Preprocessing models for speech technologies: The impact of the Normalizer and the Grapheme-to-Phone on hybrid systems

Bruna Carriço[1], *Christopher Shulby*[2], *Helena Moniz* [1,3]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]Defined.ai, Seattle WA, United States
[3]INESC-ID, Lisboa, Portugal

Abstract

This paper describes the linguistic preprocessing methods on hybrid systems provided by an Artificial Intelligence (AI) international company, Defined.ai. The startup focuses on providing high-quality data, models, and AI tools. The main goal of this work is to enhance and advance the quality of preprocessing models by applying linguistic knowledge. Thus, we focus on two introductory linguistic models in a speech pipeline: Normalizer and Grapheme-to-Phone (G2P). To do so, two initiatives were conducted in collaboration with the Defined.ai Machine Learning team. The first project focuses on expanding and improving a European Portuguese Normalizer model. The second project covers creating G2P models for two different languages – Swedish and Russian. Results show that having a rule-based approach to the Normalizer and G2P increases its accuracy and performance, representing a significant advantage in improving Defined.ai tools and speech pipelines. Also, with the results obtained on the first project, we improved the normalizer in ease of use by increasing each rule with linguistic knowledge. Accordingly, our research demonstrates the added value of linguistic knowledge in preprocessing models.

Keywords: Speech technologies, Normalizer, Grapheme-to-Phone, linguistic knowledge, models.

Resumo

Este artigo descreve os métodos de pré-processamento linguístico em sistemas híbridos fornecidos por uma empresa internacional de Inteligência Artificial (IA), a Defined.ai. A *startup* concentra-se em fornecer dados, modelos e ferramentas de IA de alta qualidade. O objetivo principal deste trabalho é aprimorar e avançar a qualidade dos modelos de pré-processamento aplicando conhecimento linguístico. Assim, focamo-nos em dois modelos linguísticos introdutórios numa arquitetura de fala: o Normalizador e o Grafema-para-fone (G2P). Para isso, foram realizadas duas iniciativas em colaboração com a equipa de *Machine Learning* da Defined.ai. O primeiro projeto concentra-se em expandir e melhorar um modelo de Normalizador para o Português Europeu. O segundo projeto cobre a criação de modelos G2P para duas línguas – Sueco e Russo. Os resultados mostram que ter uma abordagem baseada em regras para o Normalizador e G2P aumenta a sua precisão e o seu desempenho, representando uma vantagem significativa na melhoria das ferramentas e das arquiteturas de fala da empresa. Além disso, com os resultados obtidos no primeiro projeto, melhoramos o normalizador em termos de facilidade de uso, aumentado cada regra com conhecimento linguístico. Dessa forma, a nossa pesquisa demonstra o valor do conhecimento linguístico em modelos de pré-processamento.

Palavras-chave: Tecnologias de fala, Normalizador, Grafema-Fone, conhecimento linguístico, modelos.

1. Introduction

Data preprocessing is a crucial step in building a machine learning model. If data is preprocessed, the results are consistent and of high quality (García et al., 2016). For example, modern speech recognition systems model linguistic entities at multiple levels, sentences, words, phones, and other units, using various statistical approaches (Jurafsky & Martin, 2022). The parameters of these models are usually trained on data, but their accuracy attempts to capture linguistic knowledge.

This text discusses the importance of data preprocessing in building high-quality machine learning models, particularly in speech recognition systems. The Linguistic Processing Module provided by Defined.ai is described, which includes the Normalizer and Grapheme-to-Phone pipelines. Two projects conducted by the Machine Learning Team at Defined.ai have also been discussed: Normalizer expansion for European Portuguese and Grapheme-to-Phone model creation for Swedish and Russian. The study aims to gain insight into how linguistic knowledge impacts speech technologies and how to upgrade and validate Normalizer and G2P models in different languages.

2. Speech technologies: Automatic Speech Recognition

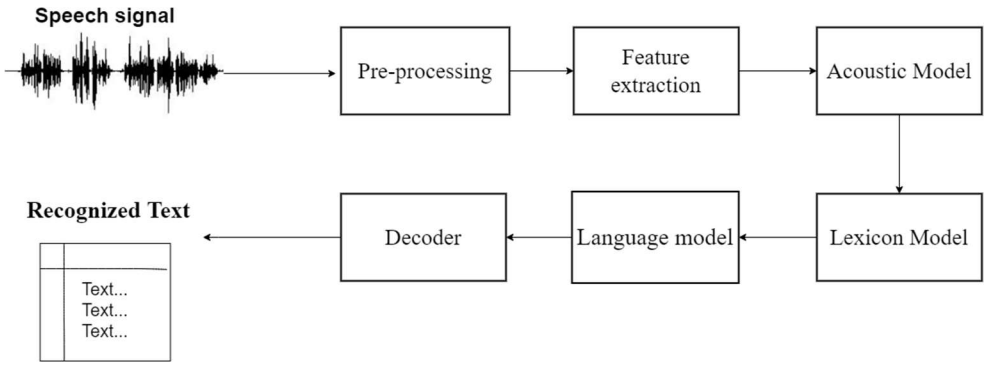
As stated by Alasadi and Deshmukh (2018), speech recognition is an important field that is constantly being developed as an interdisciplinary subfield of computational linguistics. Speech recognition creates technology and methods that empower the acknowledgment and understanding of natural language into a computer-understandable language. This is most generally known as Automatic Speech Recognition (ASR).

Automatic Speech Recognition has its origins back in the 1950s. Early ASR systems had a limited lexicon and were focused on numbers. In 1966, Hidden Markov Models (HMM) became a breakthrough in ASR and have remained state of the art for some time (Hennebert et al., 1994). Over the years, ASR technology has become more reliable and easier to handle due to computer technology and informatics advancements. By the 2000s, speech recognition technology had reached an accuracy rate of approximately 80%, and commercial applications, such as Siri and virtual smartphone assistants, became highly popular.

Classical machine learning models can be classified into two different approaches: Generative[1] and Discriminative[2] approaches. Speech recognition has mostly used the Generative approach in the past decades, and two popular methods based on this approach are HMMs and Gaussian mixture models (GMMs). The basic framework of a speech recognizer system includes three major stages: capture, transducing, and decoding. The acoustic model, lexicon model, and language model are combined during the transducing step. A typical ASR framework includes the components described in Figure 1.

[1] Generative approach: used to learn each language and determine which language the speech belongs to.
[2] Discriminative approach: used to determine the linguistic differences without learning any language.

Figure 1. Basic framework of a speech recognizer system



The field of automatic speech recognition has seen two main approaches over the years: a traditional hybrid approach and an end-to-end (E2E) deep learning approach. The hybrid approach has been dominant for the past decade due to extensive research and training data available. However, the end-to-end approach has gained popularity due to its simplicity, reduced training and decoding time, and comparable accuracy to the hybrid approach (Vielzeuf & Antipov, 2019).

On the other hand, in accordance with (Kurata et al., 2019) an end-to-end Deep Learning approach is a new paradigm in neural network-based speech recognition that has several advantages. Traditional hybrid ASR systems, which consist of an acoustic model, a language model, and a lexicon model, each of which might be sophisticated, require independent training of these components. In contrast, E2E ASR is a unified method with a much simpler training pipeline and models that perform at low audio frame rates. This reduces the amount of time spent on training and decoding. A common end-to-end Deep Learning architecture is the encoder-decoder, which is implemented with RNNs (Recurrent Neural Network).

Errattahi and El Hannani (2017) concluded that the benchmarks have shown that the Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) have proven their efficiency on several Large Vocabulary Continuous Speech Recognition (LVCSR) tasks by outperforming the traditional Hidden Markov Models. LVCSR is characterized by the authors as being the most challenging task in ASR. Recently, Rajadnya (2020) developed an application for continuous speech recognition based on DNN-HMM and Deep Belief Network (DBN) algorithms. This proved that using DNN-HMM with DBN supplies better accuracy than using typical GMM-HMM systems. Deep learning is quickly becoming a standard approach for speech recognition, having effectively replaced all the classical approaches.

In recent years, ASR systems have significantly improved in accuracy, but they still cannot reach human-level accuracy. Factors such as speaker-dependent or speaker-independent models, acoustic models, vocabulary, and language models can influence the performance of ASR systems. System failures can also occur due to issues such as noisy environments and different pronunciations by speakers.

Recently, bias and gender issues in data have also deserved attention from the community. Gender and racial bias are a concept where models and algorithms do not provide optimal services to people of a specific gender or dialect. The disparity of accessible data for both genders and dialects has been proven in recent studies to be its main cause. Brasoveanu and Dotlacil (2020) found bias against women in the performance of speech recognition systems by analyzing the gender representation in different corpora. Bias was also identified against dialect groups; dialect speakers have lower ASR performance than speakers of standard pronunciations (Wassink et al., 2022).

3. Linguistic preprocessing

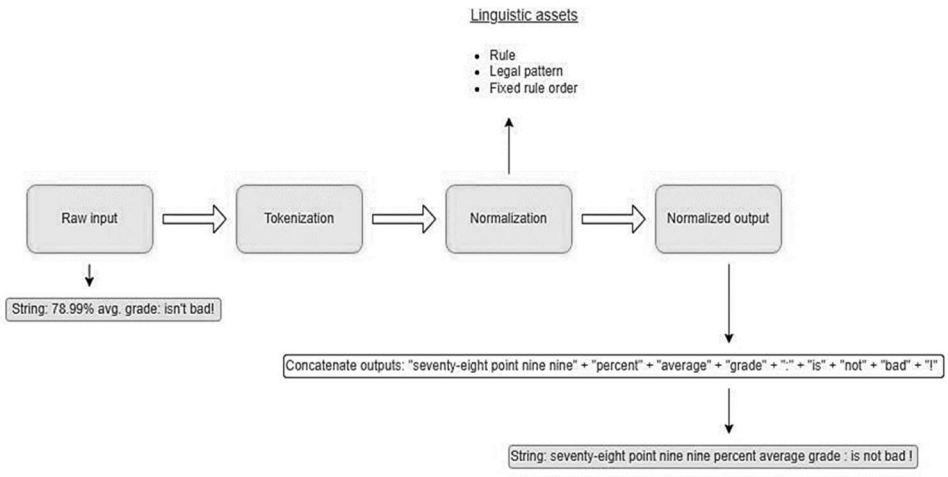
Defined.ai uses a Linguistic Processing Module (LPM) to treat raw text as input and structure it to make it usable for speech technologies that require linguistic knowledge. The LPM is used in many technologies, such as ASR and Text-to-Speech (TTS) systems. Therefore, LPM includes the development of pronunciation lexicons that will provide a mapping between a word's orthographic form and its pronunciation form. Pronunciation lexicons are mainly used to build G2P models to provide pronunciation of OOV words.

In the LPM module are included the Grapheme-to-Phone (G2P) and the Normalizer pipelines. These introductory linguistic models are essential to ensure the quality of data processing and measure its quality. A crucial step since it ensures high-quality data processing throughout the pipelines. By describing these models, we explain the process of converting written text into its spoken form and how we transform graphemes into phonetic transcription, in a stepwise perspective.

Such linguistic models are frequently used as pre-requirements for the preparation of any corpus for ASR systems. At Defined.ai these pre-requirements are one of the company's initiatives that support ASR systems and DefinedData (one of Defined.ai products). The Machine Learning (ML) team ensures the creation and assistance of in-house G2P conversion and Normalization solutions. Consequently, the outcome of this initiative is to create a more robust data collection pipeline and to develop pre-requirements for training data.

The Normalizer pipeline (Figure 2) converts written text into a spoken form that can be used for ASR systems as training data. The normalizer pipeline consists of two main steps - text preprocessing and normalization. Text preprocessing involves cleaning the text by removing unwanted characters and normalizing punctuation, whitespaces, and Unicode characters. Next, tokenization separates the raw input into tokens by considering non-printable characters, spaces, and logical semantic breaks. Normalization involves applying rules to each word of the sentence to produce the spoken form output. The current Normalizer supports 12 natural languages, including English, Portuguese, German, and French, and provides normalized forms for numbers, dates, measurements, time, durations, and symbols. The normalization process is rule-based, making it possible to predict the output of the Normalizer accurately, but new rules can be written to cover new constructions.

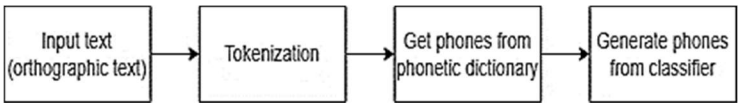
Figure 2. Normalizer pipeline



The G2P pipeline, illustrated in Figure 3, is used to predict the phonetic transcription of a given written word to create and improve ASR systems. The pipeline requires Tokenization and checks vocabulary before

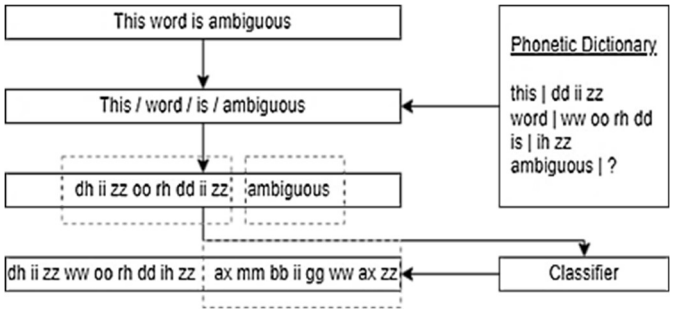
extracting the pronunciation of each word from the lexicon. If a word is not recognized by the lexicon, the Classifier predicts its phonetic transcription.

Figure 3. Grapheme-to-Phone pipeline



The process of pronunciation prediction, shown in Figure 4, starts with the tokenization of orthographic text as input. In the first line, we can see the orthographic text as the input, which is tokenized. After that, we will get the corresponding phonetic transcription of each word that was taken from the pronunciation lexicon. If a word is not recognized by the lexicon (e.g., ambiguous) the Classifier will then predict the phonetic transcription of that given word. The G2P currently supports 18 languages, including English, Portuguese, and Japanese.

Figure 4. Pronunciation prediction process



4. **Phonetic and phonological concepts**

In this section, we introduce the main phonetic and phonological aspects of two distinct languages – Swedish and Russian. We start by introducing the main linguistic concepts used to make a description of each language: phoneme, phone, free variation, and contextual variation.

A phonological segment, also known as a phoneme, is the smallest distinctive unit of sound in a language that can change the meaning of a word. Phonemes are abstract representations of speech sound that exist in the mental grammar of speakers of a particular language. For example, in English, the sounds represented by the letters *p* and *b* are considered different phonemes because they can change the meaning of a word (e.g., *pat* vs. *bat*), even though the actual acoustic properties of the sounds may vary depending on their context or speaker. A phone, on the other hand, is a physical realization of a speech sound. It refers to the actual sound produced by a speaker, including all the variations in pronunciation due to factors like accent, speaking rate, and context.

Free variation refers to a linguistic phenomenon where two or more different sounds can occur in the same linguistic context without affecting the meaning of a word or utterance. In other words, when sounds are in free variation, they can be used interchangeably by speakers without any change in the word's meaning or grammatically. On the other hand, contextual variation is another linguistic phenomenon where the pronunciation of a sound form depends on its position within a word or sentence or its surrounding linguistic context. Contextual variation is typically rule-based and can affect the meaning or grammaticality of words. Resulting from contextual variation of Russian and Swedish, the following processes were deemed important: devoicing at the end of a word, assimilation of voicing between two obstruents.

For the context of this work, we focus mainly on phonetic aspects of such languages since we aim to describe and identify speakers' real pronunciations to implement this knowledge on the G2P model. Thus, this

chapter will mostly cover phonetic transcription and phonological processes that were strictly chosen regarding their relevance to the G2P. In the process of determining the selection criteria, the following factors were considered: (1) Frequency and occurrence – prioritize phonological processes and transcription conventions that are commonly encountered in real-world languages and writing systems; (2) Phonemic variability – consider phonological processes that involve phonemic variability and may pose challenges for G2P conversion. These processes can include allophony, assimilations, and neutralization, which can affect how graphemes are pronounced, and (3) Cross-linguistic applicability – consider phonological processes and transcription conventions that have applicability across multiple languages or writing systems, as they can enhance the versatility of the G2P system.

We begin with a key element of both speech recognition and text-to-speech systems: how words are pronounced in terms of individual speech units called phones. Following (Jurafsky & Martin, 2022) a speech recognition system needs to have a pronunciation for every word it can recognize, and a text-to-speech system needs to have a pronunciation for every word it can say. We model the pronunciation of a word as a string of symbols that stand for phones or segments. A phone is a speech sound and phones are represented with phonetic symbols that bear some resemblance to a letter in an alphabetic language, like English or Portuguese. In the context of this work, we use two different alphabets for describing phones: the International Phonetic Alphabet (IPA),[3] and the DC-Arpabet[4] (phonetic alphabet created by Defined.ai based on the arpabet alphabet). IPA was created in the late 19th century to describe the sounds of all human languages while using a set of transcription principles and its phones. On the other hand, the DC-Arpabet was specifically designed by Defined.ai for US English, while using ASCII symbols. Table 1 shows some examples of an American English phone set using DC-Arpabet symbols for transcribing its consonants, semivowels, vowels, and diphthongs together with their IPA equivalents.

Table 1. Examples of the American English phone set using IPA, and DC-Arpabet symbols

Phonetic Alphabet Symbols		**Classification**
IPA	**DC-Arpabet**	
[b]	[bb]	consonant
[ð]	[dh]	consonant
[s]	[ss]	consonant
[ʃ]	[sh]	consonant
[ʒ]	[zh]	consonant
[j]	[yy]	semivowel
[w]	[ww]	semivowel
[ə]	[ax]	vowel
[ʊə]	[ux]	diphthong

The DC-Arpabet currently contains 239 indexed unique symbols, each composed of at least 2 alphanumerical characters, plus an optional 1 to 3 alphanumerical characters. The symbols are represented by '##' and '_', and whenever an audio includes silence, short pause, unintelligible sounds, vocal noise, or music we represent it using the respective abbreviation shown in Table 2. Building the G2P model for a language also requires a list of characters used to spell that language, that is, its graphemes.

[3] Armstrong and Meier (2005).
[4] Shoup (1980).

Table 2. Symbols and non-phonetic sound representation

Phonetic Alphabet Symbols			Classification
IPA	**DC-Arpabet**	**X-SAMPA**	
##	##	##	sentence break
	_		skipped phone
	Sil		silence
	Sp		short pause
	Zun		unintelligible
	Zvn		vocal noise
	Zmu		

4.1. **Swedish**

This section deals with the phonetics and phonology of Standard Swedish from a synchronistic perspective. We want to focus on describing the authentic pronunciation of speakers. First, we decided to choose one single variant among all existing ones in the Swedish language – Standard Swedish. This variant evolved from the Central Swedish dialects, and most Swedes speak it. Herewith we can predict how most speakers pronounce Swedish. Thus, in this section, we start by briefly describing the Swedish vowel and consonant inventory, where we discuss phenomena crucial to measuring the impact on the G2P. Following, we introduce two frequent features of this language – quantity and retroflexion – which were important in the decisions we had to make for the Swedish G2P model. Also, resulting from the contextual variation of Swedish, the following processes were deemed important: neutralization, vowel reduction, and vowel lengthening.

The Swedish orthographic alphabet consists of nine vowels: <a, e, i, o, u, y, å, ä, ö>. Riad (2014) considers nine distinct vowel phones. Each vowel occurs with long and short variants. The author finds these variants allophones of the same phoneme. The long vowels are [iː, yː, eː, ɛː, øː, ʉː, uː, oː, ɑː] and the short vowels [ɪ, ʏ, ɛ, ø, ɵ, ʊ, ɔ, a], respectively, in Table 3.

Table 3. Standard Swedish orthographic vowels

Orthography	IPA		
	Phoneme	**Long vowel**	**Short vowel**
<i>	/i/	[iː]	[ɪ]
<y>	/y/	[yː]	[ʏ]
<e>	/e/	[eː]	[ɛ]
<ä>	/ɛ/	[ɛː]	[ɛ]
<ö>	/ø/	[øː]	[ø]
<u>	/ʉ/	[ʉː]	[ɵ]
<o>	/u/	[uː]	[ʊ]
<å>	/o/	[oː]	[ɔ]
<a>	/ɑ/	[ɑː]	[a]

In Standard Swedish, several vowel qualities are distinguished in unstressed syllables. The phones /e/ and /ɛ/ neutralize in the short variant, as [ɛ]. The alternations between long and short vowels provide important cues to the phonemic system. Thus, Riad (2014) defends that the lowering of /ø/, and /ɛ/ before a retroflex motivates

the height separation between /a/, which is [low], and /ø/ and /ɛ/ which are [mid]. The two main short allophones of /e/ and /ɛ/ are neutralized as [ɛ]. This results in eight short vowel allophones to match the nine long vowel allophones. However, many dialects have nine long vowels and nine short vowels (Riad, 2014). In some varieties of Standard Swedish, a similar neutralization occurs for short /ø/, where some young speakers have neutralization as [ɵ].

Also, in many cases <e> and <ä> as in *sett* and *sätt* coincide and are both pronounced /e/. This can lead to the mistaken belief that there are only eight short vowels and that [e] and [ɛ] are allophones. Yet, in Standard Swedish, /e/ and /ɛ/ are treated as phones in Standard Swedish. See the vowel phonetic inventory used for Standard Swedish in Table 4.

Table 4. Standard Swedish phonetic vowels and respective examples

Phonetic symbols		**Example (DC-Arpabet)**
IPA	**DC-Arpabet**	
[a]	[aa]	[kk **aa** ll ee nn]
[ɑ:]	[ah]	[**ah** vv ss nn ih tt]
[ɛ]	[eh]	[bb aa ng kk **eh** rr]
[æ:]	[ae]	[ll vv **ae** gg aa rr]
[e:]	[ee]	[bb ll **ee** kk tt]
[ɪ]	[ih]	[ss **ih** vv ii ll aa]
[i:]	[ii]	[dd ee ll tt **ii** dd]
[ɔ]	[oh]	[ee gg **oh** nn]
[o:]	[oo]	[bb **oo** gg eh]
[œ]	[oe]	[ff **oe** ll yy eh tt]
[ø:]	[eu]	[hh **eu** rr ss aa mm mm aa tt]
[ʊ]	[ug]	[yy **ug** nn ss oh nn]
[u:]	[uu]	[kk **uu** nn kk uo rr ss]
[ʏ]	[iu]	[kk ll **iu** ff tt aa]
[y:]	[iy]	[cc **iy** ll rr uo mm]
[ɵ]	[uo]	[ll **uo** dd vv ih gg]
[ʉ]	[uc]	[mm ih nn **uc** tt eh nn]
[j]	[yy]	[oo gg **yy** ih ll tt ih kk tt]

The Swedish orthographic alphabet consists of 14 consonants: <b, d, f, g, h, l, m, n, p, q, r, s, t, v>. Table 5 shows Standard Swedish orthographic consonants and their respective pronunciations. On the other side, the consonant phonetic inventory contains 17 different phones (see Table 8).

Table 5. Standard Swedish orthographic consonants

Orthography	Phonetic symbols	
	IPA	DC-Arpabet
<b>	[b]	[bb]
<d>	[d]	[dd]
<f>	[f]	[ff]
<g>	[g]	[gg]
<h>	[h]	[hh]
<l>	[l]	[ll]
<m>	[m]	[mm]
<n>	[n], [ŋ]	[nn], [ng]
<p>	[p]	[pp]
<k>	[k]	[kk]
<r>	[r], [ʂ]	[rr], [rr ss]
<s>	[s], [ɧ]	[ss], [shx]
<t>	[t]	[tt]
<v>	[v]	[vv]

In contrast to vowels, the opposition length of consonants does not carry any meaning. The consonant system has a double specification of aspiration and voicing in the obstruents. Riad (2014) focuses on the qualitative contrasts of the Swedish consonant system, arranged according to the place of articulation and manner of articulation (see Table 6).

Table 6. Standard Swedish consonants inventory modified and extracted from Riad (2014, p. 49)

	labial,	dental,	alveolar,		
	labiodental	alveolar	palatal	velar	glottal
oral stop	p	t		k	
	b	d		g	
fricative.	f	s	ç		h
fricative/retroflex			ʂ		
fricative/approximant	v				
nasal stop	m	n		ŋ	
Lateral		l			
apical trill		r			

Standard Swedish has palatal and velar voiceless fricatives: /s/, /ʂ/, /ɕ/. The phones [ɧ] and [ɕ] are very similar, although the most prominent phonetic difference lies in its place of articulation. While /ɕ/ has a stable place of articulation, /ʂ/ is subject to contextual conditioned allophony in Standard Swedish, as well as a wide-ranging sociolinguistic variation. In onset position, /ʂ/ can have four different variations – [ɧ], [ɧʷ], [x], [ʂ] –

while in postvocalic position the realization of /ʂ/ is mostly [ʂ] or [ʂː]. Thus, the most common prevocalic realization is [ɧ]. /ʃ/ is also used in some Swedish dictionaries (*e. g.*, *Svensk Ordbok*[5]).

One of the most salient features of North Germanic standard varieties is the quantity system. Stressed syllables are invariably heavy, due to a prosodic condition. This condition is met in either the vowel alone or in a combination of the vowel and the following consonant. In a stressed syllable one segment must be long, either the vowel or the consonant, but not at the same time. Vowels and consonants thus occur in long and short variants, and it is primarily in terms of quantity that these segmental distinctions are made and described.

There are qualitative differences within vowel pairs. Each long vowel has a short counterpart. The long and short consonants in a pair are naturally much more similar in quality. Most consonants have long and short pairs, but there are a few that exhibit a defective quantitative distribution. Two phones never occur directly after a stressed vowel, namely /h/ and /ɕ/, and hence lack long variants altogether. The segments /j/ and / ŋ/, on the other hand, are always long in a postvocalic coda position, provided that the syllable is stressed. The phoneme /ŋ/ never occurs word-initially, but may occur intervocalically and as onset in unstressed positions.

Syllable weight in stressed syllables is phonetically and phonologically clear. Any stressed syllable is bimoraic, where a long vowel is bimoraic, and a short vowel monomoraic. A long consonant is (mono)moraic, and a short consonant is non-moraic. If the vowel is long, then all is fine. If the vowel is short, then the following consonants must be long. In accordance with (Riad, 2014) most of the other Germanic languages lost consonant quantities early on and this has led to rather different quantitative phonologies. In this section, we shall assume that quantity is distinctive in consonants, but some consonants have lexical length, while others become grammatically lengthened or shortened by syllabification. For the vowels, quantity is predictable from prosodic context, when a syllable is stressed or when there is quantitative information (*e. g.*, lexical or positional) in the following consonant.

Another striking feature of Standard Swedish is retroflexion. The retroflexion rule shown in Table 7, creates retroflex sound when two contiguous segments (/s, t, d, n, l/ with a preceding /r/) converge into one element. The output /ʂ, ʈ, ɖ, ɳ, ɭ/ is phonologically distinct from the input segments. Thus, retroflex consonants can appear in most simple words (*e.g.*, *framfar**t**, rampaging),* but can also occur in other articulatory patterns – word boundaries, inflections, compounds, and derivations. In word boundaries, a retroflex consonant emerges if the final letter of a word is an <r> and the initial letter of the following word is <t, d, s, l, n>.

Table 7. Standard Swedish retroflexion rule

Swedish form	**Phonetic transcription**	**English translation**
vår triumf	/voːʈriɵmf/	our victory
hur mår du	/hʉːrmoːɖɵ/	how are you
under sängen	/ɵndeʂɛŋen/	under the bed
eller nej	/ɛleɳɛj/	or not
hur ledsam	/hʉːlesam/	how sad

As for flections, when the genitive <s> is attached to a word ending with <r>, the retroflex /ʂ/ is used (e.g, Peters hus, /peteʂhʉːs/, Peter's house). When a verb ends with a final <r> the retroflex consonants /ɖ/, /ʈ/ occur (e.g., stör-de, /støːɖ /; stör-t, /støːʈ/). Furthermore, this rule also applies to past participles and nouns. Thus, retroflex consonants also occur in compound words (e. g., vårdag, /voːɖɑːg /, spring day) and derived words (varsam, /vɑːʂam/, careful). The phonetic consonants for Standard Swedish are listed in Table 8 along with some examples.

[5] https://svenska.se

Table 8. Standard Swedish phonetic consonants and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[b]	[bb]	[**bb** ee ff uu gg aa dd]
[d]	[dd]	[**dd** ee ll tt uu gg]
[f]	[ff]	[**ff** aa mm ih ll yy]
[g]	[gg]	[**gg** rr oe nn tt]
[h]	[hh]	[**hh** ug tt eh ll]
[l]	[ll]	[**ll** ii nn dd rr ih gg]
[m]	[mm]	[**mm** oh ng aa]
[n]	[nn]	[**nn** ae rr aa]
[ŋ]	[ng]	[rr ii **ng** nn ee rr]
[p]	[pp]	[oe **pp pp** eh nn]
[k]	[kk]	[bb eh **cc** eh nn eh rr]
[r]	[rr]	[**rr** eu kk]
[ʂ]	[rr ss]	[bb aa **rr ss** eh bb eh kk]
[s]	[ss]	[**ss** vv oo rr]
[ɧ]	[shx]	[aa gg aa rr **shx** ih ss mm]
[t]	[tt]	[**tt** eh nn dd eh]
[v]	[vv]	[**vv** ae rr dd]

4.2. **Russian**

This section deals with the phonetics of Standard Russian from a synchronistic perspective. We start by briefly outlining the Russian vowel inventory and consonant inventory. Thus, resulting from the contextual variation of Russian, the following processes were deemed important: neutralization, devoicing at the end of a word, assimilation of voicing between two obstruents, and palatalization.

Finally, we describe and discuss the value of two special signs in Russian phonetic and phonological system. Therefore, to make a plan for training a Russian G2P model, we also need to discuss which phonetic representations are best represented of Russian graphemes, how to treat iotated vowels, and if we should make a contrast between palatalized consonants and non-palatalized consonants.

The modern Russian alphabet consists of 33 letters: 20 consonants (<б, в, г, д, ж, з, к, л, м, н, п, р, с, т, ф, х, ц, ч, ш, щ>), ten vowels (<а, е, ё, и, о, у, ы, э, ю, я), one semivowel (<й>), and two modifier letters or signs (<ь, ъ>) that alter pronunciation of preceding consonants or a following vowel. Table 9 shows the Standard Russian orthographic vowels and their respective pronunciations.

Table 9. Standard Russian orthographic vowels, semivowel, iotated vowels, and signs

Orthography	Phonetic symbols	
	IPA	**DC-Arpabet**
<а>	[a]	[aa]
<е>	[e], [ja], [je], [jo], [ju]	[ee], [yya], [yye], [yyo], [yyu]
<э>	[e]	[ee]
<и>	[i]	[ii]
<о>	[o]	[oo]
<у>	[u]	[uu]
<ы>	[ɨ]	[ie]
<й>	[j]	[yy]
<ё>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<ю>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<я>	[ja], [je], [jo], [ju]	[yya], [yye], [yyo], [yyu]
<ь>	no phonetic value	
<ъ>	no phonetic value	

In most analyses, the Russian vowel inventory contains five vowel phones: <i, e, a, o, u>. However, studies on Modern Russian (Timberlake, 2014; Yanushevskaya & Bunčić, 2015) claim a sixth vowel <ɨ>. Each one of these vowels is realized as a rich set of allophones ruled by stress and phonological environment. In most cases, these vowels merge into two or four vowels when stressed: /i, u, a/ after hard consonants and /i, u/ after soft consonants.

In orthography, each vowel is represented by two letters. This happens so we can distinguish the not-iotated vowels and the iotated vowels. Vowels in Russian do not have a phonemic distinction of quantity; there are no words distinguished by, for example, a long [aː] as opposed to a short [a]. Thus, Table 10 provides the 11 phonetic vowels in Standard Russian, together with a semivowel.

Table 10. Standard Russian phonetic vowels and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[a]	[aa]	[bb **aa** nn kk uu]
[e]	[ee]	[vv ss tt rr yye zz **ee**]
[i]	[ii]	[vv yye txj yye rr nn **ii** xx]
[o]	[oo]	[mm ii nn **oo** vv aa ttj]
[u]	[uu]	[ss oo oo bb cx tx **uu**]
[ɨ]	[ie]	[tt **ie** ss yya txj yye]
[j]	[yy]	[tt uu tt aa kk oo **vv** aa]
[ja]	[yya]	uu vv aa zj ee nn ii **yya**]
[je]	[yye]	[ii ss tt **yye** ts]
[jo]	[yyo]	[ll **yyo** xx kk aa vv aa]
[ju]	[yyu]	[ll uu txj su uu **yyu**]

As mentioned before, Standard Russian has 20 consonants (see Table 11). However, its phonetic inventory has 32 different consonants, as shown in Table 12. Most of these occur with both a palatalized and a non-palatalized version. The remaining consonants have a single version – the velars are palatalized before front vowels; palatals are either invariably palatalized (e.g., /j/) or invariably not.

Table 11. Standard Russian orthographic vowels

Orthography	Phonetic symbols	
	IPA	DC-Arpabet
<б>	[b], [bʲ]	[bb], [bbj]
<в>	[v], [vʲ]	[vv], [vvj]
<г>	[g]	[gg]
<д>	[d], [dʲ]	[dd], [ddj]
<ж>	[ʒ], [ʒʒ]	[zj], [zx zx]
<з>	[z], [zʲ]	[zz], [zzj]
<к>	[k]	[kk]
<л>	[l], [lʲ]	[ll], [llj]
<м>	[m], [mʲ]	[mm], [mmj]
<н>	[n], [nʲ]	[nn], [nnj]
<п>	[p], [pʲ]	[pp], [ppj]
<р>	[r], [rʲ]	[rr], [rrj]
<с>	[s], [sʲ]	[ss], [ssj]
<т>	[t], [tʲ]	[t], [ttj]
<ф>	[f]	[ff]
<х>	[x]	[xx]
<ц>	[t͡s]	[ts]
<щ>	[t͡ɕ]	[tx]
<ш>	[tʂ]	[txj]
<ч>	[ʂ]	[sj]

Thus, ш and щ share a very similar way of pronouncing their sounds. The letter ш has the phonetic value of [ʃtʃ] (a "palatalized /ʃ/" followed by a sound similar to [tʃ], in which the stop element, represented by t, is weak) or [ʃʃ] (a long "palatalized /ʃ/"). Either of these pronunciations of ш is regarded as correct, but it is common for any speaker to use only one of them. The letter щ has a phonetic value that can switch between [tʃ] (a weak t followed by a "palatalized /ʃ/") and [tɕ] (a weak t followed by a [ɕ]). Either of these is correct but it is pronounced differently depending on the region of Russia (Timberlake, 2014).

One of the most characteristic features of Russian consonantal phonology is that most sounds have both a palatalized and a non-palatalized phonological segment. In accordance with (Bondarko, 2005), palatalized consonants are referred to as soft and non-palatalized consonants as *hard*. Palatalization is an articulation of a consonant in which the blade of the tongue moves toward the hard palate. For example, when a non-palatalized consonant is pronounced, the tip of the tongue is touched near the teeth, while the middle of the tongue lies low in the mouth. In contrast, when the palatalized consonants are pronounced, the tip of the tongue touches behind the upper teeth, and the blade and the middle of the tongue are raised towards the hard palate. Most consonant articulations in Russian have two forms, with or without palatalization. Thus, palatalization in Russian is indicated by adding a diacritic to the phone. According to IPA, we use a palatalized diacritic (/ʲ/) when referring to palatalized consonants (e.g., /bʲ, dʲ, gʲ/). Palatalization is contrastive in word-final and in heterogenic medial coda positions.

Some examples illustrating the palatalization contrast extracted from (Padgett, 2001) are given in (1). Contrasts (1a–b) are prevalent in the language, while (1c) is more limited due to assimilations and neutralizations in that context.

(1a) before back vowels

Mat	foul language	mat'	crumpled
Rat	glad	r'at	row
Vol	ox	v'ol	he led
Nos	nose	n'os	he carried
Suda	court of law	s'uda	here, this way

(1b) word-finally

Mat	foul language	mat'	mother
Krof	shelter	krof'	blood
Ugol	corner	ugol'	(char)coal
v'es	weight	v'es'	entire

(1c) before another consonant

polka	shelf	pol'ka	polka
tanka	tank		
v'etka	branch		
Gorka	hill		

Russian has two modifier signs with no phonetic value. The soft sign *Ь* signals the presence of a soft consonant, and the hard sign *Ъ* signals the presence of a hard consonant. Therefore, they can be used between a consonant or a vowel (*Ь*) and between a consonant and a vowel, between two consonants, or at the end of a word after a consonant (*Ъ*).

In Russian *Ь* has much wider usage than *Ъ*; *Ь* can be used at the end of words or in between two consonants, and it indicates that the preceding consonants are soft. Neither *Ь,* or *Ъ* can be a stand-alone letter or the first letter in a word.

At the same time, there are some letters in Russian that are always hard or always soft and will sound the same way, whether there is a soft sign (Ь), or not. The consonants <ж, ш, ц> are always hard (if these consonants are followed by a soft sign, the sign cannot soften the consonant and serves a purely grammatical purpose. The consonants <ч, щ, й> are always soft. A soft sign following these consonants, once again, serves only a grammatical purpose.

Table 12. Standard Russian phonetic consonants and respective examples

Phonetic symbols		Example (DC-Arpabet)
IPA	**DC-Arpabet**	
[b]	[bb]	[gg **aa** zz bb aa nn kk aa]
[bʲ]	[bbj]	[tt uu rr **bbj** yye rr nn]
[v]	[vv]	[tt uu tt aa kk oo **vv** aa]
[vʲ]	[vvj]	[sj yye **vvj** yye vv]
[g]	[gg]	[sj pp ii **gg** yye llj]
[d]	[dd]	[ee **dd** gg aa rr]
[dʲ]	[ddj]	[bb uu **ddj** tt yye]
[ʑ]	[zj]	[vv rr aa **zj** nn aa yy]
[ʑʑ]	[zx zx]	[pp oo bb yye rr yye **zx zx** yyu]
[z]	[zz]	[bb yye **zz** aa llj]
[zʲ]	[zzj]	[vv oo **zzj** mm yyo ts yya]
[k]	[kk]	[gg rr uu zz oo vv ii **kk** oo vv]
[l]	[ll]	[gg rr yya nn uu **ll**]
[lʲ]	[llj]	[gg uu bb ii tt yye **llj** nn oo]
[m]	[mm]	[dd yye ll aa **mm** ii]
[mʲ]	[mmj]	[vv ii ts yye pp rr yye **mmj** yye rr]
[n]	[nn]	[ii mm yye **nn** nn oo]
[nʲ]	[nnj]	[dd yye **nnj** gg aa mm]
[p]	[pp]	[mm ii tt rr oo **pp** oo ll ii tt]
[pʲ]	[ppj]	[**ppj** rr yye mmj yye rr oo mm]
[r]	[rr]	**[rr** yye zz aa ll]
[rʲ]	[rrj]	[**rrj** aa dd ii tt yye ll yya mm]
[s]	[ss]	**[ss** vv yye tt yya tt]
[sʲ]	[ssj]	[ss vv yya zz aa ll oo **ssj]**
[t]	[t]	[ss mm yye **tt** uu]
[tʲ]	[ttj]	[ss mm oo tt rr yye **ttj]**
[f]	[ff]	[ss pp yye ts ii **ff** ii kk aa]
[x]	[xx]	[ss rr oo kk aa **xx**]
[t͡s]	[ts]	[dd yye mm oo pp pp oo zz ii **ts** ii ii]
[t͡ɕ]	[tx]	[tt yye kk uu **tx** ii xx]
[tʂ]	[txj]	[nn ii nn **txj** yye]
[ʂ]	[sj]	[tt ii **sj** ii nn aa]

5. **Methodology**

To improve version 2 of the European Portuguese normalizer, we first analyze how version 1 of the normalizer performed in comparison to how we want version 2 to work. Secondly, we outline how to build two new G2P models for different languages. Briefly, two transitional questions served as the basis for our research:

(1) How to upgrade a Normalizer into one that covers most of the normalizable tokens? And (2) How to create and validate new G2P models in two different languages?

The Normalizer Linguistic Expansion (NLE) project aimed to expand the rules of the Replacement Maps (RMs) in the Normalizer to cover Real numbers, Symbols, Abbreviations, Ordinals, Measurements, Currency, Dates, and Time, respectively. Table 13 shows an example of the input and its respective normalized output.

Table 13. Normalization process – input and output example

Input	Normalized Output
não mais **cm2** de território inacessível para a polícia eles somaram um buraco de **1577 cm2** na parede	não mais **centímetros quadrados** de território inacessível para a polícia eles somaram um buraco de **mil e quinhentos e quarenta e sete centímetros quadrados** na parede

The goal was to have a more consistent and simpler Normalizer with greater coverage of unambiguous inputs. We added new symbols and abbreviations to the RMs and created Unit Tests (UTs).[6] The Symbols rule was designed to convert non-alphanumeric symbols into their spoken forms, but only when they were not handled by other rules. The Abbreviations rule was created to convert miscellaneous alphabetic or mixed alphabetic-symbolic sequences to their spoken forms (see Table 14). In version 2 of the Normalizer, all abbreviated forms were added to allow for multiple possible expansions due to lexical or inflectional reasons.

Table 14. Normalization process – abbreviations input and output

Input	Normalized output
Wi-Fi	**Uaifai**
r/c	**rés do chão**
O **ap.** fica longe	o **apartamento** fica longe
O **D.r** Duarte trabaha no Hospital Santa Maria	o **doutor** Duarte trabaha no Hospital Santa Maria
Ela é vegetariana, **i. e.**, ela não come carne e peixe	ela é vegetariana, **isto é**, ela não come carne e peixe
O **núm.** de erros é alto	o **número** de erros é alto

The Real Numbers rule converted numeric values of integers and/or decimal values into their spoken forms. The Ordinals rule converted a sequence of a real number and an ordinal marker into their expanded spoken forms, while the Measurement rule converted a sequence of a real number and a measurement unit into their expanded spoken forms. The Currency rule covered currency expressions in their expanded forms. In version 2 of the Normalizer, the Measurement and Currency rules introduced several new abbreviations and symbols that were expanded to their spoken form only when they were unambiguous.

Throughout the NLE project, we expanded several normalizable tokens (symbols, abbreviations, etc.), implemented rules on the various rules, and fixed issues related to varieties confusion between pt-PT and pt-BR. We ensured that pt-PT rules and pt-BR rules were always separate in version 2. The main problem was that several PN rules and assets were shared between different regions of the same language. This had been done in pt-PT and pt-BR for several rules, but mainly due to some orthographic differences between pt-PT and pt-BR.

[6] A Unit Test (UT) is a set of one or more examples to ensure that a unit of code behaves as expected. More specifically, Unit Tests are sets of examples of inputs and expected outputs that test whether the normalizer produces the correct - or expected - output given the input.

To prepare a G2P model for two different languages - Swedish and Russian – an overview of the language had to be performed. Once this task was completed, the next step involved developing a phone set for each language. This implicated identifying the distinct sounds present in the language and creating a set of symbols (phones) to represent them. This phone set serves as the basis for the phonetic transcriptions used in the G2P model. Following this, the phonetic lexicon for each language was mapped and automatically converted to the DC-Arpabet format. DC-Arpabet is a standardized phonetic transcription system used for English, but it can also be applied to other languages. This step involved mapping the phones from the initial phone set to the corresponding DC-Arpabet symbols. The phonetic lexicon was then revised and corrected by native speakers. This step is crucial to ensure the accuracy and quality of the G2P model. During this stage, errors in the phonetic transcriptions were identified and corrected, and missing words or pronunciations were added. Finally, the G2P model was evaluated to assess its performance and accuracy. This involved testing the model on a set of known words and comparing the model's predicted phonetic transcriptions to the correct pronunciations. If necessary, adjustments were made to the model to improve its accuracy.

6. Results and Discussion

6.1. Evaluation metrics

This study discussed the metrics used to evaluate the Normalizer and G2P tools. For the Normalizer, we used accuracy, WER, precision, recall, and F1 score. However, WER may not be an adequate metric since most tokens in a sentence are likely to be irrelevant to the normalizer. We introduced the WERnorm metric to address this issue, which calculates the edit distance over the number of normalized reference tokens. Concerning the G2P, precision, recall, and F1 score were used to evaluate the proportion of correctly predicted grapheme to phone correspondences. It is important selecting appropriate evaluation metrics to accurately assess the performance of these models.

In accordance with Jurafsky and Martin (2022), the word error rate is based on how much the word string returned by the recognizer (the hypothesized word string) differs from a reference transcription. Thus, WER is the proportion of transcription errors that the ASR system makes relative to the number of words that were said. The lower the WER, the more accurate the system.

Equation 1. Word Error Rate formula

$$Word\ Error\ Rate\ = 100\ \times \frac{Insertions\ +\ Substitutions + Deletions}{Total\ Words\ in\ the\ Correct\ Transcript}$$

When evaluating the Normalizer tool, a problem related to using WER as a metric for assessing the normalizer is that in any given sentence – even if we filter only for sentences containing some normalizable expression – most of the tokens are likely to be irrelevant to the normalizer. Thus, WER divides edit distance over all tokens in the reference; however, most tokens we would not expect to be normalized anyway, so applying WER to the Normalizer distorts the results. Instead, using the WERnorm[7] (Word Error Rate over Normalizable tokens) metric, we divide by the number of normalized reference tokens.

Equation 2. Word Error Rate over Normalizable Tokens metric formula

$$WER_{norm}\ = \frac{\sum_{x,y\ \epsilon\ X,Y}\ edit\ distance\ _{x,y}}{\sum_{y\ \epsilon\ Y}\ \max\ (1, normalizable\ tokens_{x,y}\ )}$$

[7] The WERnorm was an evaluation metric proposed and elaborated by the ML team.

Accuracy is also metric for evaluating classification models. It answers the question: "Overall, how often is our model correct?". Thus, Accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally, where 1 represents total accuracy.

Equation 3. Accuracy metric formula

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

Precision is defined as the number of true positives divided by the number of true positives plus false positives. This formula is used to understand the model's accuracy.

Equation 4. Precision metric formula

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall is described as the number of true positives divided by the number of true positives plus false negatives. That is, it calculates the true positives by anything that should have been predicted as positive.

Equation 5. Recall metric formula

$$Recall\ = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 score, also known as F1, is a measure of a model's accuracy on a dataset. F1 score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. A perfect model has an F1 score of 1.

Equation 6. F1 metric formula

$$F1 = \frac{2(Precision * Recall)}{Precision + Recall}$$

6.2. **Normalizer and G2P evaluation**

In the evaluation of version 2 of the normalizer, a set of around 1000 manually normalized and tagged reference sentences were used. The sentences were processed by the normalizer and compared with the reference output. The overview of the evaluation set included the number of sentences, normalizable reference tokens, and total reference tokens used. It can be observed that out of the 23428 reference tokens, only 31.2% were normalizable. All the 1000 sentences had various normalizable tokens, and a corresponding rule was applied to each token. However, the sentences lacked ordinals, resulting in a lower percentage of normalizable ordinals (1.4%) than real numbers (49.9%). This could potentially impact the results.

The normalizer results show that there are statistically significant differences among the rules in both versions of the normalizer. For example, the ordinals rule exhibits the highest percentage in both versions, while abbreviations have the lowest percentage. Real numbers, measurements, currency, time, symbols, and dates show considerable variations in their performance. The analysis of the F1-score results shows a significant performance increase in some of the rules, such as abbreviations and currency. The improvement is higher for abbreviations, currency symbols, and symbols than new real numbers, data formats, or ordinals. Concerning the WER and WERnorm metrics, WERnorm shows that version 2 of the normalizer has, on average, a 4 pp. lower error rate over normalizable tokens compared to version 1. The use of the ordinals and real numbers rule is the largest source of improvement on normalizer WER. Regarding accuracy, version 2 shows an improvement of, on average, 28 pp compared to version 1. The results shown in Table 15 indicate a significant improvement in the normalizer's performance in all the rules.

Table 15. Normalizer's version 1 and version 2 performance regarding WER, WERnorm, and Accuracy metric

Normalizer pt-PT	**WER**	**WERnorm**	**Accuracy**
Normalizer version 1	40,13%	12,47%	46,96%
Normalizer version 2	35,29%	10,58%	74,09%

Regarding the G2P model for Swedish, after doing an analysis of the phonetic lexicon (see Table 16) and the final G2P model, the revision of the lexicon shows that 99% of the orthographic words are correct, with only 1% being incorrect.

Table 16. Swedish phonetic lexicon overview

Swedish Phonetic Lexicon Overview		
	#	**%**
Lexicon entries	25248	100
Corrected words	25067	99
Incorrect words	181	1
Correct transcriptions	22339	88
Incorrect transcriptions	2909	11
WER		11

The most common errors regarding orthographic words are in (1) Double letters (e.g., *u**tt**lardet - u**t**lardet*); (2) Diacritics (e.g., *genéve - genève*). Thus, due to errors in the transcribed audio used to create the initial lexicon, we can observe 1. Segment position alteration - Metathesis (e.g., *vädning - vending*), and (3) Segment suppression in the middle of the word - Syncope (e.g., *lövs**t**edt - löv_tedt*). Consequently, errors in the orthographic words will be followed in the phonetic transcriptions as well.

Regarding phonetic transcriptions, we find 88% of transcriptions are correct and 11% incorrect. The most common errors of incorrect transcription are in the following vowels:

1. Mapping of graphemes <ä, o> - these graphemes have different equivalent phones; however, their occurrence does not always depend on phonological rules (e.g., [**eh** gg oo] – [**ae** gg ug]). /o/ was mostly transcribed as [oo], and /ä/ as [eh] instead of [ae].
2. Vowel reduction in /i/ - weakening of a vowel in an unstressed position (e.g., [hh **ih** ll **ih** ng shx eu] – [hh **ii** ll **ii** ng ss eu]). The generated transcription reduced the vowel /i/ whereas the correct transcription should be [ih] instead of [ii].

Regarding the phonetic lexicon performance, we achieved a WER of 11%. We consider this lexicon to be of good quality and ready to use for the G2P model.

Regarding per phone evaluation, we tested each phone in terms of Precision, recall, and F1-score. We conclude that the best performed phones are consonants [bb, dd, ff, hh, kk, ll, mm, nn, rr, ss, tt, vv]. Phones with a perfect F1 score are the following two: [ff, rr]. Therefore, consonants are performing better than vowels. The phone [ug] has the worst performance with an F1-score of 81%, following [ii] (83%), [uu] (85%), [ah] (86%), and [oo] (87%). The average Precision, Recall, and F1-score present a good result of 96% per phone. Thus, in total, all phones present 97% of accuracy.

Considering the G2P model for Russian, the analysis of the lexicon revision (see Table 17) showed that most of the Russian orthographic words were correct (96%), with only 4% being incorrect.

Table 17. Russian phonetic lexicon overview

Russian Phonetic Lexicon Overview		
	#	**%**
Lexicon entries	27340	100
Corrected words	26286	96
Incorrect words	1054	4
Correct transcriptions	23989	88
Incorrect transcriptions	3351	12
WER		12

We observed the following two errors in orthographic Russian words: (1) Segment suppression in the middle of the word - Syncope (e. g., а_нализа – а**з**нализа), and (2) Diacritics suppression (e. g., взл**е**ты - взл**ё**ты). Consequently, errors in the orthographic words will be followed in the phonetic transcriptions as well.

Regarding phonetic transcriptions, we find 88% of transcriptions are correct and 12% incorrect. The most common errors of incorrect transcriptions are due to:

i. Sonorization - sound change where a voiceless consonant becomes voiced (*e. g.,* [**ss** oo zz dd aa nn ii yye] - [**zz** oo zz dd aa nn ii yye]).

ii. Distinction between vowels and iotated vowels – vowels before <*ж*, *ш,* н> are not iotated (*e. g.,* [aa vv aa nn ss ts **yye** nn uu] - [aa vv aa nn ss ts **ee** nn uu])

iii. Mapping of vowel <и> - its regular equivalent phone is [ii], however it can also be [ie] (*e. g.,* [mm oo ss kk vv **ii**] - [mm oo ss kk vv **ie**]).

We got a reasonable WER of 11% for the phonetic lexicon performance. This lexicon is of good quality and ideal for the G2P model. Regarding per phone evaluation, we tested each phone regarding Precision, Recall, and F1-score. We conclude that the best performed phones are mainly consonants [gg, zx zx, ll, mm, nn, rrj, ts, txj]. Phones with a perfect F1 score are predominantly palatalized consonants [zj, llj, mmj, nnj]. Indeed, consonants are performing better than vowels. The phone [ii] has the worst performance with an F1-score of 83%, following [ie] (85%), iotated vowels [yya, yye, yyo, yyu] (86%), and consonant [zz] (87%). The average Precision, Recall, and F1-score present a good result of 95% per phone. Thus, in total, all phones present 96% of accuracy.

In conclusion, the analysis of the lexicon revision showed that 96% of the words were correct, with 4% of incorrect words mostly due to errors in orthographic Russian words and phonetic transcriptions. The phonetic lexicon performance had a reasonable WER of 12%, and the best-performing phones were predominantly consonants, especially palatalized consonants. The Swedish G2P model showed better performance with a 98% F1-score compared to the Russian model's 96% F1-score, indicating that Swedish orthography is more phonetically based and has a simpler phone set.

7. Conclusion and future work

The present paper describes the work done on the Normalizer and Grapheme-to-Phone models in Speech Technologies, focusing on the importance of linguistic knowledge in preprocessing models. The Normalizer was evaluated, and version 2 showed better performance than version 1. The phonetic lexica for Swedish and Russian were obtained and evaluated, with the Swedish G2P model showing better accuracy than the Russian model. The results obtained led to the expansion and improvement of the Normalizer and G2P models, now

supporting 14 languages. The rule-based approach used in the Normalizer and G2P models increased their accuracy and performance, demonstrating the importance of linguistic knowledge in preprocessing models. The work done on the Normalizer also contributed to discussions on specific vs. language-generic rules. In this sense, we see future possibilities of expanding the coverage of the current normalizer by implementing the real numbers and ordinals rule (best-performed rules) to a module that can be used for various languages. Languages in which real numbers and ordinals are represented similarly (e.g., magnitude/decimal separators, number base). The same could be done with other rules, although this still needs thorough research.

Overall, the study shows the relevance and meaningfulness of having linguistic knowledge in preprocessing models for Speech Technologies.

References

Alasadi, Abdumalik & Ratnadeep Deshmukh (2018) Automatic speech recognition techniques: A review. In *Signal Processing and Computer Vision*, pp. 464–470. Available at https://www.researchgate.net/publication/325296232

Armstrong, Eric & Paul Meier (2005) IPA Chart. Available at https://www.ipachart.com/

Bondarko, Liya (2005) Phonetic and phonological aspects of the opposition of "soft" and "hard" consonants in the modern Russian language. *Speech Communication* 47 (1–2), pp. 7–14. https://doi.org/10.1016/j.specom.2005.03.012

Brasoveanu, Adrian & Dotlacil Jakub (2020) Production-based cognitive models as a test suite for reinforcement learning algorithms. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, pp. 28–37. Available at https://aclanthology.org/2021.cmcl-1.pdf

Errattahi, Rahhal & Asmaa El Hannani (2017) Recent advances in LVCSR: A benchmark comparison of performances. *International Journal of Electrical and Computer Engineering* 7 (6), pp. 3358–3368. http://doi.org/10.11591/ijece.v7i6.pp3358-3368

García, Salvador, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez & Francisco Herrera (2016) Big data preprocessing: Methods and prospects. *Big Data Analytics* 1 (1), pp. 1–22. https://doi.org/10.1186/s41044-016-0014-0

Hennebert, Jean, Martin Hasler & Hervé Dedieu (1994) Neural networks in speech recognition. In *Proceedings of the 6th Microcomputer School of Neural Networks, Theory and Applications*. Swiss Federal Institute of Technology, pp. 23–40. Available at https://www.researchgate.net/publication/2249623

Jurafsky, Daniel & James Martin (2022) *Speech and language processing* (3rd ed.) [Draft]. Stanford University. Available at https://web.stanford.edu/~jurafsky/slp3/

Kurata, Gakuto, Kartik Audhkhasi & Benedict Kingsbury (2019) IBM Research advances in end-to-end speech recognition at INTERSPEECH 2019. *IBM Research Blog*. Available at https://www.ibm.com/blogs/research/2019/10/end-to-end-speech-recognition/ [accessed on 23/12/2022].

Padgett, Jaye (2001) Contrast dispersion and Russian palatalization. In Elizabeth Hume & Keith Johnson (eds.), *The role of speech perception in phonology*. Academic Press, pp. 187–218.

Rajadnya, Kirti (2020) Speech recognition using Deep Neural Network Neural (DNN) and Deep Belief Network (DBN). *International Journal for Research in Applied Science and Engineering Technology*, 8 (5), pp. 1543–1548. https://doi.org/10.22214/ijraset.2020.5359

Riad, Tomas (2014) *The phonology of Swedish*. Oxford University Press.

Shoup, John (1980) Phonological aspects of speech recognition. In Wayne A. Lea (ed.), *Trends in speech recognition*. Prentice-Hall, pp. 125–138.

Timberlake, Alan (2014) *A reference grammar of Russian*. University of California at Berkeley.

Vielzeuf, Valentin & Grigory Antipov (2019) *Are E2E ASR models ready for an industrial usage?*. Cornell University.

Wassink Sophie Groot, Jessica Wingerden van & Poell Rob (2022) Correction: Meaningful work and resilience among teachers: The mediating role of work engagement and job crafting. *PLoS ONE* 17 (5). https://doi.org/10.1371/journal.pone.0269347