Development and Improvement of Phonetic Transcribers for Catalan Applied to Speech Synthesis and Recognition
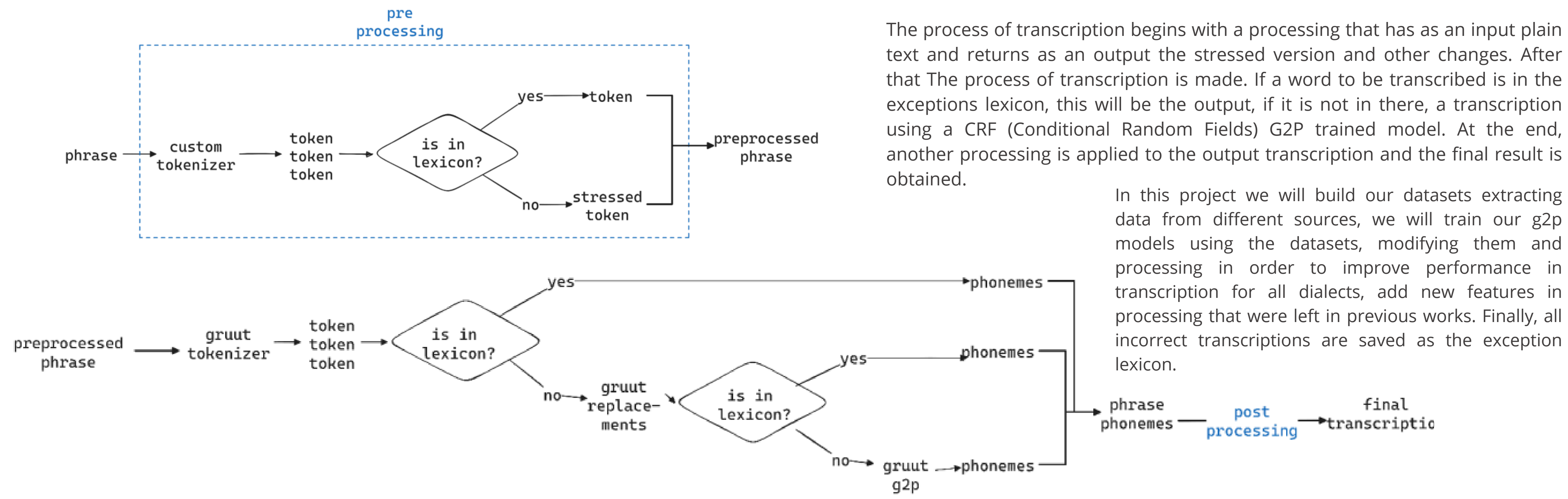
By Miquel Sánchez Queralt
Supervisor: Ignasi Esquerra Llucià

Abstract

Phonetic transcription is a process involved and used in fresh applications such as text-to-speech (TTS) and automatic speech recognition (ASR). This project aims to improve previous work about automatic phonetic transcription in Catalan language. In this version more research has been made in different dialects since the previous work was essentially made for Central Catalan.

Gruut will be used, an open-source library designed to aid in speech recognition and synthesis applications that includes deep learning techniques and different processing paths to obtain a phonetic transcription from a text input.

1 Introduction

The process of transcription begins with a processing that has as an input plain text and returns as an output the stressed version and other changes. After that The process of transcription is made. If a word to be transcribed is in the exceptions lexicon, this will be the output, if it is not in there, a transcription using a CRF (Conditional Random Fields) G2P trained model. At the end, another processing is applied to the output transcription and the final result is obtained.

In this project we will build our datasets extracting data from different sources, we will train our g2p models using the datasets, modifying them and processing in order to improve performance in transcription for all dialects, add new features in processing that were left in previous works. Finally, all incorrect transcriptions are saved as the exception lexicon.

2.Dataset

Different sources have been used for obtaining data. Wictionary [1], and Festcat [2] for both words and transcription, and Espeak [3], a rule based transcriptor that can obtain the transcription of the previous extracted words.

After having all the words transcribed in the different versions, a first processing has been made in order to change them all to the same format. Then coincidence criteria have been used, the words that had the same transcription in all the systems were added to the initial datasets. Then some discrepancies were solved, choosing arbitrarily one of the results following notes in previous works or languages rules. Finally, in the cases where 2 out of 3 transcriptions were the same, this was also added to the datasets.

In case there is a problem with this selection, we are able to tell it from the output in the trained model.

"Exemple de transcripció generada pel model i excepcions del dialecte balear amb processament"

"{ə d z ˈe m p l ə}{d ə}{t r ə n s k r i t s i ˈo}{ʒ ə n ə r ˈa ð ə}{p ə l}{m o ð ˈɛ l}{i}{ə t s ə t s i ˈo n **z}{d** ə l} **[**d i ə l ɛ ˈt t ə**]**{b ə l e **ˈa}{m** b}{p r o s ə s ə m ˈe n t}"

3.Training models and modifyng processing

In this section of the project, we have analised the files that contained the words that do not have an exact coincidence between the expected and generated transcription. We group the types of error, and we try to correct it, even changing the dataset (characters distributions in languages are obviously far from uniform), there are combinations that appear more frequently than others so we need to add transcriptions with words that contain them. If this strategy does not work, the origin of the problem is searched in the code, and if found is changed, if it is not solved it will be included in the lexicon. Pre-processing and post-processing functions and classes will also be modified in order to include new features left from previous words, such as contact between words effects in all dialects.

4. Results

As it can be seen in the first figure, the WAR increases in each iteration of the improvement for all dialects. The growth is faster at first, solving the most frequent errors but becomes slower at lasts iterations since problems are small and not so frequent.

The other measures also grow, but the iteration where the growth is bigger changes depending on the dialect. Generally good results have been achieved and differences between performance in dialects, at first, have been reduced.
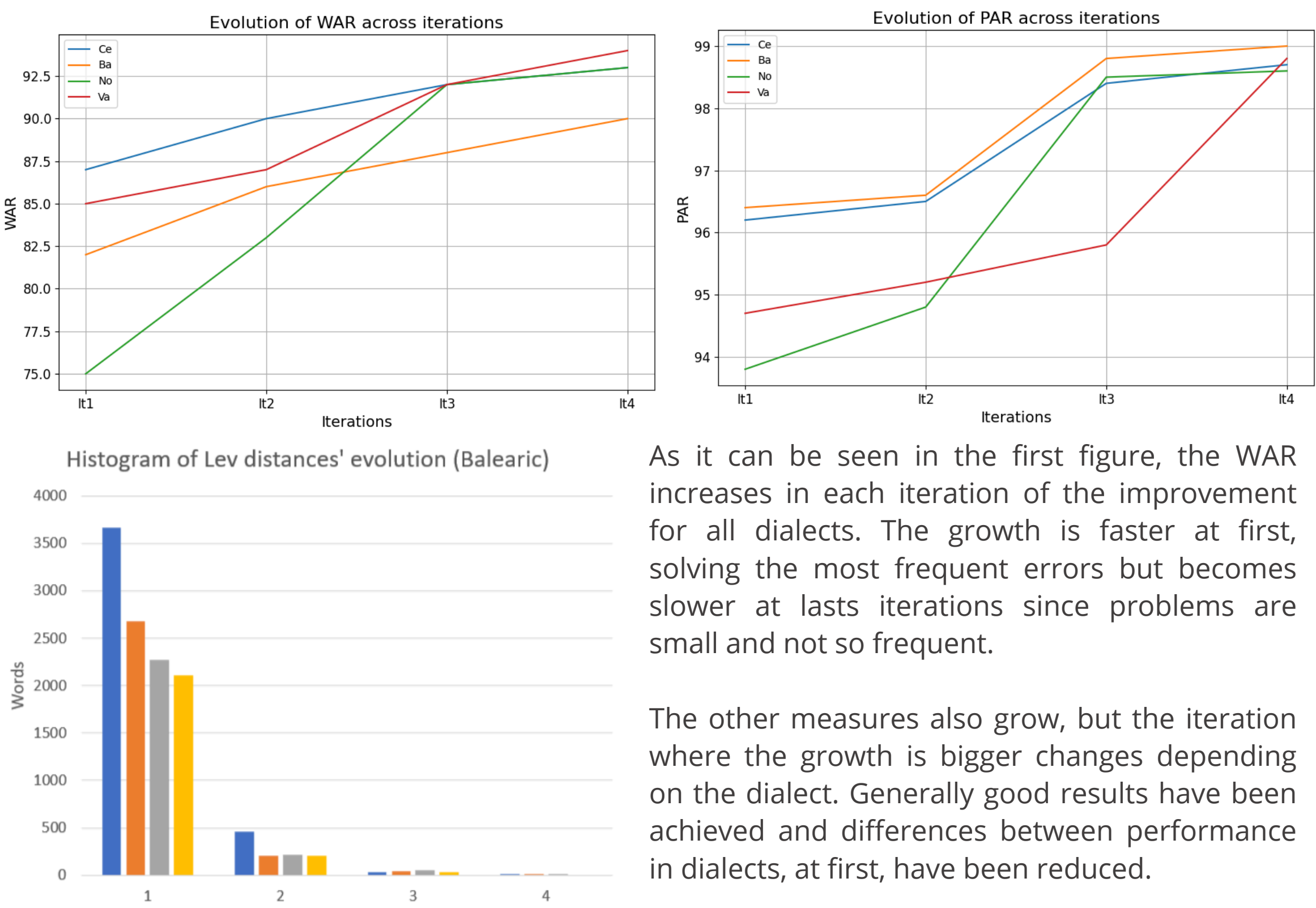
5. Conclusions

A satisfactory result has been achieved, having improved the performance of the g2p models in all dialects, and obtained a good performance in Word Accuracy Ratio and Phoneme Accuracy Ratio which is crucial in order to do speech recognition and synthesis.

The problems left in the previous works that were solvable (taking in account time, resources, the need of category detection models) were fixed with exit. Other indications for later approaches were also left.

References

[1] Wiktionary. (n.d.). Category terms with IPA pronunciation
[2] Festcat - Síntesi de la parla en català | corpus upc_ca_prompts-1.0.tar.bz2.
[3] Espeak text to speech | eSpeak. Available at: https://espeak.sourceforge.net/