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

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

HTR Handwritten Text Recognition

1 Introduction

1.1 Motivation

Comparing languages and studying their similarities and differences is part of a well-established branch of traditional linguistics called comparative linguistics.

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

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

The text group of the Language and Space lab at the University of Zurich maintains a project that provides 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.

The collection a corpus of phonetic transcriptions can be useful for different tasks other than for this current thesis. Phonetic transcriptions of written text can be an important step in the process of speech technologies (like speech synthesis). add more on usage of phonetic transcriptions

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?

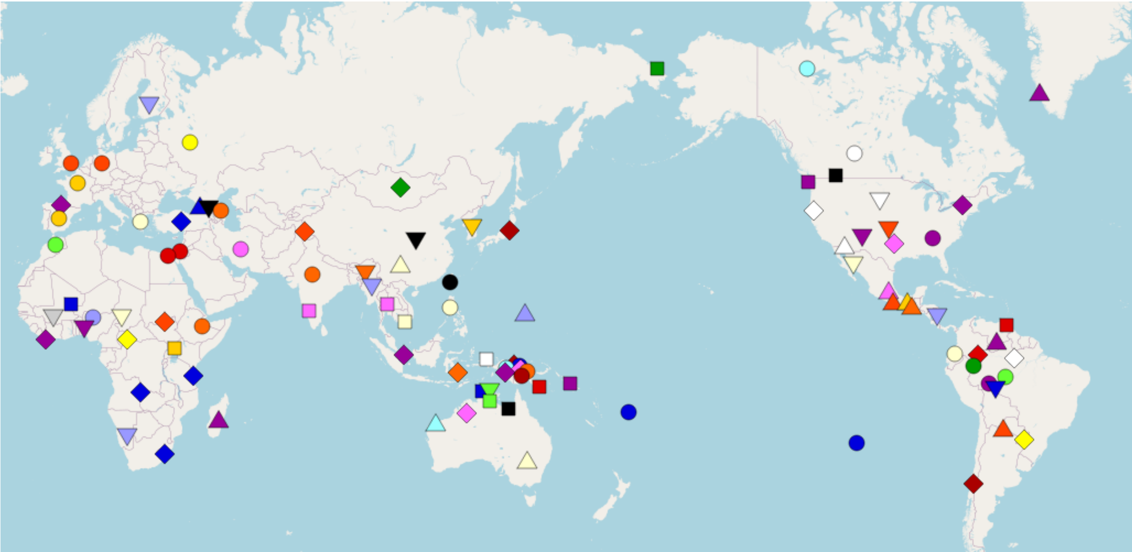

Figure 1: WALS - 100 Language Sample

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*. 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 is clear while the definition on the computational side is often less strict. definition of phoneme / phone, i.e. phonemic, phonic, correct linguistic one and then the one that is used e.g. in Lee et al. [2020], foot note 4

Writing systems

Unlike spoken language that was a part of human interaction all the time, writing systems only developed over time. A single grapheme can represent a sound ... (video)

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, 19.07.2019]. Many methods and tools used in corpus phonetics are based on automatic speech recognition (ASR) algorithms or simple programming [Chodroff, 19.07.2019].

2.5 Automated phonetic transcription

Creating phonetic transcriptions is essentially a Sequence-to-sequence (seq2seq) task. Like other Natural Language Processing (NLP) tasks its goal is to transform a sequence of characters into another sequence of characters.

2.5.1 Rule-based models

The first systems of 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 a few language 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. A problem with rule-based approaches is mainly 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 which is caused by constant language change and some flexibility in language use, all systems must be able to deal with rare and unseen words. 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 examples considered only one language (see e.g. [Toma and Munteanu, 2009]).

2.5.2 N-gram Models / Statistical models

These are sometimes referred to as traditional models.

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].

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.

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 has one of the most common phonetic transcription conventions used in linguistics.

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.

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

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".

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.

3.3 Transkribus

In order to make use of as much data as possible, a software is used 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 Word Error Rate (WER) of 34.52% and a Character Error Rate (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.

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 Rule-based systems

added section on literatur background in chapter two, not sure what to do with this here...

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)

4.2 Neural G2P

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. The definition of phoneme in this context does not satisfy the precise linguistic definition. It is therefore to note that phoneme used here simply refers to the transcription symbols used to represent a specific sound in a specific language.

Other ways of grapheme-to-phoneme conversion will be described here (apart from the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON) tasks which will be described below).

As with many NLP tasks, research focussed mainly on English or other languages where a lot of data is easily available [Gorman et al., 2020].

The SIGMORPHON [Sigmorphon, 2021] regularly organizes shared tasks concerned with morphology and phonology. For the years 2020 and 2021 they organized a grapheme-to-phoneme 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 4.2.2 and 4.2.3. 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.

4.2.1 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.

4.2.2 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. The system is not available online, but I asked the author if I could use it for my experiments.

4.2.3 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. 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]. add more details on data preparation, p.127

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.

Results in the low-resource setting are still worse compared to the medium-resource setting.

Error analysis

Error analysis is important for future models.

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 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 and training a model to do OCR

5.2 Creating full texts out of pronunciation dictionaries

In order to create full texts out of pronunciation dictionaries, I created a simple python script. There were several problems that needed to be addressed in order to create 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 Part-Of-Speech (PoS) tags. Although this is generally possible it would mean an significantly greater effort which might exceed this thesis' scope.
- Not all words are in the dictionaries. This problematic as the texts cannot be fully transcribe in this way. A solution would be to transcribe only sentences and then pick those sentences that are fully transcribed. Another solution could be to have statistics of the most frequent missing words (in English 'and' is missing) and either transcribe them manually if possible, find transcriptions in the internet or check in the narrow / broad pronunciation dictionary and use this word instead. The latter possibility might corrupt the data, but as there are not that many missing words, it is worth a try. Another way to deal with missing words is to iteratively split the words in two or possibly three

or more parts and check whether the subwords are in the dictionary. This depends on the specific language. In the case of English it was necessary to exclude splitting of the first and the last character separately as those are in the dictionary but with their alphabet reciting pronunciation which is typically not used within a word. For this approach it is necessary to have some language specific expertise.

- The IPA allows to transcribe intonation segments. In German, those correspond mostly to punctuation like end of sentence symbols or commas. But this must not be true for every case. In order to include those a close examination of the

There are several decisions to be made.

5.3 Automatic G2P

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.

E. Chodroff. Corpus Phonetics Tutorial, 19.07.2019. 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`.

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.

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.

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 on 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