It sounds like Greek to me

Performance of phonetic representations for language identification

Johannes IJpma[1] , **Marco Loog**[1] , **Tom Viering**[1] , **Stravos Makrodimitris**[1] , **Arman Naseri Jahfari**[1]

[1]TU Delft

{d.j.ijpma}@student.tudelft.nl, {m.loog, t.j.viering, s.makrodimitris, a.naserijahfari}@tudelft.nl

Abstract

This paper compares the performance of two phonetic notations, IPA and ASJPcode, with the alphabetical notation for word-level language identification. Two machine learning models, a Multilayer Percerptron and a Logistic Regression model, are used to classify words using each of the three notations. With both models the IPA notation outperforms the other two notations.

1 Introduction

Over the past decades, breakthroughs have been achieved in the field of Natural Language Processing enabling automatic translation, classification and text generation[15] [4]. And with it, the field of Computational Historical Linguistics which concerns itself with tasks such as automatic assessment of genetic relatedness and cognate detection [8]

This field makes use of different phonetic representations of words. Two major phonetic notations are The International Phonetic Alphabet (IPA) and ASJPcode. ASJP is a simplification of IPA introduced by [3]. The theoretical advantages and disadvantages of using ASJP over IPA have been discussed by [16]. But their performance difference in language identification task has, to our knowledge, never been tested.

In this paper, we investigate the effects of these representations on the task of single word language identification. We focus on the following question: "How well can machine learning models predict the language of input words written in IPA and ASJP?"

IPA is an alphabetic system designed to represent the speech sounds of any spoken language. It is a useful representation since it gives us a way to study sounds of all possible human languages in a single universal alphabet. This system has, among others, the following useful properties:

1. Each sign should have a distinctive sound.
2. The same sign should be used for the same sound across all languages.

These properties give additional information compared to the word written in their original alphabets. Where in, for example, the English language the pronunciation of the of *ch* differs in each of the words *chest, chef* and *chemist* these differences are represented when written in IPA ([/tst/, /f/ and /kmst/]). [2]. This extra information could be used to achieve more accurate language classification with machine learning models. To investigate the main research question this paper answers the following sub-questions:

Q1. Does either the IPA, ASJPcode or the original alphabetical representation consistently outperform the other for language identification?

Q2. How dependent are machine learning models on the order of the characters of the word and their IPA and ASJPcode representation for language identification? Is there a difference between the three notations?

We use Logistic Regression and Multilayer Perceptron models to find answers to these questions.

This paper is structured as follows: section 2 describes the necessary background and section 3 discusses previous work. In sections 4 and 5 describe the methodology and a discussion of the results. Conclusions are drawn in section 6 and finally, section 7 discusses responsible research and reproducibility.

2 Background

International Phonetics Alphabet

The International Phonetics Alphabet is an alphabetic system describing distinct sounds used in human spoken language. It is largely based on the Latin alphabet and consists of It consists of 107 letters, 52 diacritics, and four prosodic marks. Every distinctive sound has its own character.

Automated Similarity Judgment Program

The Automated Similarity Judgment Program (ASJP) database was introduced by [3] to aid the classification of languages through automated lexical comparison. The database consists of 40-item lists of words for more than half of all languages. Together with the database came the introduction of ASJPcode, a standard orthography and simplification of IPA that exclusively uses symbols found on a QWERTY keyboard.

ASJPcode consists of 41 symbols (7 for vowels and 34 for consonants). This orthography has been criticized for its oversimplification. According to [7] "it is clear that what are being compared are tiny subsets of linguistic entities that are structurally simplified compared with their source forms, and these subsets are then asked to stand as useful representative samples of their respective namesakes"

Despite its shortcomings this database is used in computational historical linguistic research [12] [5]

3 Previous Work

Previous work on language identification has focused on identifying the language of single words by training an MLP using data scraped from Wikipedia [6], or on using audio fragments by training a CNN on their corresponding spectrograms [13] [11].
The IPA and ASJP phonetic notations are often used in Computational Historical Linguistic research. For example in language classification [3], automated dating of language families [7], and cognate classification [9]. However, they have not yet been applied to the task of language identification.

4 Methodology

The ipa-dict and CogNet datasets are used for the experiments. This section describes them and the architectures of the machine learning models used in more detail. The results can be found in the next section.

Ipa-dict datatset

The data used for this research is acquired from the ipa-dict Github repository [10] which provides lists of words paired with their corresponding pronunciation for 24 languages. These include different versions of the same language for English (British and American), Spanish (Spanish and Latin American), Chinese (Original and Simplified) and Vietnamese (North, South and Central). An example of the data can be found in Figure 1

Word	IPA
Aasgeier	/aːsˈgəiːɐ̯/
Abakus	/ˈʔapˌʔɑːkus/
abändere	/ˈʔapˌʔɛndɐ̯ə/

Figure 1: Example ipa-dict German

The dataset is far from balanced. Figure 2 shows the number of word-IPA entries for each available language.

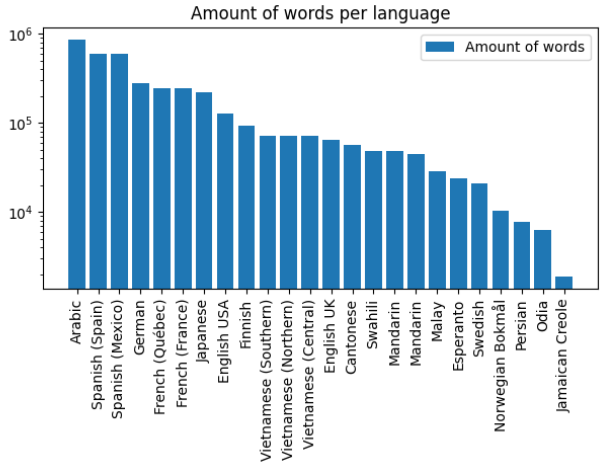

Figure 2: Amount of words available in each language in ipa-dict

CogNet Database

The CogNet Database [1] is a large database of cognates, which are words with a common origin, either from having the same parent language or from borrowing from other languages. This database comes with phonetic transcriptions of words in both IPA and ASJP. A sample of this dataset can be found in Figure 3. The language field of this dataset contains 884 different values. And while a lot are valid ISO 639-2 or ISO 639-1 language codes, a lot of these entries are not useful for this research. Apart from that a lot of the languages in this database have less than one hundred data points. Because of this, only the 25 languages with the most entries are used in the experiments. These languages have between 4000 and 56000 data points.

lemma	IPA	ASJP	language
dictionary	dɪkʃ(ə)n(ə)ɹi	dikS3n3ri	en
gratis	ˈɣraːtɪs	xratis	nl

Figure 3: CogNet word pronunciation data example

Data preprocessing

Most machine learning models need a fixed feature vector as their input. To achieve this with the data consisting of strings with varying lengths, words longer than 30 characters are removed after which the data is one-hot encoded. Since each character at each of the 30 possible positions now represents an individual feature it is important to know the size of each alphabet for each of the languages classified, since a large alphabet can cause the high dimensionality of the one-hot encoded feature vector.
While the majority of the languages are made up of less than 200 characters, Mandarin (original and simplified), Cantonese and Japanese have several (tens of) thousands of characters. This made one-hot encoding these languages infeasible. How many characters make up the alphabets of each of the 24 languages is shown in Figure 4

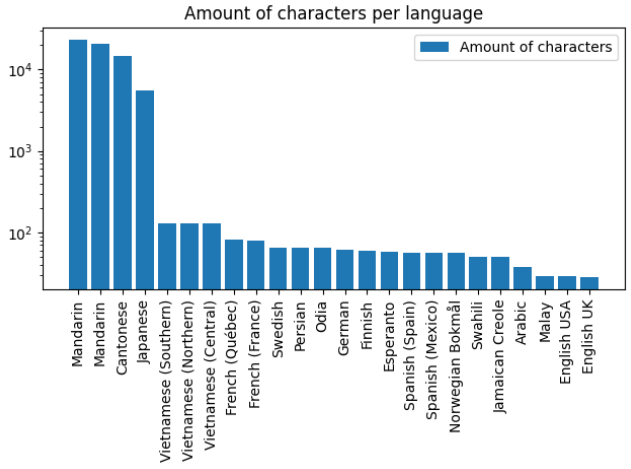

Figure 4: Amount of characters compiling each language in their own alphabet

When limited to the IPA, this was not an issue because it is limited to 163 characters. See Figure 5 for the number

of characters in each language after converting them to IPA. Since the ASJPcode only consists of 41 symbols, it cannot cause this issue either.

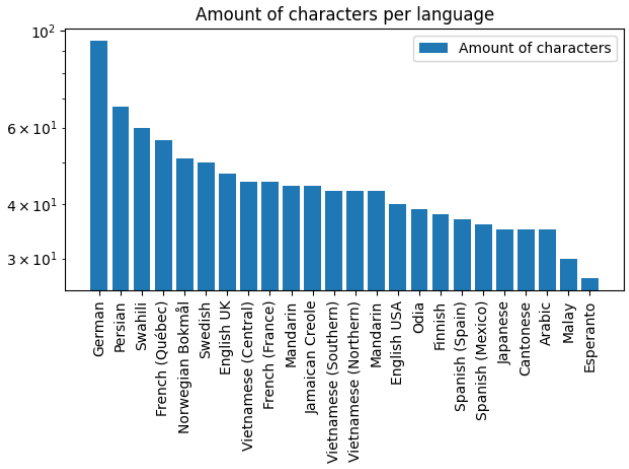

Figure 5: Amount of characters compiling each language in the IPA

Experiment 1: Performance of different representations To answer Q1 the Logistic Regression model and MLP have been fed the on-hot encoding of either the normal words, IPA or ASJP of a set of five languages, [German, English, Finnish, French, Esperanto] languages. When possible the IPA data from the ipa-dict is preferred because its pronunciations have been manually checked [10]. These languages were chosen because they were well represented in both the ipa-dict and CogNet datasets. The data used in the experiment is an intersection of these five languages from the ipa-dict and CogNet's wikt.phonetic_words.tsv. For each common entry, the word and IPA transcription from ipa-dict are used and the ASJP transcription from the CogNet dataset. A random sample of 5000 words for each language was chosen to create a balanced sample of the data, with replacement since some of the languages were just short of this threshold. A stratified sample of 20% of this subset was used as a test set.
The logistic regression models used L1 regularization and Scikit-learn's liblinear solver.
The MLP consists of two hidden layers of 150 nodes with a Sigmoid activation function. Because of the limited amount of data a dropout layer is added between the two hidden layers with a dropout rate of 0.5. The output layer makes use of the soft-max activation function. The model makes use of an Adam optimizer and categorical cross-entropy loss function and is trained for 7 epochs.
10-fold cross-validation is used for both models Next, three Logistic Regression models and three Multilayer perceptrons were trained using these same words, but each of the three on a different representation (Alphabetical, API or ASJP). The hypothesis is the IPA representation will outperform the ASJPcode since the latter is a simplification. The big questions are how significant the impact of these simplifications are and whether ASJPcode can outperform the normal text in this task. The results are shown in Figures 6 and 7.

Experiment 2: Impact character ordering To answer Q2 the models are trained on the number of occurrences of a character in the words, meaning that any information to the order of the characters is lost. This experiment aims to see whether the occurrence of certain symbols is enough to correctly identify the language. The same models and data are used as in the previous experiment. The words are now represented as a feature vector containing integer values. How often a character occurs in a word is denoted by the integer value at the corresponding position in the array. Do these models just learn that specific characters are most likely to belong to a certain language or do they also use the position and relation to other characters, and how big is this impact for the different word representations? The results are shown in Figure 9 for the MLP and Figure 13 for Logistic Regression.

Experiment 3: Dependence of accuracy on common characters of two alphabets According to the outcomes of experiment 1, the IPA notation outperforms the ASJPcode, which in turn has higher accuracy than the alphabetical notation. It could be that, since most languages use a form of the Latin script as their alphabet, languages in the alphabetical form are harder to distinguish simply because they have more characters in common when written this way.
To test this hypothesis the MLP is trained on 300 language pairs. This is repeated three times, one for each notation. The resulting accuracies of each language pair are plotted against the percentage of characters the two languages have in common. This experiment does not make use of cross-validation because it would make the training of the, in total, 900 language pairs too time-consuming. Figure 10 displays a scatterplot of the results. The MLP has the same architecture as in experiment 1. A smaller sample of 4000 characters per language is used and the model is trained for just three epochs. This experiment uses just 25 languages from the CogNet phonetic_words database.

5 Results

Performance phonetic representation

The results of the first experiment are shown in Figures 6 and 7. Figure 6 shows the mean and variance of the accuracies of the MLP for the words (Alpha stands for the original script), their IPA and ASJPcode transcriptions after 10-fold cross-validation. It can be seen that the IPA transcription indeed outperforms the ASJPcode, and that both outperform the original scripts. While this is an indication that IPA could be preferable these results are not conclusive proof that IPA will always outperform the other notations in language identification. The accuracies of the words and of the ASJPcode might achieve similar performance when using a larger dataset.

	Word	IPA	ASJP
Accuracy mean	0.765	0.976	0.813
Accuracy variance	4.32e-05	0.59e-05	8.07e-05

Figure 6: Mean and variance of the MLP accuracy after 10-fold cross validation on words, their IPA and ASJPcode representation

Figure 7 shows the results of the first experiment for the Logistic Regression model. The results of the Logistic Regression model are in line with the results from the MLP.

	Word	IPA	ASJP
Accuracy mean	0.795	0.980	0.830
Accuracy variance	4.1e-05	0.5e-05	5.3e-05

Figure 7: Mean and variance of the Logistic Regression model accuracy after 10-fold cross validation on words, their IPA and ASJPcode representation

Performance without character ordering

To determine whether the order of the characters mattered for the performance of the models they have been trained on just the occurrences of the characters. The mean and variance of the accuracy of the MLP after 10-fold cross-validation can be found in Figure 8. The Difference in average accuracy compared to the one-hot encoded data of the previous experiment can be found in Figure 9 The corresponding figures of the logistic regression model can be found in Figures 12 and 13 in Appendix B.

	Word	IPA	ASJP
Accuracy mean	0.662	0.965	0.737
Accuracy variance	1.83e-04	1.32e-05	8.62e-05

Figure 8: Mean and variance of the MLP accuracy after 10-fold cross validation on the amount of characters occurring in the words, their IPA and ASJPcode representation

As can be seen in figure 9 the impact of the loss of information of the order of characters barely affects the performance of the MLP. The accuracy using this representation is just a percentage lower. The words in their alphabetical notation are affected the most, with the accuracy decreasing about 10%. The accuracy using the ASJPcode decreased about 8%.

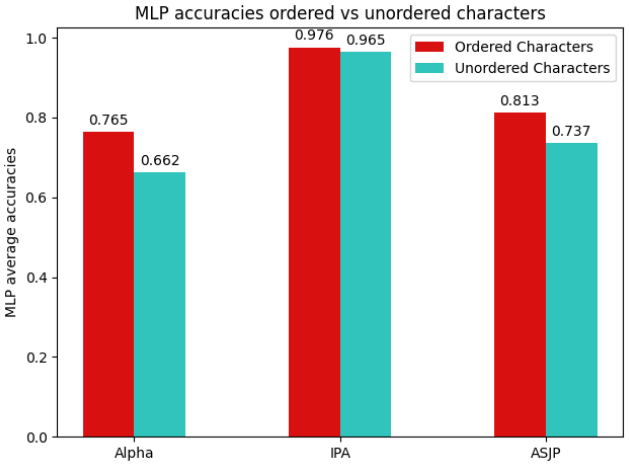

Figure 9: Accuracy MLP on words, their IPA and ASJPcode representation with and without information as to the order of the characters

Performance vs common characters

To determine whether the difference in performance can be caused by the amount of characters languages have in common, the accuracies of pairs of languages are compared with the ratio of characters common in both languages to the number of unique characters. Figure 10 displays a scatter plot in which each data point is a language pair. The figure shows no correlation between the accuracy and the ratio of common characters to the total amount of unique characters. What is interesting about this figure is the small blue cluster in the top left which consists of languages with 0 to 3 per cent characters in common and the words represented in their alphabet. While not visible in the scatter-plot this cluster contains 105 out of the 300 language pairs for this representation. Using a phonetic notation for language identification might not be beneficial if all languages to classify have different alphabets.

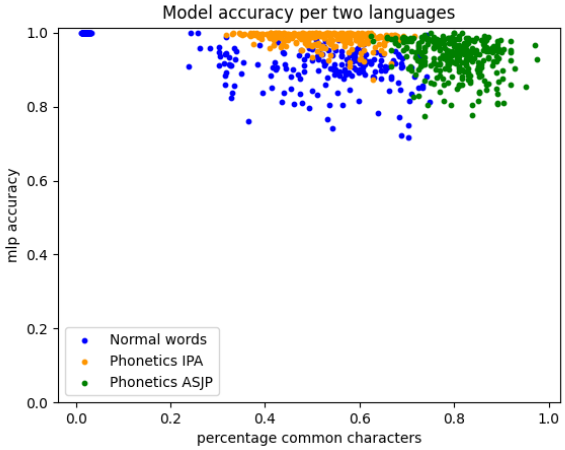

Figure 10: Each data point is the accuracy of the MLP trained on two languages against the ratio of characters used by both languages to the total amount of unique characters of both languages combined for each of the representations (normal words, IPA and ASJPcode)

6 Discussion & Future Work

This research compares the performance of data in three different representations for language identification, how recognizable languages are by just the occurrence of the symbols and one possible explanation of the different accuracies.
The accuracy of both machine learning models experiment 1 for the alphabetic representation is lower than accuracies achieved in previous work [6]. This might be because the size of the data sample used in our experiments is than 10% of that used to train their model. Previous work on language identification using a CNN trained on 3.75 second long audio clips achieved an 89% accuracy [13]. The CNN classified six languages and was trained on 5000 audio samples per language.
The third experiment aims investigates a possible reason for the high accuracy of the IPA data compared to the other two representations. There are, however, a lot of contributing factors that have not been taken into account. One of them being the distribution of the characters of the two languages classified. If, for example, a character appears only a single time in the fist language and in all instances of the second, that character by itself is a strong identifier as to which language a word belongs.

7 Conclusion & Future Work

This research aims to compare the performance of alphabetical, IPA and ASJPcode notations for word-level language identification. The experiments conducted in this paper indicate that the IPA phonetic notation might lead to better results for language identification than the alphabetical notation or ASJPcode phonetic notation.
The open question is still why the IPA notation performs so much better than the other two. This could be studied using the permutation feature importance measure or by applying the Lime explanation technique [14]. Another question is whether the difference in performance still exists when the models are trained on more data. Previous work suggests that the MLP can perform better using the alphabetical notation than our model has when trained on a larger dataset [6].

8 Responsible Research

The research in this paper has been done for the course CSE3000 at Delft University of Technology. The data used for this research has been acquired from public Github repositories with an MIT or Attribution-NonCommercial-ShareAlike 4.0 International license. Data samples and model weights have been saved for each experiment and can be found at the Gitlab repository [1]
All changes made to the data and removal of data have been described in sections 4 and 5.

References

[1] Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Adam Brown. International phonetic alphabet. *The Encyclopedia of Applied Linguistics*, 2012.

[3] Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308, 2008.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Volker Gast and Maria Koptjevskaja-Tamm. The areal factor in lexical typology. In *Aspects of linguistic variation*, pages 43–82. De Gruyter Mouton, 2018.

[6] Tom Ham. Language recognition using deep neural networks, Aug 2018.

[7] Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875, 2011.

[8] Gerhard Jäger. Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182, 2019.

[9] Gerhard Jäger and Pavel Sofroniev. Automatic cognate classification with a support vector machine. In *Proceedings of the 13th Conference on Natural Language Processing*, volume 16, pages 128–134. RUB Bochum, 2016.

[10] Doherty Liam. ipa-dict - monolingual wordlists with pronunciation information in ipa. https://github.com/open-dict-data/ipa-dict, 2016. Accessed: 2021-03-16.

[11] Gregoire Montavon. Deep learning for spoken language identification. pages 1–4, 2009.

[12] Taraka Rama and Lars Borin. N-gram approaches to the historical dynamics of basic vocabulary. *Journal of Quantitative Linguistics*, 21(1):50–64, 2014.

[13] Shauna Revay and Matthew Teschke. Multiclass language identification using deep learning on spectral images of audio signals. *arXiv preprint arXiv:1905.04348*, 2019.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[15] Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[1] https://gitlab.ewi.tudelft.nl/cse3000/2020-2021/rp-group-35/rp-group-35-common

[16] Søren Wichmann, Taraka Rama, and Eric W Holman. Phonological diversity, word length, and population sizes across languages: The asjp evidence. 2011.

A MLP achitecture

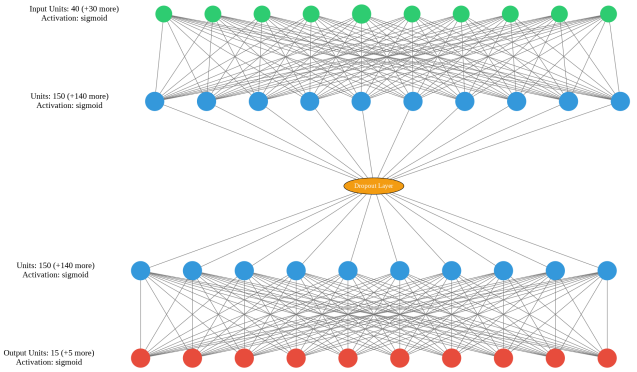

Figure 11: The achitecture of the mlp used in the experiments

B Results Logistic Regression model experiment 2

	Word	IPA	ASJP
Accuracy mean	0.665	0.968	0.741
Accuracy variance	9.92e-05	1.20e-05	4.20e-05

Figure 12: Mean and variance of the Logistic Regression model accuracy after 10-fold cross validation on words, their IPA and ASJP-code representation

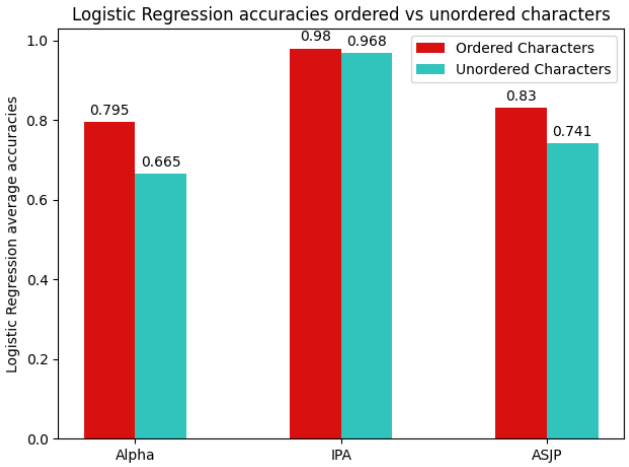

Figure 13: Accuracy Logistic Regression model on words, their IPA and ASJPcode representation with and without information as to the order of the characters