OTEANN: Estimating the Transparency of Orthographies with an Artificial Neural Network

Xavier Marjou
Lannion, Brittany, France
`xavier.marjou@gmail.com`

Abstract

To transcribe spoken language to written medium, most alphabets enable an unambiguous sound-to-letter rule. However, some writing systems have distanced themselves from this simple concept and little work exists in Natural Language Processing (NLP) on measuring such distance. In this study, we use an Artificial Neural Network (ANN) model to evaluate the transparency between written words and their pronunciation, hence its name Orthographic Transparency Estimation with an ANN (OTEANN). Based on datasets derived from Wikimedia dictionaries, we trained and tested this model to score the percentage of correct predictions in phoneme-to-grapheme and grapheme-to-phoneme translation tasks. The scores obtained on 17 orthographies were in line with the estimations of other studies. Interestingly, the model also provided insight into typical mistakes made by learners who only consider the phonemic rule in reading and writing.

1 Introduction

An alphabet is a standard set of letters that represent the basic significant sounds of the spoken language it is used to write. When a spelling system (also referred as *orthography*) systematically uses a one-to-one correspondence between its sounds and its letters, the encoding of a sound (also referred as *phoneme*) into a letter (also referred as *grapheme*) leads to a single possibility; similarly the decoding of a letter into a sound leads to a single possibility as well. Such orthography is thus *transparent* with regards to phonemes with the advantage of offering no ambiguity when writing or reading the letters of a word, as illustrated in Figure 1.

In real life, no existing orthography is fully transparent phonemically. One reason is that a word spoken alone is sometimes different from a word spoken in a sentence. An even more consequen-

tial reason is that some orthographies like English[1] and French[2] have incorporated deeper depth rules that have moved them away from a transparent orthography (Seymour et al., 2003); this has created ambiguities when trying to write or read phonemically, as illustrated in Figure 2.

Many studies have discussed the degree of transparency of orthographies (Borleffs et al., 2017). These studies are mainly motivated by the estimation of the ease of reading and writing when learning a new language (Defior et al., 2002). Finnish, Korean, Serbo-Croatian and Turkish orthographies are often referred as highly *transparent* (Aro, 2004) (Wang and Tsai, 2009), (Turvey et al., 1984), (Öney and Durgunoğlu, 1997), whereas English and French orthographies are referred as *opaque* (van den Bosch et al., 1994). However, little work exists in NLP about measuring the level of transparency of an orthography. One noticeable exception is the work of van den Bosch et al. (1994) who have created grapheme-to-phoneme scores and tested them on three orthographies (Dutch, English and French).

This study extends such work with a method called OTEANN, which models a word-based *phoneme-to-grapheme* task and a word-based *grapheme-to-phoneme* task using an ANN. For the sake of simplicity, the former task is called a *writing* task while the latter task is called a *reading* task. The goal is not to build a perfect spelling translator or a spell checker. Instead the goal is to build a translator which can indicate a degree of phonemic transparency and thus make it possible to rank orthographies according to this criterion.

Interestingly, recent years have seen tremendous progress regarding NLP with ANNs (Otter et al., 2018). Sutskever et al. (2014) proposed an ANN

---

[1] `https://en.wikipedia.org/wiki/English_orthography#Spelling_patterns`
[2] `https://fr.wiktionary.org/wiki/Annexe:Prononciation/français`
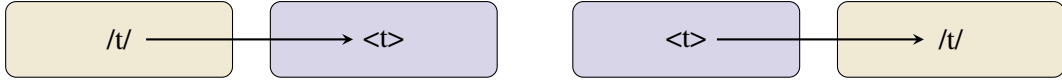
1

Figure 1: Example of unambiguous correspondence during writing and reading tasks in Esperanto.
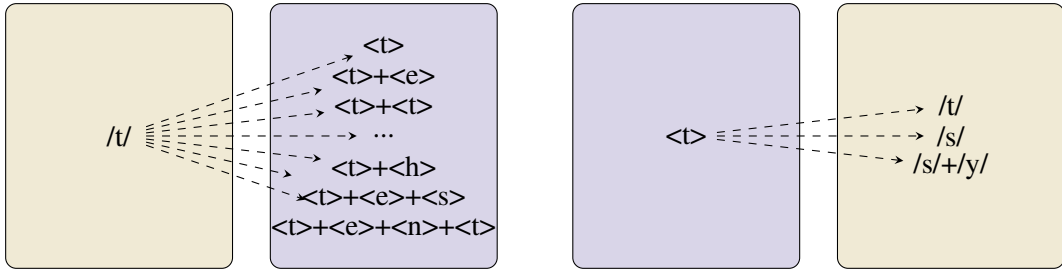

Figure 2: Example of ambiguous correspondence during writing and reading tasks in French. The /t/ phoneme can correspond to multiple graphemes, depending on the nature of the word and also depending on the nature of neighboring words in the sentence or even in a previous sentence. Similarly, the <t> grapheme can correspond to multiple phonemes.

called a Sequence-to-sequence (seq2seq) model that has proven to be very successful on language translation tasks. More recently, ANNs based on as *attention* (Bahdanau et al., 2014), (Vaswani et al., 2017) and *transformers* like Bidirectional Encoder Representations from Transformers (BERT), (Devlin et al., 2018) and Generative Pre-Training (GPT) (Radford, 2018) have again enhanced and outperformed seq2seqs. Considering writing a word and reading a word as two translations tasks allows re-using the transformers for our work. To this purpose, we used a minimalist GPT implementation (Karpathy, 2020) called *minGPT*. Notice that since we don't aim at building a perfect spelling translator, we do not have to translate a sequence of words into another sequence of words; our model only requires translating a spoken word into a spelled word (*writing task*) and a spelled word into a spoken word (*reading task*). In other words, our ANN operates at the character level within a sequence of characters of single words. The pronunciation and spelling of the word are both encoded as a sequence of UTF-8 characters; a pronounced word is encoded with the characters belonging to the set of phonemes of the target language, whereas a spelled word is encoded with the characters belonging to the alphabet of the target orthography. We directly re-used minGPT code with no modification. The only differences were the training data and the code for extracting the prediction at inference time.

We used OTEANN to test seventeen orthographies in order to evaluate their degree of phonemic transparency. Sixteen of them are the official orthographies of their respective language (Arabic, Breton, Chinese, Dutch, English, Esperanto, Finnish, French, German, Italian, Korean, Portuguese, Russian, Serbo-Croatian, Spanish, and Turkish) while the seventeenth is a phonemic orthography proposed for French.

A unique multi-orthography ANN model instance was trained to learn the writing and reading tasks on all languages at the same time. In other words, we used a single dataset containing samples of all studied orthographies. The multi-orthography ANN model was then tested for each orthography and each task with new samples, which allowed calculating an average percentage of correct translations. A score of 0% of correct translations represented a fully opaque orthography (no correlation between the input and the target), whereas a score close to 100% represented a fully transparent orthography (full correlation between the input and the target).

Our study first confirms that orthographies like Arabic, Finnish, Korean, Serbo-Croatian and Turkish are highly transparent whereas other ones like Chinese, French and English are highly opaque. For example, when solely based on a phoneme-grapheme correspondence, we estimated the chances of correctly writing a French word at 28%; similarly, when solely based on a grapheme-phoneme correspondence, we estimated the chances of correctly pronouncing an English word at 31%. For Dutch, English and French reading tasks, our obtained ranking is in line with the one of van den Bosch et al. (1994). One unexpected finding is that OTEANN also allows discovering

2

Orthography	Task	Input	Output
en	write	dʒɒb	job
en	read	job	dʒɒb

Table 1: Features of the multi-orthography dataset

certain mistakes performed by a new learner during writing and reading.

Remarkably, our method should apply to any orthography, provided a dataset is available.

2 Methodology

In order to evaluate a level of transparency of some orthographies two main steps were necessary: obtaining datasets and carrying out the training and testing experiments with the ANN.

2.1 Datasets

As displayed in Table 1, we needed a multi-orthography dataset with four features per sample: the orthography, the task (write or read), the input word (pronunciation or spelled word) and the output word (spelled word or pronunciation). A spelled word was represented by a sequence of graphemes whereas a pronunciation was represented by a sequence of phonemes. The characters representing phonemes are also called International Phonetic Alphabet (IPA) characters. Having a single dataset with multiple orthographies and tasks allows a single multi-orthography ANN model to learn to read and write all orthographies; otherwise, it would require one ANN model per orthography-task pair.

In order to build such dataset, we first generated one sub-dataset per orthography (e.g. one 'en' sub-dataset for English), each containing the pronunciation and the spelled word (e.g. 'dʒɒb' and 'job').

2.1.1 Baseline Orthographies

We first created baselines representing a fully transparent orthography and a fully opaque orthography.

Regarding a fully transparent orthography, we created a new artificial orthography called Entirely Transparent ('ent') orthography. We generated its samples by using the IPA pronunciation of real Esperanto words both as the pronunciation and as the spelled word, which resulted in a sub-dataset containing an 'ent' *bijective* orthography.

Regarding a fully opaque orthography, we also created a new artificial orthography called Entirely Opaque ('eno') orthography. We generated its samples by taking the IPA pronunciation of real Esperanto words mapping each of theirs phonemes to a random grapheme from a list of 25 graphemes, which resulted in a sub-dataset containing an 'eno' orthography with no correlation between the pronunciation and the spelled word.

2.1.2 Studied Orthographies

A sub-dataset was created for each of the following orthographies: Arabic ('ar'), Breton ('br'), German ('de'), English ('en'), Esperanto ('eo'), Spanish ('es'), Finnish ('fi'), French ('fr'), Italian ('it'), Korean ('ko'), Dutch ('nl'), Portuguese ('pt'), Russian ('ru'), Serbo-Croatian ('sh'), Turkish ('tr') and Chinese ('zh').

We incorporated the words from the corresponding Wiktionary[3] dump[4], with the exception of the following ones:

- Words containing space characters;

- Words containing more than 25 characters;

- Words containing capital letters (except for German words);

- Words containing non-standard characters with regard to the orthography's alphabet.

Two orthographies required additional processing:

- For German, proper nouns were discarded and the capital letter of common nouns was transformed into lower case;

- For Korean, the syllabic blocks words were converted in a series of two or three letters (one vowel and one or two consonants) pertaining to the Korean alphabet with *ko_pron*[5] Python library.

Regarding pronunciation, we directly extracted the IPA pronunciation when available in the associated Wiktionary dump, which was the case for 'br', 'de', 'en', 'es', 'fr', 'it', 'nl', 'pt' and 'sh'. The Esperanto ('eo') pronunciation came from the French Wiktionary. For the others ('ar', 'ko', 'ru', 'fi', 'tr'), we had to derive it from the spelled word with additional software. For Russian, the Russian Wiktionary dump did not contain the IPA. We thus used

---

[3] https://wiktionary.org
[4] https://dumps.wikimedia.org/
[5] https://pypi.org/project/ko-pron

wikt2pron ru_pron module[6] to obtain a pronunciation similar to the one displayed in the Russian Wiktionary web pages. For Chinese, we only selected Mandarin words in simplified Chinese and limited to one or two symbols (a.k.a. Hanzis); we then obtained their pronunciation from the *CEDICT*[7] dataset.

Extracting the phonemic pronunciation from Wiktionary may raise concerns given than IPA symbols can be used both for phonetic and phonemic notations and that there is no unified consistency between the different dictionaries. When processing the IPA strings, we nonetheless took care of preserving the highest surface pronunciation as possible: most pitches were removed since they represent no useful hint during the writing task (i.e. no consequence on the spelled word) and especially since they are generally impossible to predict when translating the spelled word into a pronunciation during the reading task. Nevertheless the /ː/ pitch was noticed as indispensable for some orthographies, for instance for predicting double vowels in the spelling of Finnish words or the *alif* letter in Arabic. Regarding the /ˈ/ pitch, it can slightly influence Spanish translation scores: it can lead to a better writing score as it can be a hint for predicting accented letters, but it can also lead to a lower reading score.

Another interesting orthography was a proposal of an alternative orthography for French called French Ortofasil ('fro')[8], which seeks to be phonemically transparent. Although not fully bijective (e.g. both /o/ and /ɔ/ map to <o> letter), it indeed seems highly transparent. We therefore used it to generate a sub-dataset for the 'fro' orthography.

It is debatable whether Chinese should be included in this study given the term *alphabet* is usually reserved for largely phonographic systems that have a small number of elements. We decided to include it because our ANN model allowed for *alphabets* with thousands of graphemes.

Table 2 summarizes the sub-datasets obtained.

2.1.3 Training and test datasets

$11,000$ samples were randomly selected in each of the 17 sub-datasets. Each sample from a sub-dataset produced two samples in the multi-orthography dataset: one sample for *write* task and one sample for the *read* task, as illustrated in Table 1. This multi-orthography dataset was subsequently divided into a training dataset ($10,000$ * 17 * 2 samples) and a test dataset ($1,000$ * 17 * 2 samples).

2.1.4 ANN architecture

We used minGTP (Karpathy, 2020) which runs on PyTorch[9]. Regarding the hyper-parameters, we configured a block size of 63 characters, 4 layers, 4 heads and 336 embedding tokens, which resulted in an ANN of $9,589,536$ trainable parameters and an episode training time of 2 hours and 10 minutes on a 4 GPU node. No effort was spent to shrink or prune the ANN, so its size could still be optimized. The data and code are available on Github [10].

2.1.5 Performance metric

We used a simple score in order to assess the performance of the ANN prediction during the testing step. When all the predicted characters were equal to those of the true target, a prediction was considered successful, hence allowing to score the percentage of successful predictions performed for each orthography-task pair.

2.1.6 Training and testing

We specified an episode as:

- **Generating the training and test datasets.** At the end of this step, each character present in these datasets was provisioned in the inventory of the ANN instance.

- **Training the ANN model**. The full training dataset was processed to be used as text blocks containing the concatenation of the four features (*orthography*, *task*, *input* and *output*) separated by a comma. Therefore, a single instance of the model was used to learn to write and read all 17 orthographies in one training.

- **Testing the ANN model for each orthography-task pair**. For each orthography-task pair, $1,000$ new samples were tested. Each sample was fed into the model with the concatenation of the three first features (*orthography*, *task* and *input*) separated by a comma. The model had to

Orthography	Samples	Phonemes	Graphemes	Nb. of Phonemes	Nb of Graphemes
ar	12,057	32	47	8.0 ± 2.0	8.9 ± 2.3
br	17,343	45	29	6.6 ± 1.9	7.5 ± 2.2
de	529,740	41	30	10.2 ± 3.1	11.5 ± 3.4
en	42,206	42	29	7.3 ± 2.7	7.6 ± 2.6
eo	26,845	25	28	8.8 ± 2.6	8.6 ± 2.5
es	40,824	34	33	8.1 ± 2.7	8.7 ± 2.6
fi	105,352	28	27	10.4 ± 3.5	10.4 ± 3.5
fr	1,214,248	35	41	9.0 ± 2.7	11.2 ± 2.9
fro	1,214,262	35	32	9.0 ± 2.7	8.6 ± 2.6
it	26,798	34	32	9.1 ± 2.8	9.1 ± 2.6
ko	64,669	41	67	10.6 ± 4.0	8.3 ± 3.0
nl	13,340	45	28	7.8 ± 3.1	8.6 ± 3.4
pt	12,190	37	38	7.7 ± 2.3	7.9 ± 2.3
ru	304,514	30	33	10.5 ± 3.1	10.7 ± 3.1
sh	98,575	27	27	9.1 ± 2.8	8.9 ± 2.7
tr	117,841	36	31	10.3 ± 3.7	10.1 ± 3.6
zh	27,688	32	4813	9.9 ± 2.2	1.8 ± 0.3
eno	26,845	25	25	8.8 ± 2.6	8.8 ± 2.6
ent	26,845	25	25	8.8 ± 2.6	8.8 ± 2.6

Table 2: Summary of the sub-datasets. For each sub-dataset, a line indicates the number of samples available, the number of different phoneme UTF-8 characters, the number of different grapheme UTF-8 characters, the mean number of phonemes in words, and the mean number of graphemes in words.

predict a value equal to the *output* feature, which was the target to be found.

We performed 11 episodes to measure the mean and standard deviation of each orthography-task pair and thus assess the consistency of our results.

Future work may use more test samples to gain a statistical insight on the different types of errors depending on the orthography at hand.

3 Results

First, regarding the results of the two baseline orthographies, the 'eno' opaque orthography obtained a score of 0% in both writing and reading, which was in line with the expectations given that there was no correlation between its phonemes and its graphemes; on the other hand, the 'ent' transparent orthography scored above 99.6% on the writing and reading tasks, which indicated a high level of correlation between its phonemes and its graphemes. We thus considered our ANN model satisfactory for our objective of comparing the performance of different orthographies.

Figure 3 and Table 3 present our main results. They are significantly different between writing and reading since these tasks are generally not symmetrical. Two features are likely to influence the

symmetry, and therefore the efficiency of each task. As recalled by Figure 2, the most important feature would undoubtedly be the number of possible phoneme-to-grapheme and grapheme-to-phoneme ambiguities per tested orthography. Unfortunately we did not possess such data. Another impacting feature may be the number of possible values (graphemes or phonemes) for a given target character. The higher the number of values, the harder the prediction should be for the ANN. Future work should investigate the relative importance of these features on the OTEANN performances.

Comparing OTEANN's reading results with those of van den Bosch et al. (1994), OTEANN first seems to naturally assimilate the grapheme complexity (e.g. for French, it successfully learnt that "cadeau" should be pronounced /kado/). Regarding grapheme-to-phoneme complexity (*G-P complexity*), they ranked English (*G-P complexity*=90%) more complex than Dutch (*G-P complexity*=25%) which, in turn, was more complex than French (*G-P complexity*=15%). OTEANN results preserved the same ranking with transparency scores of 31%, 57% and 79s% for English, Dutch and French. Admittedly, OTEANN's scores were different in terms of scale but OTEANN had to deal

Orthography	Write	Read
ent	99.6 ± 0.3	99.8 ± 0.1
eno	0.0 ± 0.0	0.0 ± 0.0
ar	84.3 ± 0.8	99.4 ± 0.3
br	80.6 ± 0.6	77.2 ± 1.6
de	69.1 ± 1.0	78.0 ± 1.5
en	36.1 ± 1.5	31.1 ± 1.3
eo	99.3 ± 0.2	99.7 ± 0.1
es	66.9 ± 2.0	85.3 ± 1.3
fi	97.7 ± 0.3	92.3 ± 0.8
fr	28.0 ± 1.4	79.6 ± 1.7
fro	99.0 ± 0.3	89.7 ± 1.1
it	94.5 ± 0.8	71.6 ± 0.9
ko	81.9 ± 1.0	97.5 ± 0.5
nl	72.9 ± 1.7	55.7 ± 2.2
pt	75.8 ± 1.0	82.4 ± 0.9
ru	41.3 ± 1.6	97.2 ± 0.5
sh	99.2 ± 0.3	99.3 ± 0.3
tr	95.4 ± 0.7	95.9 ± 0.6
zh	19.9 ± 1.4	78.7 ± 0.9

Table 3: Phonemic transparency scores.
(OTEANN trained with 10, 000 samples)

with more orthographies as well as with the writing task.

Figure 3 also allows categorizing the studied orthographies with respect to their degree of transparency:

- **Esperanto**: With scores above 99.3%, Esperanto orthography is nearly as transparent as the 'ent' baseline. The most common error occurred on a doubled letter in the input, which was incorrectly translated to a single letter.

- **Arabic, Finnish, Korean, Serbo-Croatian and Turkish**: Their scores above 80% both in writing and reading confirmed that their orthography is highly transparent as indicated in (Aro, 2004), (Wang and Tsai, 2009) and (Öney and Durgunoğlu, 1997). The Arabic score is high on in the read direction, which is likely due to the use of diacritics in the dataset; without them, the score would undoubtedly be lower. Regarding Korean, its orthography became a little less transparent during the twentieth century; its high scores suggest that further work should check the dataset and evaluate new scores.
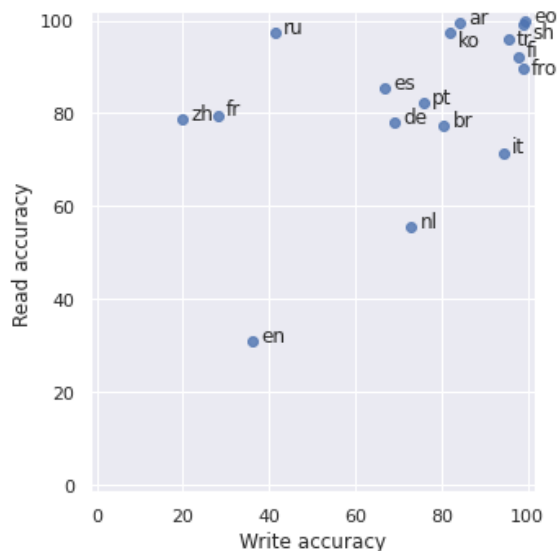

Figure 3: Scatterplot of the mean scores.
(OTEANN trained with 10, 000 samples)

- **Breton, German, Italian, Portuguese and Spanish**: With all their scores above 65% their orthography was also measured as fairly transparent. For Spanish, the detailed results showed that the most common failure during writing occurs with accents: the ANN had great difficulty predicting whether a vowel should contain an accent or not. For Italian, typical errors observed in the results were the prediction of /ɛ/ instead of a /e/ and /ɔ/ instead of a /o/, which were harder to discriminate. Future work may revise the scoring formula to reduce the cost of some of these errors in the performance calculation.

- **Dutch**: The Dutch reading score (56%) is low but might be slightly enhanced given a possible lack of consistency regarding the phonemes used in the Dutch sub-dataset.

- **Russian**: The Russian writing score (41%) may seem low. However, Russian has strong stress-related vowel reduction, which makes it hard to know how to write a word without knowing the morphemes involved. Nevertheless, future work should either study their sub-dataset more in depth or use a different data source like wikipron[11] to possibly improve its scores.

- **Chinese**: The results indicated a low writing score (20%), which is not surprising given

6

than some phonemes can have multiple corresponding graphemes and that there are thousands of graphemes (Hanzis) to be learnt. However, it turns out that its reading score is much higher (79%).

- **French**: With a low writing score (28%), the results showed that the chances of correctly writing a French word on the sole basis of its pronunciation were rare, as anticipated given the high number of phoneme-to-grapheme possibilities. Without being able to access a broader context than the word itself, the ANN was not able to reliably predict how to write a French word. With a much higher reading score (80%), the ANN obtained good reading results. As a comparison, for the same language, the alternative 'fro' orthography obtained excellent writing score (99%) and reading score (90%). Recall that the difference between its two scores is due to the fact that the 'fro' orthography is not bijective. For instance, in the reading direction, the <o> letter can be translated into /o/ or /ɔ/).

- **English**: With a low writing score (36%) and a low reading score (31%), the results showed that English orthography is also highly opaque, which is consistent with most studies. As a reminder, a phonemic reading of an English word often does not work because of its high number of grapheme-to-phoneme possibilities. For instance the grapheme <u> can either correspond to /ʌ/ (as in "hug"), to /juː/ (as in "huge"), to /ɜːr/ (as in "cur") or /jʊəː/ as in "cure". As for Russian, additional work should be dedicated to check the English sub-dataset and possibly enhance it if necessary, which could improve 'en' scores by a few percent.

Observing the detailed result of each prediction also made it possible to study the phonemic correspondences learned or not learned by the OTEANN model.

- For task-orthographies with a high transparency score, the model successfully predicted most pronunciations or spellings even when the correspondences involved more than one letter. For instance, OTEANN predicted that the Italian word "cerchia" should be pronounced /tʃerkja/, hence showing that the model had successfully learned that <c>, when followed by <e>, should be pronounced as /tʃ/ and also that <c>, when followed by <h>, should be pronounced as /k/.

- For task-orthographies with a low transparency score, the model generally failed on letters involved in ambiguous correspondences (recall Figure 2). For instance, it incorrectly predicted that the pronunciation of the English word "level" was "livəl" instead of "lɛvəl", which might be a bad generalization from words like "lever" learned at training time. OTEANN also incorrectly predicted that the spelling of the French word /ale/ was "allez" when the expected target was "aller" (another French homophone); this type of error is inevitable since the OTEANN model intentionally use single word input samples and therefore cannot rely on neighboring words as additional context to discriminate between homophones with different spelling.

- Surprisingly, the model also predicted spellings that do not exist but who could have existed, in the same vein as *ThisWordDoesNotExist.com*[12]. For instance, OTEANN predicted that the spelling of the French word /swaʁe/" was "soirer", which does not exist but looks like a French infinitive verb that would mean "to celebrate at a party".

In addition, the results in Table 3 also showed that the ANN has less than a 30% chance of correctly writing a word in French or Chinese after training on 10000 samples while Figure 9 shows that the same ANN has more than a 85% chance of correctly writing a word in Finnish, Italian, Serbo-Croatian or Turkish after training only on 1000 samples. Such a discrepancy highlights the enormous additional cost in terms of time and energy for learning a non-transparent orthography.

4 Discussion and Conclusion

Among the tested orthographies, some shared the grapheme inventory. Given that they are all trained together, there might be an impact on performance. Although some of our preliminary experiments with a single ANN instance per orthography did not seem to lead to significant differences, it could be interesting to formally compare both approaches.

---

[12] https://www.thisworddoesnotexist.com

The accuracy metric we used is all or nothing. Additional work could also study alternative accuracy metrics and compare their results on the different orthographies.

Although Wiktionary data may be inconsistent in quality and therefore positively or negatively impact the measured metric, the results obtained for Dutch, English and French orthographies reasonably extended those of van den Bosch et al. (1994) while the other results reflected the perception of several other studies. Consequently, our OTEANN model showed that an ANN can convincingly estimate a level of phonemic transparency for multiple orthographies both for the phoneme-to-grapheme and grapheme-to-phoneme directions.

This method should be easily applicable to other orthographies beyond those tested in this study. However, since the superfluous IPA symbols slightly influence the score results, future work should closely examine and discuss the phonemes to use depending on the orthography to be tested.

As OTEANN also points out some possible grapheme or phoneme errors when writing or reading phonemically, it could also be used to detect possible errors in the dictionaries of transparent orthographies; it could also be used to evaluate proposals for improving opaque orthographies.

Finally, it would be beneficial to investigate if our ANN and its artificial neural units somehow imitate the way a beginner learns to write and read a language. If so, it might suggest that a transparent orthography would be easier and faster to learn than an opaque orthography.

References

Mikko Aro. 2004. *Learning to read: The effect of orthography*. 237. Jyväskylän yliopisto.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Elisabeth Borleffs, Ben AM 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.

Sylvia Defior, Francisco Martos, and Luz Cary. 2002. Differences in reading acquisition development in two shallow orthographies: Portuguese and spanish. *Applied Psycholinguistics*, 23(1):135–148.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Andrej Karpathy. 2020. mingtp. Available at `https://github.com/karpathy/minGPT`, MIT licence.

Banu Öney and Aydin Yücesan Durgunoğlu. 1997. Beginning to read in turkish: A phonologically transparent orthography. *Applied psycholinguistics*, 18(1):1–15.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2018. A survey of the usages of deep learning in natural language processing. *arXiv preprint arXiv:1807.10854*.

A. Radford. 2018. Improving language understanding by generative pre-training.

Philip HK Seymour, Mikko Aro, Jane M Erskine, and Collaboration with COST Action A8 Network. 2003. Foundation literacy acquisition in european orthographies. *British Journal of psychology*, 94(2):143–174.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

MT Turvey, Laurie B Feldman, and G Lukatela++. 1984. The serbo-croatian orthography constrains the reader to a phonologically analytic strategy. *Status Report on Speech Research: A Report on the*, page 17.

Antal van den Bosch, Alain Content, Walter Daelemans, and Beatrice de Gelder. 1994. Analysing orthographic depth of different languages using data-oriented algorithms.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yu-Chun Wang and Richard Tzong-Han Tsai. 2009. Rule-based korean grapheme to phoneme conversion using sound patterns. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 843–850.

A Additional Experiments and Results

In addition to testing our OTEANN model trained on $10,000$ samples, we also tested the same OTEANN model but trained with fewer samples ($1,000$, $2,000$, $3,000$, and $5,000$), each time following the methodology described in section 2. We then aggregated the results to summarize them in Figure 9, which shows the learning curve of the studied orthographies as a function of the number of training samples.
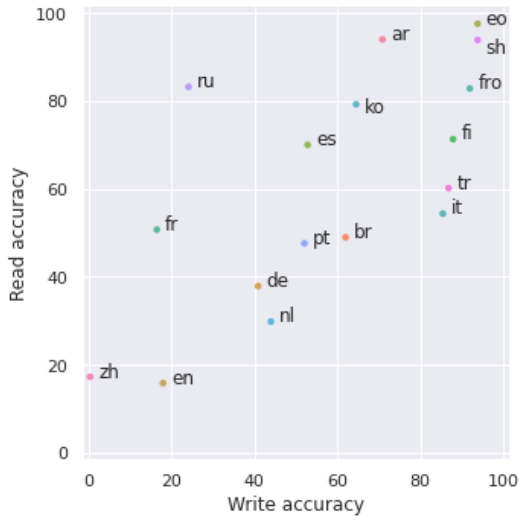
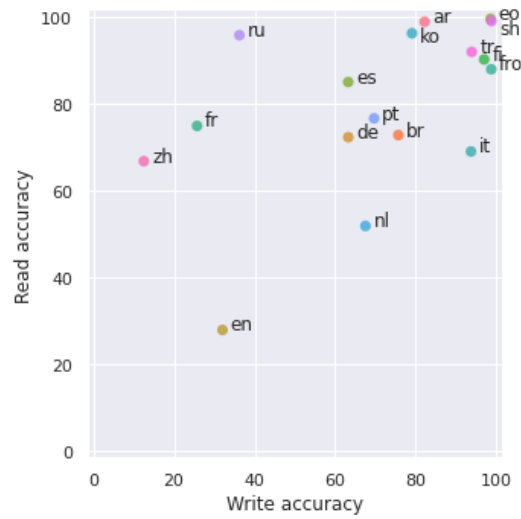Figure 4: Scores with $1,000$ training samples.

Figure 7: Scores with $5,000$ training samples.

Figure 5: Scores with $2,000$ training samples.
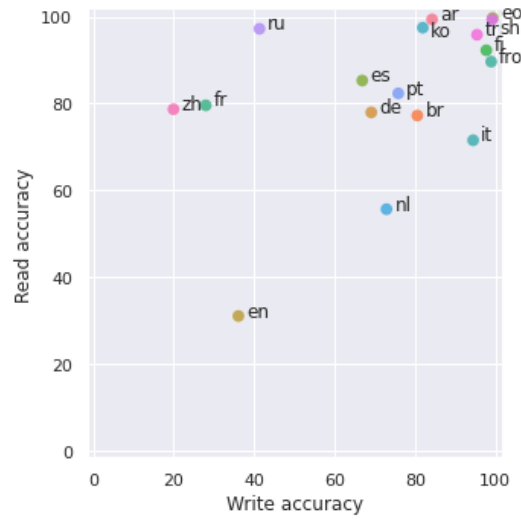

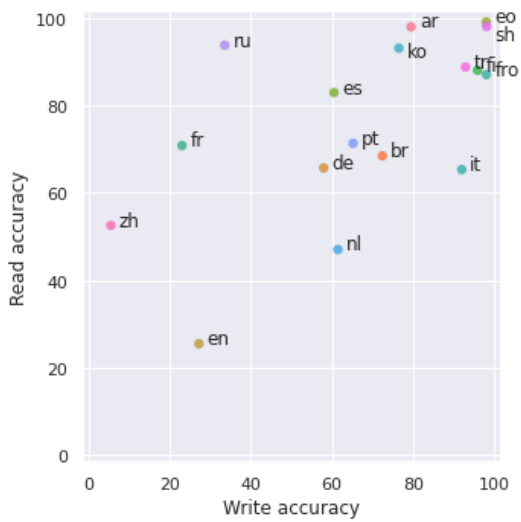Figure 8: Scores with $10,000$ training samples.
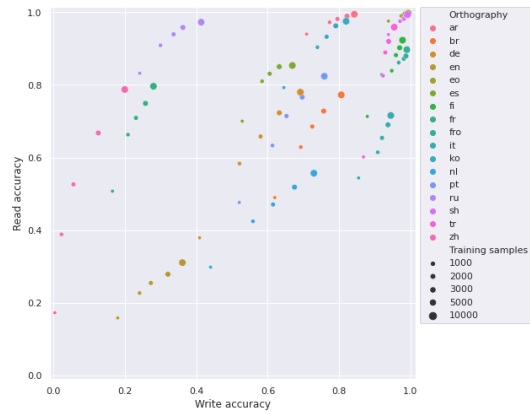

Figure 6: Scores with $3,000$ training samples.

Figure 9: Scores according to the number of training samples