Examining the Relation Between Orthographic Complexity and Automatic Speech Recognition Performance

Annika Shankwitz
Indiana University, Department of Linguistics

Abstract

While the relations between orthographic complexity (OC), reading performance, and visual\auditory word processing have been examined, the relation between OC and automatic speech recognition (ASR) performance has yet to be explored. The current project investigates this relationship by comparing the OC and ASR performance of 11 languages. Ultimately, no obvious relation between OC and ASR was found. This lack of obvious relation can be explained by considering that ASR models operate on phonetic, not phonemic, forms.

1 Introduction and Background

The relation between orthographic complexity (OC), reading difficulties (Katz and Frost, 1992), and visual\auditory word processing (Frost and Katz, 1989) have been examined; however, to the best of my knowledge, the relation, if one exists, between OC and automatic speech recognition (ASR) performance has not been explored. The goal of this project is to explore said relation.

OC, also known as orthographic depth, describes the reliability of mappings between graphemes and phonemes in a given language. A language with low OC has mainly 1-to-1 mappings, whereas a language with high OC may have a number of 1-to-many or many-to-1 mappings. An example of a 1-to-many mapping can be seen in Figure 1.

ASR refers to the transcription of spoken utterances into a written form, most frequently into graphemes. In some ways then, ASR mirrors a phoneme-to-grapheme (p2g) conversion task. Because of this, it is reasonable to question if the OC of a language affects its ASR performance. Intuitively, it would be expected that languages with a higher OC, i.e. more complex/less reliable grapheme-to-phoneme (g2p) mappings, would perform worse during ASR than languages with lower a OC, i.e. less complex/more reliable g2p mappings. Worse performance could possibly surface as lower learning rates, or as higher character error rates.

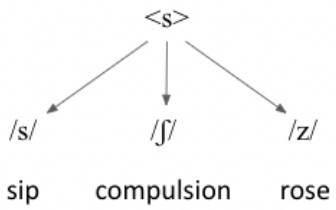

Figure 1: Grapheme to Phoneme Correspondence of English <s> (Carr, 2019)

To answer this question, the OC of 11 languages were calculated. Then, for each language, three ASR models were implemented which successively doubled in amount of training data; namely from 2.5 hours, to 5 hours, to 10 hours. OC scores were then compared with ASR performance across increments of training hours.

2 Methods & Materials

2.1 Selected Languages

In order to be selected, a given language had to meet three criteria.

1. The language must have at least 7,500 words with IPA transcriptions on WikiPron (Lee et al., 2020), or a corpus containing at least 7,500 words and Epitran[1] support. (Mortensen et al., 2018).

[1]Epitran transliterates orthographic forms into phonemic forms.

2. The language must have at least 10 hours of training data available through Common Voice (Ardila et al., 2020).

3. The language must be part of openAI's multilingual Whisper model (Radford et al., 2022).

The languages which fulfilled these requirements were Catalan, Dutch, English, Farsi, German, Georgian, Italian, Japanese, Polish, Portuguese, Russian, Spanish, Turkish, Thai, Tamil, Swahili, and Welsh. Due to time constraints, only Dutch, Farsi, Georgian, Polish, Portuguese, Russian, Turkish, Thai, Tamil, Swahili, and Welsh. were completely processed.

2.2 Orthographic Complexity

While there are a number of theoretical approaches to examine OC (Borleffs et al., 2017; Katz and Frost, 1992; Schmalz et al., 2015), these approaches do not produce explicit scores, and therefore pose difficulty for the comparison at hand. For this reason, OC was calculated with the Orthographic Transparency Estimation Artificial Neural Network model (OTEANN) (Marjou, 2021), which does produce explicit scores. OTEANN treats g2p and p2g conversion as translation tasks. The model is trained on *(orthographic form*, *IPA form*) pairs, from which p2g and g2p mappings are learned. During testing, the model uses these learned mappings to predict IPA forms given orthographic forms and vice-versa. The model produces two scores: a read score, the proportion of correctly predicted IPA forms (g2p), and a write score, the proportion of correctly predicted orthographic forms (p2g). A predicted form is considered correct only if it is completely correct - partial matches are not counted.

A lexicon of 7,500 pairs was used for each language. Pairs were mainly collected from WikiPron (Lee et al., 2020), with a priority on broad transcriptions (underlying representations) (Dutch, Polish, Portuguese, Welsh); however, if broad transcriptions weren't available, narrow transcriptions (surface representations) were used instead (Georgian, Russian). If more than 7,500 pairs were available for a given language, 7,500 pairs were selected at random. If there were less than 7,500 pairs available, 7,500 orthographic forms were sourced from text corpora (Tamil, Turkish, Farsi, Swahili) (Ramasamy and Žabokrtský, 2012; Marşan et al., 2022; Rasooli et al., 2020; Leipzig Corpora Collection, 2017), at random and IPA transcriptions were generated with Epitran (Mortensen et al., 2018). Due to inconsistencies in the transcription of tone, Thai orthographic forms were sourced from WikiPron, but IPA forms were generated with Epitran.

To determine read and write scores, each lexicon was processed by OTEANN 7 times. Each time, 6,500 pairs were randomly selected as training data and 1,000 pairs as training data. The resulting read and write scores were then averaged to determine final read and write scores.

2.3 Automatic Speech Recognition

The ASR model used was openAI's Whisper model (Radford et al., 2022), specifically the small multilingual model. Lodagala (2024) was used to fine-tune this model for each language, and gather character error rates (CER) and word error rates (WER).

Common Voice data (Ardila et al., 2020) was used to train ASR models. Three ASR models were implemented for each language, the first was trained on 2.5 hours of speech data, the second on 5 hours, and the last on 10 hours[2]. Each model received 1 additional hour of data in the form of a dev set. Each dev set was restricted to at most 150 unique voices [3]. All models were evaluated on a 1-hour testing set.

3 Results

3.1 Orthographic Complexity

The OC of every language meeting the criteria presented in section 2.1 can be seen in Table 1, and the OC of completely processed languages can be seen in Figure 2.

The results from Marjou (2021) have been included in Table 1 for comparison purposes. The OC scores reported in Marjou (2021) are largely higher overall, e.i. the languages appear to have lower OC. This discrepancy likely reflects methodological differences between the current implementation and Marjou (2021). Marjou (2021) trained models on a multilingual data set, whereas the current implementation trained language-specific models on monolingual data sets. The lower OC scores reported in Marjou

[2]The data sets build on each other. All data present in the 2.5 hour training set was present in the 5 hour training set, and so on for the 10 hour training set.

[3]Dev sets of over 150 speakers resulted in poor ASR performance.

(2021) may possibly be explained by the learning of cross-linguistic patterns. For example, <t> maps to /t/ in en, pt, de, nl, es, it, and tr. In order to firmly draw this conclusion, OC scores would need to be calculated for all languages evaluated in Marjou (2021) and compared.

These results show that of the fully processed languages, Thai has the highest OC and Swahili, Georgian, and Turkish the lowest OC.

3.2 Automatic Speech Recognition

The word error rates (WER) and character error rates (CER) across hours of training data for each language can be seen in Table 2.

These results show that by 10 hours of training data, all languages reach relativity low CER rates, namely rates of under 15%. Farsi exhibited the highest CER at 14.35%, and Georgian exhibited the lowest CER at 6.17%.

4 ASR Performance, OC, and Training Hours

The individual evaluation of OC and ASR performance allows for the possible relation between these two factors to be determined. As ASR mirrors p2g conversion, OTEANN write scores (also a p2g task) were used to represent OC. Additionally, as the present focus is on orthography, CER scores were used to represent ASR performance. Finally, number of training hours acted as a reference point. OTEANN write scores, CERs, and training hours can be seen plotted together in Figure 3.

In general, all languages learn at a comparable rate and reach a similar CER at 10 hours of training data. Even the Farsi curve aligns with the general pattern at and after 5 hours of training data. This suggests that there is no obvious relation between OC and ASR performance, that is, all languages exhibit similar ASR performance regardless of their OC.

5 Discussion

The above results indicate that the hypothesis was not supported; there is seemingly no relation between OC and ASR performance. This conclusion can be understood when considering the data ASR models process.

While ASR mirrors a p2g task, in reality, ASR is more akin to a phone-to-grapheme conversion task. It is then not shocking that the examined languages performed similarly. All languages have allophonic variation. In an ASR task, where phonetic forms are being processed, this allophonic variation becomes more significant than the complexity of the phonological system. As OC is concerned with p2g mappings, not phone-to-grapheme mappings, OC bears no obvious relation to ASR performance.

Moving forward, there might be merit in considering if a relation exists between phonemic inventory size and ASR performance. Languages with a larger phonemic inventory potentially have more allophones, and therefore might pose more challenge to ASR models. The languages surveyed here have phonemic inventories between 30 (Thai, Farsi) and 48 (Welsh) phonemes (Moran and McCloy, 2019). While there does not seem to be a performance difference between phonemic inventories of this size, there may be performance differences for languages with much smaller (Hawaiian, n=13) or much larger (Hindi, n=74) phonemic inventories.

Another possible explanation for these results lies in the limited variability of OC among the fully processed languages. All fully processed languages have OCs of .4 or greater. As no languages under this value were fully processed, there could exist a threshold at or below .4 where ASR performance begins to change. For this reason, languages with OC scores of under .4, such as English, should be evaluated.

Moving now into methodological limitations, this project created ASR data sets by choosing Common Voice files at random. As a result, most data sets had large number-of-voice variability. For example, the 2.5-, 5-, and 10-hour English training sets contained 1046, 1854, and 3222 voices respectively. When evaluating the resulting models, CER and WER increased as training data was incremented. Models of German and Japanese, which also had high number-of-voice variability, demonstrated the same trend. It was thought that these results were not representative, and were thrown out. Examination of Welsh ASR models outlined the errors with this method.

Welsh has only 10 hours of training data available on Common Voice. Because of this, random file selection resulted in low number-of-voice variability; specifically 36 or 37 voices per training set. CER and WER decreased as training data in-

Language	Read Score	Write Score	OTEANN Read Score	OTEANN Write Score
en	.161	.203	.311	.361
th	.666	.438	-	-
pt	.473	.508	.824	.758
ru	.46	.523	.972	.431
de	.304	.557	.78	.69
ca	.285	.567	-	-
cy	.448	.584	-	-
nl	.567	.584	.557	.729
fa	.966	.75	-	-
pl	.865	.783	-	-
ta	.634	.864	-	-
es	.675	.909	.853	.669
it	.637	.913	.716	.945
sw	.937	.945	-	
ka	.949	.954	-	-
tr	.962	.966	.959	.955

Table 1: Orthographic Complexity Scores

note. Languages within shaded rows were not completely processed.

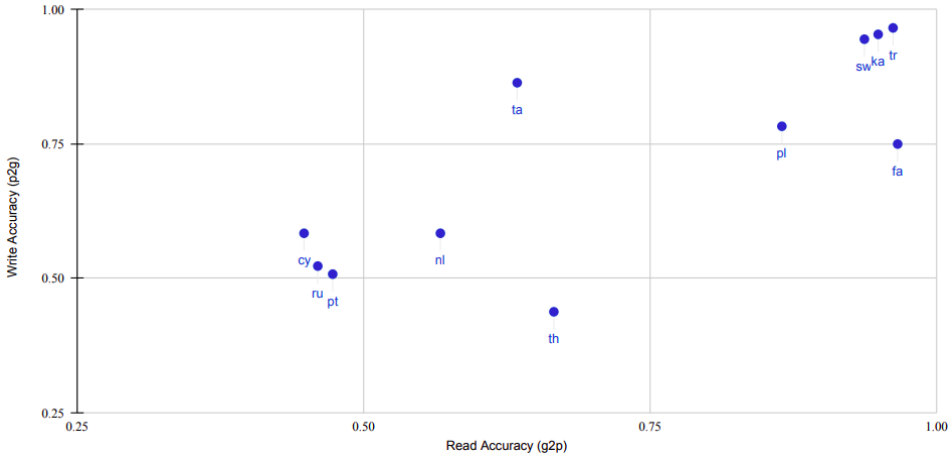

Figure 2: Orthographic Complexity Scores

note. 0 = high complexity (no predictability), 1 = low complexity (high predictability)

creased, as is expected. The decision was then made to focus on languages with less voices in training data, and to manually limit the dev data to a maximum of 150 voices. This method generated the results reported here. While the present conclusion can be drawn from these results, moving forward, composition of ASR data sets should be explicitly controlled, namely in terms of number-of-voices.

Number-of-voices in all ASR data sets should be standardized, both within and across languages, i.e. all data sets of the same type (train, test, dev) should contain the same number of speakers. While a limit of 150 voices seems to be a successful threshold for dev sets, the recommended number of voices in training and testing sets is unclear. Although, a smaller amount of voices in training sets might facilitate learning, and a larger amount of voices in testing sets might generate more representative results. Ultimately, this new methodology should be used to evaluate all languages listed in section 2.1.

Language	Training Hours	WER	CER	Language	Training Hours	WER	CER
Welsh	2.5 Hours	46.79	15.83	Georgian	2.5 Hours	52.5	11.61
	5 Hours	36.11	12.37		5 Hours	43.07	7.95
	10 Hours	28.47	9.48		10 Hours	33.03	6.17
Dutch	2.5 Hours	29.16	10.75	Turkish	2.5 Hours	54.65	18.31
	5 Hours	24.97	9.08		5 Hours	42.92	14
	10 Hours	24.01	9.44		10 Hours	40.5	13.36
Farsi	2.5 Hours	93.91	89.07	Tamil	2.5 Hours	22.63	12.22
	5 Hours	46.35	15.67		5 Hours	23.36	12.09
	10 Hours	42.37	14.35		10 Hours	17.34	8.91
Polish	2.5 Hours	37.01	17.57	Portuguese	2.5 Hours	38.88	15.05
	5 Hours	31.58	9.22		5 Hours	29.41	11.17
	10 Hours	25.15	7.32		10 Hours	26.02	10.36
Swahili	2.5 Hours	50.62	18.82	Russian	2.5 Hours	36.8	13.64
	5 Hours	40.97	15.06		5 Hours	29.2	9.68
	10 Hours	33.71	12.88		10 Hours	23.16	7.09
Thai	2.5 Hours	43.69	19.05				
	5 Hours	31.78	9.22				
	10 Hours	26.63	9.13				

Table 2: ASR Performance

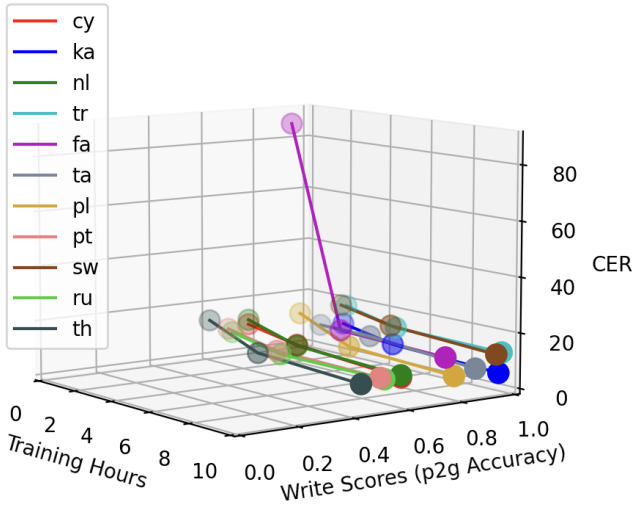

Figure 3: CER as Related to Number of Training Hours and p2g Accuracy

6 Conclusion

This project quantified the OC and ASR performance of 11 languages in order to determine the possible relation between these factors. Ultimately, no obvious relation was found. The fact that ASR models process phonetic, not phonemic, forms can explain the lack of obvious relation. Additional languages should be processed to confirm this finding, ideally with methodology that explicitly controls for number-of-voices in ASR data files.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and

Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Elisabeth Borleffs, Ben Maassen, Heikki Lyytinen, and Frans Zwarts. 2017. Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and writing, 30(8), 1617–1638*.

Phillip Carr. 2019. *English Phonetics and Phonology*. Wiley-Blackwell.

R Frost and L Katz. 1989. Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory cognition, 17(3), 302–310.*

Leonard Katz and Ram Frost. 1992. Chapter 4 the reading process is different for different orthographies: The orthographic depth hypothesis. In Ram Frost and Leonard Katz, editors, *Orthography, Phonology, Morphology, and Meaning*, volume 94 of *Advances in Psychology*, pages 67–84. North-Holland.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Leipzig Corpora Collection. 2017. Swahili Community Corpus. Leipzig Corpora Collection. Dataset.

Vasista Lodagala. 2024. Fine-tuning and evaluating Whisper models for Automatic Speech Recognition.

Xavier Marjou. 2021. OTEANN: Estimating the transparency of orthographies with an artificial neural network. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 1–9, Online. Association for Computational Linguistics.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. *arXiv preprint arXiv:2207.11782*.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague dependency style treebank for Tamil. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1888–1894, İstanbul, Turkey.

Mohammad Rasooli, Pegah Safari, Amirsaeid Moloodi, and Alireza Nourian. 2020. The persian dependency treebank made universal.

Xenia Schmalz, Eva Marinus, Max Coltheart, and Anne Castles. 2015. Getting to the bottom of orthographic depth. *Psychon Bull Rev 22, 1614–1629*.

Appendix

Language	Type	Hours	Number of Voices	Language	Type	Hours	Number of Voices
	Train	2.5	37		Train	2.5	32
	Train	5	37		Train	5	32
Welsh	Train	10	37	Dutch	Train	10	32
	Dev	1	109		Dev	1	146
	Test	1	109		Test	1	502
	Train	2.5	121		Train	2.5	22
	Train	5	151		Train	5	22
Thai	Train	10	172	Tamil	Train	10	22
	Dev	1	124		Dev	1	32
	Test	1	666		Test	1	242
	Train	2.5	77		Train	2.5	125
	Train	5	81		Train	5	134
Polish	Train	10	83	Farsi	Train	10	145
	Dev	1	127		Dev	1	131
	Test	1	585		Test	1	629
	Train	2.5	23		Train	2.5	28
	Train	5	23		Train	5	28
Turkish	Train	10	23	Georgian	Train	10	28
	Dev	1	117		Dev	1	73
	Test	1	529		Test	1	291
	Train	2.5	280		Train	2.5	306
	Train	5	292		Train	5	315
Russian	Train	10	296	Portuguese	Train	10	316
	Dev	1	103		Dev	1	71
	Test	1	532		Test	1	627
	Train	2.5	147				
	Train	5	147				
Swahili	Train	10	147				
	Dev	1	74				
	Test	1	349				

Table 3: Break Down of Voices

Parameter	Value
sampling rate	16000
num proc	2
train strategy	epoch
learning rate	1.25e-5
warmup	1000
train batchsize	16
eval batchsixe	8
num epochs	5

Table 4: ASR Hyperparameters