Universität Zürich UZH

Master thesis

zur Erlangung des akademischen Grades

Master of Arts

der Philosophischen Fakultät der Universität Zürich

(Titel)

Author: Deborah N. Jakobi

Matriculation number: 16-054-165

Supervisor:

Institut für Computerlinguistik

Abgabedatum: (xx.xx.xxxx)

Abstract

This is the place to put the English version of the abstract.

Zusammenfassung

Und hier sollte die Zusammenfassung auf Deutsch erscheinen.

Acknowledgement

I want to thank X, Y and Z for their precious help. And many thanks to whoever for proofreading the present text.

Phillip Ströbel from the CL institute at the UZH for his help with the OCR technologies. Lysander Jakobi for writing the Hebrew transcription. Florina Vogel for helping with the Farsi transcription. Si-En? Tanzil?

Contents

List of Figures

List of Tables

List of Acronyms

IPA International Phonetic Association

IPA International Phonetic Alphabet

WER word error rate

CER character error rate

G2P grapheme-to-phoneme

seq2seq sequence-to-sequence

SIGMORPHON Special Interest Group on Computational Morphology and Phonology

NLP natural language processing

PoS Part-Of-Speech

WALS World Atlas of Language Structures

ASR automatic speech recognition

TTS text-to-speech

HTR Handwritten Text Recognition

DISC distinct single characters

1 Introduction

1.1 Motivation

With the advent of technologies that can process huge amounts of data, many linguistic tasks that were originally very tiresome and expensive to do, can now be accomplished much faster. Well known examples for this branch called natural language processing (NLP) are machine translation or search engines. While much of this research is done with written language, technologies like automatic speech recognition (ASR) or text-to-speech (TTS) require the processing of large amounts of spoken language. In many such technologies it is necessary that phonetic transcriptions of written or spoken text are available. They are needed such that a TTS model can learn how to produce them as an intermediate step and eventually produce speech based on phonetic transcriptions. While this can be done and was typically done for one language only, there are efforts made to create multilingual models. In order to do that, there needs to be a lot of knowledge about how language works. Comparing languages and studying their similarities and differences is part of a well-established branch of traditional linguistics called comparative linguistics. The analysis of large amounts of text in any language is commonly referred to as corpus linguistics. Corpus linguistics allows for both qualitative and quantitative analysis of text. Although text can refer to written or spoken language, most corpora contain written text [McEnery and Hardie, 2011]. Multilingual corpora can be used to compare languages. If all of these different approaches are combined, we end up by what we could call comparative corpus phonetics.

The text group of the Language and Space lab at the University of Zurich maintains a project that provides such a multilingual corpus consisting of 100 language text samples [SPUR project]. Those 100 languages are meant to be representative for all the world's languages which is explained in more detail in section 2.1. It is therefore meant to give insight on relations, similarities, differences or properties of individual languages or language families. Specifically, their goal is to use quantitative methods like statistical modelling, machine learning and information theory to study language variation and compare languages. While there are many different types of analyses that can be performed on those text samples, the question keeps coming up if analyses of speech versions of those languages might not be more accurate or give better insight. Although this is not sure and simply a suggestions, it needs to be proven that working on text versions only represents languages well enough to present generalizable results. This present thesis ties in with this open research question. The goal is to collect phonetic transcriptions of the corpus. The same analyses can be performed on the phonetic texts which can be compared to the text analysis. In order to add a phonetic corpus to the already existing one, various steps need to be performed which are outlined in section 1.3.

Add quick intro into corpus linguistics, quantitative analysis, this is essentially what is done with the corpus. [McEnery and Hardie, 2011]

Introduction to comparative and historical linguistics at some place. [Hock and Joseph, 2019]

1.2 Goals & methods

The primary goal of this thesis is to create phonetic transcriptions of as many languages as possible which are in the already existing corpus. Given the explanations above, the steps I aim to conclude to reach this goal are the following:

1. Data collection: The given dataset contains no phonetic transcriptions of those 100 languages. The first step is to find already existing data.

2. Phonetic transcriptions: As existing data will not be available in sufficient amounts to perform meaningful analysis, the next step is to actually create phonetic transcriptions of as many languages as possible of the corpus.

3. Calculations and Analysis: Once the transcriptions have been obtained, the newly created phonetic corpus can be analysed and calculations can be performed.

4. Based on the steps before, I will answer the following final question: Is it essential for the study of multilingual corpora to perform analyses on phonetic text (i.e. speech representations) rather than only written text? depends on what can be done before...

1.3 Research questions

1.4 Thesis structure

The thesis is subdivided into six chapters including a final conclusion. Chapter 2 sets the boundaries of the theoretical background. It presents the linguistic foundation of phonetics and phonology, an introduction to corpus linguistics or rather corpus phonetics and finally an overview of the possibilities for automated creation of phonetic transcriptions. Chapter 3 introduces to the struggle of data collection. It explains the various data types and how those can be used. Chapter 4 dives deeper into the possibilities for creating phonetic transcriptions and what models can be used to create those. Chapter 5 presents my own experiments to create phonetic transcriptions of the corpus.

2 Research Background

2.1 The corpus

The corpus contains 100 languages which are proposed by Comrie et al. [2013]. This online book contains different chapters each of which shows a different linguistic feature including a map which shows the distribution of that feature over the world's languages. While the number of languages presented on the individual maps depends on the amount of research done in a specific area, the sum of all maps gives quite an impressive overview on the structure of nearly half of the world's languages. Out of the 2676 languages a sample of 100 languages was chosen. This sample does not contain too many languages from one area, neither does it contain too many languages from one family. Not considering the aforementioned criteria of maximizing genealogical and areal diversity can lead to misleading results. Figure 1 shows the distribution of the corpus on a world map. The different icons show the genus of the languages which is a classification of languages defined by the World Atlas of Language Structures (WALS) team that maintains the language collection. The interactive map can be viewed online [100-language-sample]. Table 0 in the appendix A shows all languages that are in the 100 language corpus.

2.2 Corpus linguistics and quantitative analysis

The relation between spoken and written language. Remember that writing systems came only much later compared to language in general. Can they capture language as such well enough? Computational linguistics deals mostly with written languages, what does linguistics say and do?

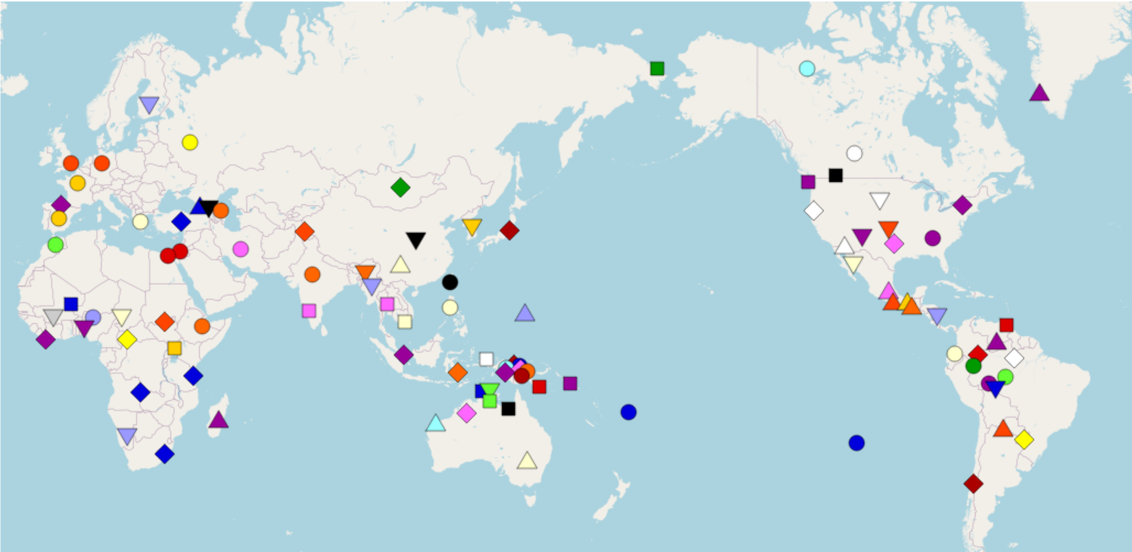

Figure 1: WALS - 100 Language Sample

2.3 Introduction to phonetics and phonology

Given that phonetics and phonology is a sub-area of traditional linguistics and often only touched on superficially in computational linguistics, I will summarise the most important assumptions and terms concerning said field. A very important terminological distinction is between phonetics and phonology. While phonetics refers to the study of actual sounds, phonology refers to the study of sound *systems*. In phonetics, it is not so much important what the different sounds mean, but how they are produced and perceived and what different sounds a human being can produce and perceive at all. When it comes to human communication using spoken language, many of these sounds are not actually used to produce distinguishable meaning. This is why on the other hand phonology is important to describe the set of distinguishable sounds that make up a language. For example: the letter /r/ in English can be pronounced in many different ways. None of those pronunciations produces a change in meaning. This means that there exist many different *phonetic* sounds but only one *phonological* or *phonemic*. Those sounds are referred to as phone and phoneme respectively. While there are infinitely many phones there are only finitely many phonemes in a language. Not all different possible sounds are actually considered qualitatively 'good' sounds of a language. Usually there is a subset of all possible phones that is accepted as 'good quality sounds' within all different dialects of a language [Kracht, 2007]. An obvious example being loudness: Although very silent speech produces correct phones, these are not 'good quality' as they simply cannot be understood. Or speaking in English with hardly any mouth and tongue movement. Although this produces understandable sound, it is not generally considered good speech.

The alphabets used to represent sounds in different languages do not uniquely map a letter to one specific phoneme. Most of the time, there is a standard pronunciation of each letter that is trained by reciting the alphabet. However, in reciting the alphabet there is a vowel added to the consonants in order to pronounce them more easily. These explanations make clear that the mapping of written text to spoken text in various languages is complex. In order to make things easier, there is the International Phonetic Alphabet (IPA) that can be used to transcribe any text in any language to a phonetic text [Kracht, 2007].

It is important to note at this point that the terms phonetic and phonemic respectively phone and phoneme are sometimes used interchangeably. Their linguistic definition as given above is clear while the definition on the computational side is often less strict. definition of phoneme / phone, the one that is used e.g. in Lee et al. [2020], foot note 4

add explanation of allophones, monophthongs, diphthong, suprasegmental they appear quite often in the lit (maybe make a glossary

Writing systems

Unlike spoken language that was a part of human interaction all the time, writing systems only developed over time. There are different writing systems that developed in different places at different times. The structure of the spoken language, the cultural context or the tools that were at hand to write are a few of many factors that influenced the emergence of a specific writing system. A single grapheme can represent either a phoneme, a syllables or words. In German, for example, the writing system consists of an alphabet, the Latin alphabet. The Latin alphabet is used for many different languages in western Europe and those languages that were influence by colonisation. There are other alphabets like the Cyrillic or the Greek alphabet. Having an alphabet means that each grapheme of the alphabet represents a phoneme. The exact phonemic realisation of the grapheme depends on the context, so there is not necessarily a one-to-one mapping. Many language use accents to slightly change the sound of a grapheme or they use more than one grapheme to represent one sound. Apart from these conventions spoken and written languages change differently over time. Spoken languages are typically more flexible and ready to change while their written representation often stays the same. This can lead to official governmental interventions like the German orthography reform of 1996 that intended to adapt the German spelling to represent the German pronunciation more adequately. Also, major inventions like printing machines gave rise to standardization of writing systems as reading and writing became more common.

A special variant of an alphabet-language is abjad. Abjad represents only consonants and no vocals. Semitic languages make use of abjad. A part from alphabets, there are also syllabic and logographic writing systems. In syllabaries, a grapheme represents a syllable instead of a single sound. Examples are the Japanese Hiragana and Katakana. Logographic systems represent entire words or morphemes as graphemes. Chinese is an example for a logographic system. We cannot break down Chinese signs into single morphemes or letters. The history and development of writing systems is an entire independent study area. For this thesis it is mostly important to be aware of the independently developing systems. Not all scripts can be treated the same and this most certainly has implications on models to create phonetic transcription.

add example

An exception to the above explained characteristics of an alphabet are phonetic alphabets like the International Phonetic Alphabet (IPA) where each grapheme represents exactly one phone [CrashCourse, 2021]. More on this special alphabet will be explained in section 3.1.

2.4 Corpus phonetics

Due to recent technological advancement it has become possible to store large digital collections of speech recordings and their aligned transcriptions. These possibilities gave rise to a wider acknowledgement of corpus phonetics. Corpus phonetics deals with an abundance of linguistic variation. In addition to language, style or vocabulary variation, there are differences in dialect and idiolect, physiological state of the speakers and their attitude [Liberman, 2019; Chodroff, 2019]. Many methods and tools used in corpus phonetics are based on ASR algorithms or simple programming [Chodroff, 2019].

2.5 Automated phonetic transcription

Today's technologies allow to build models that create phonetic transcriptions automatically given an original text. There are several approaches which will be discussed below. Creating phonetic transcriptions is essentially a sequence-to-sequence (seq2seq) task. Like other NLP tasks its goal is to transform a sequence of characters into another sequence of characters. In the present case, the input sequence is a sequence of graphemes. These can look very differently depending on the script (see section 2.3). The output sequence is a sequence of phonemes[0]. A common way to transform written text into its phonetic version is referred to as G2P. The idea behind this approach is that individual letters (graphemes) are converted into sounds represented as phonemes.

Most of the research done in this area is limited to the English language. This is not uncommon in NLP research. The overwhelming availability of English data resources and the unavailability and serious struggles to find data in other languages heavily influences this research. Ashby et al. [2021] report that the SIGMORPHON (Special Interest Group on Computational Morphology and Phonology) G2P tasks in 2020 and 2021 is the first attempt to tackle multilingual G2P. While these are very recent tasks, there are earlier models and methods that contributed to the evolving of nowadays G2P methods.

2.5.1 Rule-based models

The first systems to create phonetic transcriptions of text were rule-based systems. Rule-based transcriptions models are built using linguistic pronunciation rules. In order to be able to create such a system, one needs to collect pronunciation rules first. While there are only few languages where such rules are ready and available for the general public there are many languages where those rules need to be created first. In order to create the rules in the first place, a lot of linguistic expertise is needed. Apart from this initial effort to create the rules, a problem with rule-based approaches is the maintenance of the systems. To maintain the system, experts need to keep track of language change which is time consuming and expensive. In addition, most languages are irregular in their pronunciation and those irregularities need to be tracked. Due to the open-vocabulary situation and the impossibility to

[0]Please refer to section 2.3 in order to understand the terminological implications phoneme. As it is common in research, I will stick to the term *phoneme* although strictly speaking it is not always correct. Phoneme in this case just refers to any symbol that is used to represent that sound.

cover all possible words, all systems must be able to deal with rare and unseen words [Rao et al., 2015; Bisani and Ney, 2008]. Rule-based systems are outperformed by more recent neural systems [Gorman et al., 2020; Ashby et al., 2021]. To the best of my knowledge, there have not been published any more new rule-based systems in the last few years. Many systems published considered only one language and were not multilingual (see e.g. Toma and Munteanu [2009]).

[Rao et al., 2015]

[Bisani and Ney, 2008]

add a few examples of rule-based systems and why and by whom they where outperformed (see Ashby et al. [2021]; Gorman et al. [2020] for this purpose)

2.5.2 N-gram Models / Statistical models

N-gram models, statistical models or joint-sequence models were used before neural models took over the field. These are sometimes referred to as traditional models. One reason why they were outperformed by neural models is that it is necessary to construct alignments between grapheme and phonemes. This is necessary because one grapheme can be realized as multiple phonemes or vice versa. It is not possible to simply have a one-to-one alignment. check Lo and Nicolai [2021] they include a lot of references about this topic

2.5.3 Neural models

Neural G2P models have been reported to outperform most other models [Lee et al., 2020]. Many researchers experiment with different variants of LSTM models [Lee et al., 2020; Hammond, 2021; Gautam et al., 2021; Rao et al., 2015]. Many earlier neural LSTM models use a connectionist temporal classification layer to include alignment information [Lo and Nicolai, 2021].

transducers: those are like automaton. Unlike automaton that only tell you if a certain sequence is in a particular language, transducers output something at every state.

seq2seq: condition output sequence on entire input sequence. This does not work well for input that gets continuously longer or very long input sequences.

Neural transducers, as presented by Jaitly et al. [2016], extend previously used seq2seq models. They can treat more arriving input without having to redo the entire calculation for the entire updated sequence. At each time step, the neural transducer can output zero to many output symbols.

A problem with creating phonetic transcriptions is that the input and output segments are not always of the same length. It is difficult to align input and output.

Generally, there is a difference between models that assume conditional independence between the each output step (e.g. Hidden Markov Models) and there are models that do not make this assumption but condition the current output on the entire sequence before (seq2seq). Seq2seq models, however, have to wait until the full input sequence is processed before they can start decoding.

The Special Interest Group on Computational Morphology and Phonology (SIGMORPHON) [Sigmorphon, 2021] regularly organizes shared tasks concerned with morphology and phonology. For the years 2020 and 2021 they organized a G2P conversion task [Ashby et al., 2021; Gorman et al., 2020]. The tasks represent a first attempt at creating benchmarks for multilingual G2P conversion. Both tasks and their results will be discussed in sections 2.5.4 and 2.5.5. Although there is other research on G2P, many recent publications have been made within the SIGMORPHON shared tasks which is why there are two separate sections on those tasks. As the SIGMORPHON tasks are the most recent and probably most influential contributions to G2P research, both tasks will be discussed separately below.

2.5.4 SIGMORPHON task 2020

Yu et al. [2020] contributed to the 2020 SIGMORPHON G2P task. Their contribution is of particular interest for this thesis as it proposes a data augmentation model for low-resource settings. As there are many languages in the corpus that have only very little available data, such a model could be of great use. The methodology applied in their approach is ensemble learning combined with a self-learning strategy.

Results

Error analysis

2.5.5 SIGMORPHON task 2021

The second iteration of this G2P task attempts at outperforming the models of the previous task. An additional challenge is its separation into high-, medium- and low-resource languages. This reflects the needs of this present research well, as many languages in the corpus are low resource languages. In preparation for the task, the WikiPron data (see chapter 3) was cleaned to exclude foreign words that include phones that are not in the actual language's native phone inventory. If a word contained foreign phones, it was excluded. This was the case for words whose pronunciation was not adapted to the language at hand but the transcription of the foreign language was used. This cleaning was only applied to medium- and low-resource languages. Additionally, the lists were sorted according to scripts. There are many languages that use multiple scripts and using them in the same dataset does not produce good results. There were other steps done to ensure good quality of the datasets. I collected more on that subject in section 2.6. The final datasets have to following sizes: The high-resource subtask consisted of about 41,000 word-transcription pairs of American English only. The medium-resource task provided 10,000 word-transcription pairs for ten languages and the low-resource task another 1,000 for ten different languages [Ashby et al., 2021]. The datasets were split into 80% training data, 10% development data and another 10% test data.

The baseline for this year's G2P task is an adapted version of last year's submission by Makarov and Clematide [2020]. The baseline model has been made available for this year's task. The model they use is a neural transducer that is trained with imitation learning. The basis of the neural transducer was originally designed for morphological inflection [Aharoni and Goldberg, 2016]. Instead of just learning to output the correct string, the model learns to produce an optimal sequence of edit actions needed to transform the input string into the output string. Due to the nature of inflection (overlapping alphabets of input and output sequences), the original model was encouraged to copy the input. This does not work well for G2P tasks as the input and the output alphabet are not always the same (especially for non-Latin scripts like Korean). explain neural transducer, the model more in depth.

As explained above, the model learns to create sequences of edit actions. The problem with this approach is that there are many possible sequences of edit actions that produce the same result. Imitation learning is proposed as a solution for this problem. explain imitation learning better and more precise.

Results

The results show that there are great differences in languages. One possible explanation is that the datasets were a mix between broad and narrow transcriptions. As narrow transcriptions contain much more detail, it can be argued that this is more difficult for any system. The authors doubt the influence of this but they could not (yet) quantify this impression.

Results in the low-resource setting are still worse compared to the medium-resource setting. This means that current systems seem to be unable to achieve a good performance when only using 800 samples for training. More research needs to be done in data augmentation techniques and improving the systems to cope with only little available data.

The differing performance for various languages calls for the questions what makes a language hard to pronounce. Especially as for Georgian, all submissions and the basline reached a word error rate (WER) of 0.0.

Error analysis

In order to find the most common errors of the systems, there were two types of analysis conducted. The first one is to simply find the most common wrongly transcribed grapheme-phoneme pairs. This analysis showed that many errors are due to language internal ambiguities. Some errors go back to errors or inconsistency in the data. The other type of analysis is to create a covering grammar. This means that a grammar is created that includes all possible combinations of grapheme-to-phoneme mappings that are allowed in this language. This set is constructed manually. This error analysis was only conducted for three languages in the medium setting and four of the low-resource languages. Then, words were considered that were predicted wrongly by the system. For those words it was checked whether the prediction by the system was completely wrong or if it was one of many possible transcriptions of that word. If so, the error was that system did not guess the correct transcription for this word. These errors could be considered ambiguities in the language. Another error type that can be identified by this analysis is when the reference transcription cannot be derived from the covering grammar. This can either mean that the covering grammar is incomplete or that the reference is a word that is very atypical for the language (e.g. borrowed word but not yet adapted to pronunciation rules) or simply wrong.

2.6 Data quality considerations

not sure where to put this, but I think it makes sense to have a separate section on data quality. Most papers include some things Data quality is crucial in any machine learning application. Papers mostly include a section about their preprocessing and what should be done to ensure high quality datasets. The list given below is an incomplete list of potential problems and measures taken in different settings:

- **Exclusion of words with less than two Unicode characters or less than two phone segments** [Ashby et al., 2021] add an explanation
- **Separation by script** [Ashby et al., 2021]: It is very straightforward why this is done. There is no obvious connection between the different scripts of a language and its pronunciation. It makes sense to treat different scripts as different languages.
- **Exclude foreign words with foreign pronunciations** [Ashby et al., 2021]: Foreign words in a language with their original pronunciation can add phonemes that are not in that language's phoneme inventory. If they were to be included it would make sense to include a pronunciation adapted to the actual language.
- **Words with multiple pronunciations in word lists**: Ashby et al. [2021] excluded those words, however, it might also be possible to add Part-Of-Speech (PoS) tags or other linguistic information to distinguish these words.
- **Consistent broad transcriptions** [Ashby et al., 2021]: With broad transcriptions it is important to be consistent and not use allophones. Ashby et al. [2021] did this specifically for Bulgarian.
- **Linguistic variation and processes** [Ashby et al., 2021]: Some transcriptions include examples for monophthongization or deletion which are ongoing linguistic processes but should not be part of a dataset representing a standard variation. Ashby et al. [2021] dealt with monophthongization by choosing the longer to two transcriptions as this logically exclude the monophthonged version. This does of course only work if there are more than one pronunciations available.
- **Tie bars**: Ashby et al. [2021] notice that some languages (English and Bulgarian) have inconsistent use of tie bars. This can be correct by replacing all

inconsistencies by the tie-bar-version.

- **Errors in the transcriptions**: Gautam et al. [2021] noticed many errors in the WikiPron English data. They identified errors by looking at the least frequent phones and then check the word-pronunciation pairs where those phones occurred in. As the number of phones in a language is often known this can be used to check the phones in the datasets and identify uncommon ones.

Especially the task of finding errors in the transcriptions is quite tricky. It requires a lot of knowledge about the phonology and phonetics of a specific language.

3 Data Collection

The first important part of this thesis is concerned with data collection. Although phonetics is an important sub-area in linguistics, phonetic transcriptions are hard to find. If there are any transcriptions available, there are various hindrances that prevent it from being used as is. The following chapter outlines the different data types which are available and the different strategies that are used to convert the data into one well-formatted corpus. Apart from hindrances concerning sources and format, there are issues concerning the data itself. There are generally many more different pronunciations of a word than there are spellings. It is thus important to specify clearly what dialect or pronunciation convention a phonetic transcription follows.

3.1 Transcription Conventions

Another problem that needs be dealt with are different transcription conventions. There are different phonetic languages and within those there are different levels of transcription details. The most common are listed below.

IPA The International Phonetic Association (IPA) has one of the most common phonetic transcription conventions used in linguistics.

DISC The DISC convention is different from most of the others as it assigns exactly one ASCII code to each phone. The alphabet covers only Dutch, English and German phone inventories [R H. Baayen and Gulikers, 2021]. It is therefore very impractical for a multilingual corpus. However, it is still in use and can be found in some papers (e.g. Rao et al. [2015]).

In order to guarantee comparability, some transcriptions need to be translated into other transcription conventions.

Apart from different character sets there are different levels of detail. Not all transcriptions represent the phonetics in equal detail. Generally, there is the distinction of broad and narrow transcription. These two go back to the linguistic distinction of phone and phoneme. Broad refers to a phonemic description. Following the linguistic definition in chapter 3, this means that the transcription does not transcribe speaker specific pronunciations or dialectal variations. This kind of transcription is therefore less complex and usually easier to create and understand. Narrow transcriptions are phonetic. They present every speaker individual or dialectal sounds as exactly as possible. Although the spoken text in narrow and broad transcription sounds only minimally different, the two texts can diverge greatly. It is important to treat broad and narrow transcriptions as two different kinds of transcriptions.

(3.1) pɪˈk^{h}

(3.2) pɪˈk^{h}

Example 3.1 is a narrow (phonetic) transcription of the beginning of the Mapudungun version of the short story *The North Wind and the Sun*. The same text is transcribed broadly (phonemic) in example 3.2. As becomes clear in this example, the narrow transcriptions is much longer as it contains more different characters. The problem, with especially the narrow transcriptions, is that the transcriber still needs to define what narrow means in a specific case. This becomes tricky when given a task to automatically transcribe text, the training data might employ one definition of narrow, while there are texts in the test set that might follow another definition. However, in practice data is very rare, so in the end you would probably just use any data you can get.

3.2 Transcription Sources & Formats

Phonetic transcriptions of various languages are available from different sources in different formats. In order to use those, they have to be converted into simple text format in appropriate encoding that can easily be read and processed by a machine. The following subsections list the different data types and how they are used.

3.2.1 Full Text

For the task at hand, phonetic transcriptions in the form of fully transcribed texts would be ideal. As became clear, it is hardly possible to find those. There is plenty of material describing how different languages can be transcribed but those rarely contain fully transcribed text. If they do, it is mostly limited to one or a few sentences. The JIPA continuously published different phonetic transcriptions of a short story called "The North Wind and the Sun". A collection of those is available in a handbook of the JIPA which is only available as a pdf scan of the original book [Press, 2010]. While OCR is technically possible it turns out to be very difficult for IPA characters. The tools that exist do sometimes include IPA character recognition like the ABBYY FineReader which can be acquired for a fee. The CL institute at the UZH owns a version of the ABBYY tool but this version does not include the IPA module although ABBYY generally supports IPA character recognition. This ABBYY version was run on a JIPA pdf containing said phonetic transcriptions but the result could not be used. Mostly diacritics and special phonetic symbols were not correctly transcribed.

There are also open source tools. One of which is called tesseract. tesseract does not include the IPA alphabet. It is possible to train the model to include the IPA alphabet but this would need appropriate training data. Add quote

Some transcriptions have been published in separate issues as part of a collection of articles called "Illustrations of the IPA". While some of them are available in plain text format most of them are only available as pdfs or even images in text books. It is of course possible to manually type-write those which is what I did. More on how this is best done is explained in chapter 5 on experiments.

Additionally, some texts include short descriptions where certain pronunciations rules are explained which are not included in the transcriptions (especially stress).

3.2.2 Pronunciation Dictionaries

Another data type that is found quite often are lists of words' pronunciation. Those are sometimes referred to as pronunciation dictionaries. However, these often mean that there are words mapped to an audio representation which is not what is meant in this present case. Pronunciation dictionary in this present case refers to the mapping of an orthographic word to its pronunciation using phonetic symbols. Although such lists are very handy, especially as they can easily be used to train a transcription model, transcriptions of individual words and of entire texts are not exactly the same. There are two major problems:

- Pronunciation depends on the context of the word in question. Word forms are ambiguous and sometimes their pronunciation differs given on their specific context. add example
- Phonetic boundaries are not always equivalent with word boundaries. Spoken language sometimes merges certain words which leads to one phonetic unit.

There are phonetic symbols to represent such merging which often happens in, for example, French.

WikiPron

There exist databases of pronunciation dictionaries. Many of those do not release the mining software used to extend the database with more languages [Lee et al., 2020]. A very recent project that publishes pronunciation lists is WikiPron. The WikiPron project [Lee et al., 2020] is an open-source Python mining tool to retrieve pronunciation data from Wiktionary. Their database contains 1.7 million word/pronunciation pairs in 165 languages. Both, the database and the tool, are freely available online. Apart from the mining tool and the database, WikiPron can be used for grapheme-to-phoneme modelling. More on this subject will be discussed in chapter 4. In both G2P shared tasks organized by SIGMORPHON (see 4, data provided by WikiPron was used. For the 2021 task, WikiPron was improved and additional scripts were added based on feedback and findings in the 2020 task. One major improvement was concerned with languages written in different scripts. WikiPron supports now the detection of different scripts and languages can be sorted according to those scripts.

4 Models for Phonetic Transcription

This chapter introduces to the various methods and models that can be used to create phonetic transcription of various languages out of plain text.

4.1 Evaluation metrics

The most common metric to evaluate phonetic transcriptions is the WER. This is the percentage of predicted transcriptions that deviate from the gold standard. The lower, the better the model. The idea is that we can capture the cost that it takes to transform the system text into the reference text. If the WER is 0, this means that the texts are exactly the same. The following formula is used to calculate WER:

$$WER = \frac{S + I + D}{N} \tag{4.1}$$

In equation **??** the S stands for substitution, I for insertion, D for deletion and N denotes the total number of words in the reference sequence. If you want the percentage the number needs to be multiplied with 100. Note that the WER can be more than 100%. This happens if, for example, the are a lot of additional insertions or deletions in the system text. Another metric that is used quite often is the character error rate (CER). It is caculated in the exact same way as the WER, but instead of words everything is calculated on character basis. In a multilingual setting, it is sometimes necessary to have a score for the entire system covering more than one language. In such cases it is custom to use a macro-averaged WER or CER. explain macro average (and micro to be complete)

4.2 Neural G2P

4.3 Low-resource setting

Apart from a few well-studied examples, for most languages there is only little available data. It is therefore highly interesting and important to find solutions of how to deal with lack of data. Hammond [2021] submitted a system to the 2021 SIGMORPHON edition focusing on data augmentation methods. The primary goal of their approach was to test how successful a minimalist data augmentation model would be, knowing it would most probably not outperform any of the other models. They identified two approaches that might improve low-resource models. The first one is to use as much as possible of the development set for training. The second to train all languages together differentiating the languages only by a tag added to the word representations.

5 Experiments

This chapter presents the experiments and practical explorations that I conducted for this thesis. The previous chapters listed the different steps and problems that arise when trying to create and analyse a phonetic corpus.

5.1 Typewriting pdf phonetic transcriptions

In order to make use of as much data as possible, I used a software to manually transcribe the pdf scans.

The software allows to make use of neural Handwritten Text Recognition (HTR) models. There exists no pre-trained IPA model but I trained my own while transcribing the documents. On the website they mention that ideally training needs 5,000 - 10,000 words already transcribed. Although my available data is not nearly enough to train a reliable model, it was a great help to transcribe. As the scans where not handwritten text, the model still reached a surprisingly good quality. For the Hebrew transcription, the model reached a WER of 34.52% and a CER of 6.11%. The two main mistakes were made for two characters that were not even in the training data. The quality of the scans differed quite a lot which had an influence on the performance of the model as well. After transcribing another document I trained the model again and transcribed the remaining documents. The transcriptions got continuously better.

add more information on the model

5.2 Pronunciation dictionary coverage

In order to get an understanding of how many words are covered in the word lists, I created a script to calculate the coverage, the WER and the CER. I replaced the words in the texts with the words in the word list and compared it to the reference transcription.

In order to create full texts out of pronunciation dictionaries, I created a simple python script. While dealing with the full texts and the word lists, I noticed several things that are important when dealing with those texts.

- The pronunciation dictionaries sometimes included duplicates with different pronunciations. This is not surprising but still it needs to be handled well. A possible solution to this is to include PoS tags. Although this is generally possible it would mean an significantly greater effort which exceeds this thesis' scope. Another solution is to simply delete duplicate words. A close examination also showed that sometimes, the duplicate pronunciations are wrong. As it is the case with the English word "would". add this example indented
- For some full texts it is not clear, whether their transcription is narrow or broad. On the other hand, sometimes there is no broad or narrow word list available for a specific language but only one of those.
- The IPA allows to transcribe intonation segments. In German, those correspond mostly to punctuation marks like end of sentence symbols or commas. But this must not be true for every case. It needs to be decided if those should be kept or potentially deleted.

5.3 Automatic G2P

As there is only very little data available as full texts, we decided to use the short stories as test set for the experiments with G2P models. Those texts are all manually created and are specifically created by linguists for the purpose of studying phonetics of many languages.

References

100-language-sample. WALS Online - Languages. In M. S. Dryer and M. Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/languoid/samples/100`.

R. Aharoni and Y. Goldberg. Morphological Inflection Generation with Hard Monotonic Attention, 2016. URL `https://arxiv.org/pdf/1611.01487`.

L. F. Ashby, T. M. Bartley, S. Clematide, L. Del Signore, C. Gibson, K. Gorman, Y. Lee-Sikka, P. Makarov, A. Malanoski, S. Miller, O. Ortiz, R. Raff, A. Sengupta, B. Seo, Y. Spektor, and W. Yan. Results of the Second SIGMORPHON Shared Task on Multilingual Grapheme-to-Phoneme Conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.13.

M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2008.01.002. URL `https://www.sciencedirect.com/science/article/pii/S0167639308000046`.

E. Chodroff. Corpus Phonetics Tutorial, 2019. URL `https://eleanorchodroff.com/tutorial/index.html#`.

B. Comrie, M. S. Dryer, D. Gil, and M. Haspelmath. Introduction. In M. S. Dryer and M. Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL `https://wals.info/chapter/s1`.

CrashCourse, 2021. URL `https://www.youtube.com/watch?v=-sUUWyo4RZQ&list=PL8dPuuaLjXtP5mp25nStsuDzk2blncJDW&index=18`.

V. Gautam, W. Li, Z. Mahmood, F. Mailhot, S. Nadig, R. Wang, and N. Zhang. Avengers, ensemble! benefits of ensembling in grapheme-to-phoneme prediction. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–147, 01 2021. doi: 10.18653/v1/2021.sigmorphon-1.16.

K. Gorman, L. F. Ashby, A. Goyzueta, A. McCarthy, S. Wu, and D. You. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.2. URL `https://aclanthology.org/2020.sigmorphon-1.2`.

M. Hammond. Data augmentation for low-resource grapheme-to-phoneme mapping. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 126–130, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.14. URL `https://aclanthology.org/2021.sigmorphon-1.14`.

H. H. Hock and B. D. Joseph. *Language History, Language Change, and Language Relationship*. De Gruyter, 2019. ISBN 9783110613285. doi: 10.1515/9783110613285.

N. Jaitly, D. Sussillo, Q. V. Le, O. Vinyals, I. Sutskever, and S. Bengio. A neural transducer, 2016.

M. Kracht. *Introduction to linguistics*. Los Angeles, 2007. URL `https://linguistics.ucla.edu/people/kracht/courses/ling20-fall07/ling-intro.pdf`.

J. L. Lee, L. F. Ashby, M. E. Garza, Y. Lee-Sikka, S. Miller, A. Wong, A. D. McCarthy, and K. Gorman. Massively Multilingual Pronunciation Modeling with WikiPron. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.521`.

M. Y. Liberman. Corpus Phonetics. *Annual Review of Linguistics*, 5(1):91–107, 2019. ISSN 2333-9683. doi: 10.1146/annurev-linguistics-011516-033830.

R. Y.-H. Lo and G. Nicolai. Linguistic knowledge in multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–140, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.15. URL `https://aclanthology.org/2021.sigmorphon-1.15`.

P. Makarov and S. Clematide. CLUZH at SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–176, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.19. URL `https://aclanthology.org/2020.sigmorphon-1.19`.

T. McEnery and A. Hardie. *Corpus Linguistics: Method, theory and practice*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge, 2011. ISBN 9780511981395. doi: 10.1017/CBO9780511981395.

C. U. Press. The principles of the international phonetic association (1949). *Journal of the International Phonetic Association*, 40(3):299–358, 2010. doi: 10.1017/S0025100311000089.

R. P. R H. Baayen and L. Gulikers, 2021. URL `https://catalog.ldc.upenn.edu/docs/LDC96L14/`.

K. Rao, F. Peng, H. Sak, and F. Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229, 2015.

Sigmorphon. SIGMORPHON - Special Interest Group on Computational Morphology and Phonology, 2021. URL `https://sigmorphon.github.io/`.

SPUR project. Non-randomness in Morphological Diversity, 2021. URL `https://www.spur.uzh.ch/en/departments/research/textgroup/MorphDiv.html`.

S.-A. Toma and D. Munteanu. Rule-based automatic phonetic transcription for the romanian language. *Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, Computation World*, 0:682–686, 11 2009. doi: 10.1109/ComputationWorld.2009.59.

X. Yu, N. T. Vu, and J. Kuhn. Ensemble self-training for low-resource languages: Grapheme-to-phoneme conversion and morphological inflection. In *Proceedings* *of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 70–78, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.5. URL https://aclanthology.org/2020.sigmorphon-1.5.

A Tables

Table 1: The table shows a list of the 100 languages in the corpus including information on the language families.

Iso639-3	Name WALS	Family WALS
abk	Abkhaz	Northwest Caucasian
amp	Alamblak	Sepik
aey	Amele	Trans-New Guinea
apu	Apurinã	Arawakan
bmi	Bagirmi	Central Sudanic
bsn	Barasano	Tucanoan
gry	Grebo	Niger-Congo
eus	Basque	Basque
ape	Arapesh (Mountain)	Torricelli
bsk	Burushaski	Burushaski
ram	Canela-Krahô	Macro-Ge
tzm	Berber (Middle Atlas)	Afro-Asiatic
cha	Chamorro	Austronesian
ckt	Chukchi	Chukotko-Kamchatkan
zoc	Zoque (Copainalá)	Mixe-Zoque
dgz	Daga	Dagan
hae	Oromo (Harar)	Afro-Asiatic
arz	Arabic (Egyptian)	Afro-Asiatic
fij	Fijian	Austronesian
fin	Finnish	Uralic
gni	Gooniyandi	Bunuban
khk	Khalkha	Altaic
hau	Hausa	Afro-Asiatic
heb	Hebrew (Modern)	Afro-Asiatic
hin	Hindi	Indo-European

Iso639-3	Name WALS	Family WALS
hix	Hixkaryana	Cariban
hnj	Hmong Njua	Hmong-Mien
qvi	Quechua (Imbabura)	Quechuan
imn	Imonda	Border
ind	Indonesian	Austronesian
kal	Greenlandic (West)	Eskimo-Aleut
kyh	Karok	Karok
gyd	Kayardild	Tangkic
kio	Kiowa	Kiowa-Tanoan
cku	Koasati	Muskogean
kor	Korean	Korean
ses	Koyraboro Senni	Songhay
kgo	Krongo	Kadu
kut	Kutenai	Kutenai
lkt	Lakhota	Siouan
laj	Lango	Eastern Sudanic
lvk	Lavukaleve	Solomons East Papuan
lez	Lezgian	Nakh-Daghestanian
dni	Dani (Lower Grand Valley)	Trans-New Guinea
lue	Luvale	Niger-Congo
ayz	Maybrat	West Papuan
myh	Makah	Wakashan
cmn	Mandarin	Sino-Tibetan
mpc	Mangarrayi	Mangarrayi-Maran
mni	Meithei	Sino-Tibetan
arn	Mapudungun	Araucanian
mrc	Maricopa	Hokan
vma	Martuthunira	Pama-Nyungan
mph	Maung	Iwaidjan
ote	Otomí (Mezquital)	Oto-Manguean
ell	Greek (Modern)	Indo-European
naq	Khoekhoe	Khoe-Kwadi
scs	Slave	Na-Dene
tur	Turkish	Altaic
kat	Georgian	Kartvelian
kan	Kannada	Dravidian
mya	Burmese	Sino-Tibetan

Iso639-3	**Name WALS**	**Family WALS**
jpn	Japanese	Japanese
one	Oneida	Iroquoian
pwn	Paiwan	Austronesian
gug	Guaraní	Tupian
myp	Pirahã	Mura
crk	Cree (Plains)	Algic
plt	Malagasy	Austronesian
jac	Jakaltek	Mayan
rma	Rama	Chibchan
rap	Rapanui	Austronesian
rus	Russian	Indo-European
sag	Sango	Niger-Congo
mig	Mixtec (Chalcatongo)	Oto-Manguean
xsu	Sanuma	Yanomam
spa	Spanish	Indo-European
fra	French	Indo-European
eng	English	Indo-European
deu	German	Indo-European
spp	Supyire	Niger-Congo
swh	Swahili	Niger-Congo
tgl	Tagalog	Austronesian
tml	Asmat	Trans-New Guinea
tha	Thai	Tai-Kadai
tiw	Tiwi	Tiwian
bhq	Tukang Besi	Austronesian
vie	Vietnamese	Austro-Asiatic
wyb	Ngiyambaa	Pama-Nyungan
wba	Warao	Warao
pav	Wari'	Chapacura-Wanham
pes	Persian	Indo-European
kew	Kewa	Trans-New Guinea
kjq	Acoma	Keresan
wic	Wichita	Caddoan
mzh	Wichí	Matacoan
yad	Yagua	Peba-Yaguan
yaq	Yaqui	Uto-Aztecan
yor	Yoruba	Niger-Congo

Iso639-3	**Name WALS**	**Family WALS**
zul	Zulu	Niger-Congo