach this issue. This work aims to study the relationships between verbs at a phonetic level, so it is important that the constructed vector representations take this into account.

We will start with the presentation of the International Phonetic Alphabet (IPA), presented in Tables 2.5 and 2.6. IPA is a phonetic notation system, created by the International Phonetic Association to promote a standardization in the transcription of data from different languages. It organizes symbols that represent sound units present in human languages based on the characteristics of execution of these units.

Table 2.5 gathers the set of consonant sounds and displays in the dimension of the columns the point of articulation of the sounds, that is, the point where a disturbance of the air passage occurs. The "Bilabial" column, for example, corresponds to the obstruction that occurs on the lips during the production of their respective phones. The lines make up the different possible *modes* of articulation, that is, the forms of obstruction of the air passage. The obstruction can be total as in [**p**] or partial as in [**s**]. Finally, within the same cell a sound may occur with or without the vibration of the vocal cords, as is the case with the pair [**p b**]. The symbol on the left represents the silent sound and the symbol on the right, voiced.

The vowels are organized in Table 2.6. The columns in this table refer to the place where the sounds are reproduced (with advance or retreat of the tongue), and the lines, at

[1] Evidence in the field of psycholinguistics (Bybee (1995); Cotterell (2001)) indicators that humans learn to generalize phonological patterns based on the count of *types of words*, ignoring the frequency of use of words.

the height of the tongue in relation to the roof of the mouth during the execution of the sound. When the symbols appear in pairs, the one on the right represents a rounded vowel (Seara, I. C. (2015)). As an example of how this table works, it is interesting to observe the execution of the vowel [**u**] and compare it with the vowel [**o**]. When pronouncing them for a comparison, the roundness of the lips in both is noted, however the height of the tongue during the execution of the second is slightly lower than the height of the first. To understand the matter of anteriority/posteriority, it is interesting to observe the movement (forward and backward) of the tongue during the execution of [**e**] and [**o**]. In this case, it is also noted that [**o**] has a rounding stroke, while [**e**] does not.

	Bilabial	Labiodental	Dental	Alveolar	P-alveo.	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɣ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	β			r					R		
Tap/Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lat. Fricative				ɬ ɮ							
Aprox.		ʋ		ɹ		ɻ	j	ɰ			
Lat. Approx				l		ɭ	ʎ	L			

Table 2.5: *Consonants IPA*

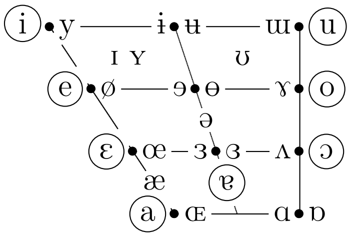

Table 2.6: *Vowels IPA*

The purpose of the IPA is to map and characterize all the sounds of human languages. In this way, each language takes advantage of only a subset of the IPA. The phones present in Portuguese-Brazilian are circulated. Taking into account only the subset of BP phones, and also keeping in mind that the phonetic alphabet symbols need to be transformed into ASCII code to be interpreted, it makes sense to build a new table adapted to the problem in question. The result of this adaptation can be seen in Tables 2.7 and 2.8. In these new tables, in addition to the exclusion of some points and modes of articulation (motivated by the very nature of BP), an alternative transcription key is also presented that includes only characters belonging to the ASCII code. With regard to the vowel table (2.8), the rounding dimension was dispensed since there was no longer any need for this markup as this information would be redundant due to the fact that all front vowels are rounded in Brazilian Portuguese and no back vowel is.

	Bilabial	Lab. dent.	Alveolar	P-alveo.	Velar
Plosive	p b		t d		k g
Nasal	m		n		N
Tap/Flap			r		
Fricative		f v	s z	x j	h
Lat. appr.			l		L

Table 2.7: *Consonants in the new representation*

	Front	Back
Close	i	u
Close-Mid	e	o
Open-Mid	E	O
Open	a	
Nasal	A (ã) 3 ($\tilde{e}$)	

Table 2.8: *Vowels in the new representation*

Table 2.9 displays the proposed transcription key using the phonetic tables created and Table 2.10 displays some examples of possible transcriptions. In order to facilitate preprocessing, some phonetic transcriptions were represented by the same symbol, such as: t ʃ et → t; u, w and ʊ→ u.

Phone (AFI) + Practical Example	Proposed Transcript
[**p**] - **p**arar	p
[**b**] - **b**otar	b
[**t**] - **t**ocar	t
[**d**] - **d**ançar	d
[**k**] - **c**asar	k
[**g**] - **g**ostar	g
[**f**] - **f**ugir	f
[**v**] - **v**oltar	v
[**s**] - **s**oltar	s
[**z**] - pre**s**enciar	z
[ʃ] - **ch**amar	x
[ʒ] - **j**antar	j
[**t**ʃ] - sen**t**ir	t
[**d**ʒ] - **d**izer	d
[**h**] - e**rr**ar	h
[r] - enca**r**ar	r
[**l**] - pu**l**ar	l
[ʎ] - espa**lh**ar	L
[**m**] - **m**orar	m
[**n**] - **n**adar	n
[ŋ] - so**nh**ar	N
[**a**] - p**a**rar	a
[**e**] - l**e**r	e
[ɛ] - esp**e**ro	E
[**i**] - r**i**r	i
[**o**] - pr**o**por	o
[ɔ] - col**o**co	O
[**u**] - c**u**rtir	u
[**ẽ**] - **e**ntreter	3
[**ã**] - pl**a**ntar	A
[**õ**] - comp**o**nho	o
[ʊ] - cas**o**	u
[**j**] - sa**i**o	i
[**w**] - vo**l**to	u

Table 2.9: *Dictionary for Proposed Transcripts*

Verb	Transcription
ressentir	hes3ntir
paro	paru
possuo	posuu
olha	oLa
sacudir	sakudir
voltar	voutar

Table 2.10: *Transcription Examples*

2.3.1 The Encoder-Decoder inputs

After the construction of the corpus, the transcribed verbs had their respective heads associated with the phonetic traits that compose them, the same traits that originated the Tables 2.7 and 2.8. Table 2.11 presents an example of this process for the verb "poder" ("can") - "*pOsu*" (I can).

Phone	Phonetic Features
p	bilabial, plosive, unvoiced
O	Open-Mid, back
s	fricative, alveolar, unvoiced
u	close, back

Table 2.11: *Phonetic features for the verb "Posso" (can)*

Note in the table 2.11 that it is possible to characterize each phone with three or two phonetic features. However, to build the vector representation for the model, we need to consider in advance all possible dimensions of the model, since each dimension will represent a single phonetic feature.

For the construction of the vectors, we will use *dictionaries*. In computing, a dictionary is a data storage structure that associates a key with a value. This structure has a mutually exclusive set of keys, each associated with a value. Thus, when querying a dictionary with a key value, the structure returns the associated value in response. In total, 20 dimensions are required to represent a headset. There are 18 to represent the features of the constructed phonetic tables (2.7 and 2.8): Plosive, Nasal, Tepe, Fricative, Lateral, Bilabial, Labio-Dental, Alveolar, Post-Alveolar Approximant , Velar, Glottal, Close, close-mid, Open-mid, Open, Front, Back; and another two to represent the beginning and end of the verb. The need for these last two dimensions will become clearer after Cap. 3.

The phones are then characterized by the presence (1) and absence (0) of the mentioned features. Thus, the built dictionary has phones on the key values and a list of 0's and 1's to represent the presence and absence of phonetic features. As seen, according to the phonetic tables developed, each phone can be described by three or two phonetic traits, so that each vector will have only three or two values marked as **1** 's. The representation of the beginning and end of the verb will be exceptions, with only a presence mark in the vector in the respective dimension.

Table 2.12 shows as an example a comparison between two similar phones (**p** and **b**) that are distinguished only by the phonetic "voiced" trait. The result is a vector representation that also carries this notion of proximity between the phones. The beginning of the verb (that is, before the first phone) is represented by a vector that marks 0 in all phonetic lines and 1 to represent the beginning (the ending can be understood in a similar way). In addition, it is worth noting that the proposed phonetic tables (2.7 and 2.8) add up to 28 possible phones. As we also have the start and end markings, we will have a total of 30 different vector representations available.

Feature	**p**	**b**
plosive	1	1
nasal	0	0
tap	0	0
fricative	0	0
l-approx	0	0
bilabial	1	1
labiodental	0	0
alveolar	0	0
p-alveolar	0	0
velar	0	0
glottal	0	0
voiced	0	1
close	0	0
close-mid	0	0
open-mid	0	0
open	0	0
front	0	0
back	0	0
<beg>	0	0
<eos>	0	0

Table 2.12: *Codification example of phones*

Putting together all the steps outlined, the complete process of transformation of the inputs can be summarized to:

1. Addition of symbols to mark the beginning and end of verbs.
2. Division of verbs in phones, following the proposed transcription key.
3. Transformation of the phones into *arrays* of 0's and 1's, following the developed dictionary of phones.

In short, we will have that the vectors of *input* are therefore binary vectors of 20 dimensions, containing from one to three dimensions occupied by **1** 's and the rest filled with zeros. The complete representation of the entire verb, in turn, will consist of the consecutive concatenation of these vectors.

Bibliography

Albright, B., A. & Hayes, “Rules vs. analogy in english past tenses: a computational/experimental study”, *Cognition*, volume 90:119–161, 2003. , 3

Bahdanau, Cho K. Bengio Y., D., “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, *arXiv e-prints*, 2014. 1, 13

Bybee, J. L., “Language and cognitive processes”, *Language*, volume 10:425–455, 1995. 20

Bybee, J. L. & Moder, C. L., “Morphological classes as natural categories. language”, *Language*, volume 59(2):251–270, 1983. 3

Chomsky, M., N. & Halle, *The sound pattern of English*, MIT Press, 1968/1991. , 1, 3, 4

Cotterell, Christo Kirov & Ryan, “Stochastic phonology”, *GLOT*, volume 5, 2001. 20

Cotterell, Ryan, Kirov, Christo, Sylak-Glassman, John, Yarowsky, David, Eisner, Jason & Hulden, Mans, “The sigmorphon 2016 shared task—morphological reinflection”, *in:* “Proceedings of the 2016 Meeting of SIGMORPHON”, Berlin, Germany: Association for Computational Linguistics, 2016. 13

Câmara Jr., J. M., *Estrutura da Língua Portuguesa*, 30 ed., Vozes, 1999. 14

Faruqui, et al., “Morphological Inflection Generation Using Character Sequence to Sequence Learning”, *arXiv e-prints*, 2015. 13

Goodfellow, Y. & Courville A., I. & Bengio, *Deep Learning*, MIT Press, http://www.deeplearningbook.org, 2016. 12

Guide, Bruno Ferrari, *Abordagem computacional para a questão do acento no português brasileiro*, Tese de Mestrado, University of Sao Paulo, 2016. 18

Kann, H, K. & Schütze, “MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection”, *in:* “Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology”, 62–70, Berlin, Germany: Association for Computational Linguistics, 2016. 13

Kirov, Christo & Cotterell, Ryan, “Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate.”, *Transactions of the Association for Computational Linguistics*, 6:651–665, 2018. , 3

Krizhevsky, Alex, Sutskever, Ilya & Hinton, Geoffrey E., “Imagenet classification with deep convolutional neural networks”, *in:* “Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1”, NIPS’12, 1097–1105, USA: Curran Associates Inc., 2012. 12

Liu, X. & Huang X., P. & Qiu, “Recurrent Neural Network for Text Classification with Multi-Task Learning”, *arXiv e-prints*, 2016. 12

Mackenzie, C.E., *Coded character sets: history and development*, The Systems programming series, Addison-Wesley Pub. Co., 1980. 16

Marcus, et al., G., *Overregularization in language acquisition. Monographs of the Society for Research in Child Development*, 1992. 7

Mitchell, R., *Web Scraping with Python*, O’Reilly Media, Inc., 2015. 18

Nakisa, Ramin Charles & Hahn, Ulrike, “Where defaults don’t help: the case of the german plural system”, *ArXiv*, volume cmp-lg/9605020, 1996. 12

Nelson, G., *English: an Essential Grammar*, Routledge., 2010. 3, 13

Peirce, C., S. & Santiago, “Prolegomena to an Apology for Pragmaticism.”, *The Monist*, volume 16:506, 1906. 19

Pinker, S., *The Language Instinct*, New York: NY: Harper Perennial Modern Classics, 1994/2007. 4

Pinker, S., *In Single Combat." Words and Rules: The Ingredients of Language*, Harper Perennial, 1999. 1, 3, 8, 17

Pinker, S. & Prince, A., “On language and connectionism: Analysis of a parallel distributed processing model of language acquisition”, *Cognition*, volume 28(2):73–193, 1988. , 1, 3, 8, 12

Plunkett, Kim & Marchman, Virginia, “From rote learning to system building: acquiring verb morphology in children and connectionist nets”, *Cognition*, volume 48(1):21 – 69, 1993. 12

Plunkett, R. C., K. & Nakisa, “A connectionist model of the arabic plural system”, *Language and Cognitive Processes*, volume 12(5-6):807–836, 1997. 12

Plunkett, V., K. & Marchman, “U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition.”, *Cognition*, volume 38(1):43–102, 1991. 8, 12

Prasada, S., S. & Pinker, “Generalisation of regular and irregular morphological patterns”, *Language and Cognitive Processes*, volume 8(1):1–56, 1993. 7

Rumelhart, D. E. & McClelland, J. L, *On learning the past tenses of English verbs*, Bradford Books/MIT Press, 1986. 1, 5, 6, 7, 8, 11, 13, 14, 15, 16, 17, 18, 19,

Schneider, W., “Connectionism: Is it a paradigm shift for psychology?”, *Behavior Research Methods, Instruments, & Computers*, volume 19(2):73–83, 1987. 8

Seara, et al., I. C., *Fonética e Fonologia do PortuguÊs Brasileiro*, editora contexto, 2015. 5, 21

Sutskever, I. et. al., "Sequence to Sequence Learning with Neural Networks", *arXiv e-prints*, 2014. 13

Westermann, R., G. & Goebel, "Connectionist rules of language.", 1997. 12

Wickelgren, W. A., "Auditory or articulatory coding in verbal short-term memory", *Psychological review*, volume 76:232–5, 1969. 16, 17

Wu, Y. et. al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", *arXiv e-prints*, 2016. 13

Wuerges, E. T., *A Aquisição da Morfologia Verbal por Crianças Falantes de Português Brasileiro e o Uso de Formas Variantes*, Tese de Mestrado, University of Sao Paulo, 2014. 9, 10

Appendix A

Phonetic Transcription by Rumelhart and McClelland

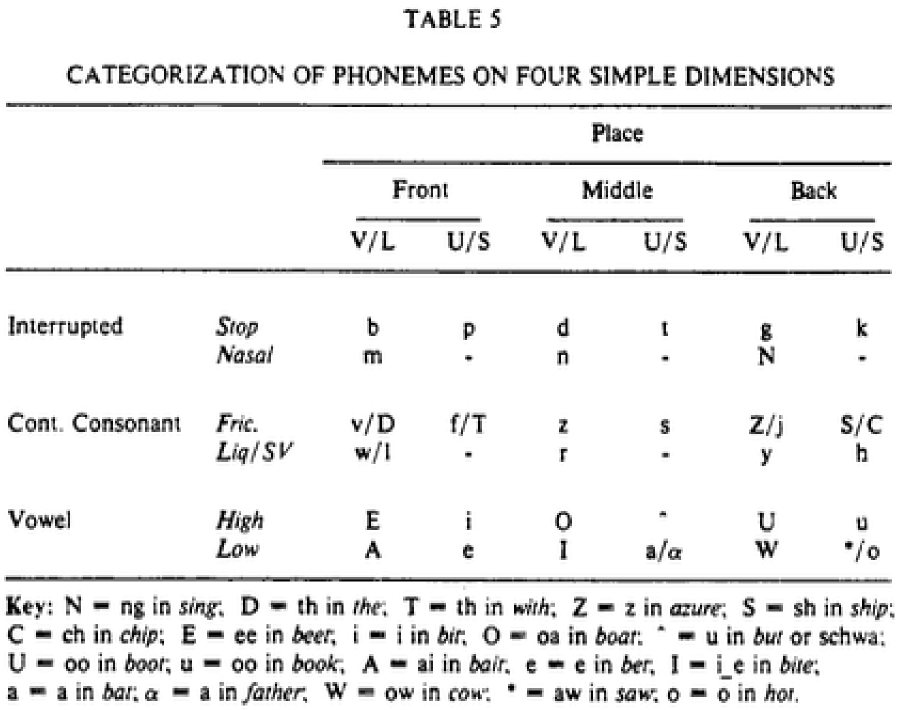

TABLE 5

CATEGORIZATION OF PHONEMES ON FOUR SIMPLE DIMENSIONS

		Place					
		Front		Middle		Back	
		V/L	U/S	V/L	U/S	V/L	U/S
Interrupted	*Stop*	b	p	d	t	g	k
	Nasal	m	-	n	-	N	-
Cont. Consonant	*Fric.*	v/D	f/T	z	s	Z/j	S/C
	Liq/SV	w/l	-	r	-	y	h
Vowel	*High*	E	i	O	^	U	u
	Low	A	e	I	a/α	W	*/o

Key: N = ng in *sing*; D = th in *the*; T = th in *with*; Z = z in *azure*; S = sh in *ship*; C = ch in *chip*; E = ee in *beer*; i = i in *bit*; O = oa in *boat*; ^ = u in *but* or schwa; U = oo in *boot*; u = oo in *book*; A = ai in *bait*; e = e in *bet*; I = i_e in *bite*; a = a in *bat*; α = a in *father*; W = ow in *cow*; * = aw in *saw*; o = o in *hot*.

Figure A.1: *Preprocessing Table taken from Rumelhart & McClelland [1986] (page 235)*