Mixed Orthographic/Phonemic Language Modeling: Beyond Orthographically Restricted Transformers (BORT)

Robert C. Gale[†] galer@ohsu.edu
Alexandra C. Salem[†] salem@ohsu.edu
Gerasimos Fergadiotis[‡] gf3@pdx.edu
Steven Bedrick[†] bedrick@ohsu.edu

[†]Oregon Health & Science University
Portland, Oregon, USA

[‡]Portland State University
Portland, Oregon, USA

Abstract

Speech language pathologists rely on information spanning the layers of language, often drawing from multiple layers (e.g. phonology & semantics) at once. Recent innovations in large language models (LLMs) have been shown to build powerful representations for many complex language structures, especially syntax and semantics, unlocking the potential of large datasets through self-supervised learning techniques. However, these datasets are overwhelmingly orthographic, favoring writing systems like the English alphabet, a natural but phonetically imprecise choice. Meanwhile, LLM support for the international phonetic alphabet (IPA) ranges from poor to absent. Further, LLMs encode text at a word- or near-word level, and pre-training tasks have little to gain from phonetic/phonemic representations. In this paper, we introduce BORT, an LLM for mixed orthography/IPA meant to overcome these limitations. To this end, we extend the pre-training of an existing LLM with our own self-supervised pronunciation tasks. We then fine-tune for a clinical task that requires simultaneous phonological and semantic analysis. For an "easy" and "hard" version of these tasks, we show that fine-tuning from our models is more accurate by a relative 24% and 29%, and improves on character error rates by a relative 75% and 31%, respectively, than those starting from the original model.

1 Introduction

Recently, large language models (LLMs) have shown notable success in capturing information across several linguistic layers, developing rich representations of syntactic and semantic structures within their hidden layers (Rogers et al., 2020). This is accomplished through the use of self-supervised techniques, in which LLMs are pre-trained on large corpora to perform generic, contrived tasks. With a well-designed task, the model can make the most of vast quantities of unlabeled text, gleaning the structural patterns of the language(s). For example, in masked language modeling (MLM), a model is trained to restore partially obscured text to its original form. Following this relatively generic pre-training task, the resulting model can then be used as a starting point, and its weights fine-tuned (or its architecture augmented with additional output layers) in a task-specific manner. Crucially, the task-specific training can be accomplished using orders of magnitude less data than is required for the original pre-training step.

This approach is well-suited to many linguistic tasks, particularly those that rely on syntax and semantics. However, tasks that also depend on explicit representation of *phonology* are under-served by the current paradigm. Examples of such tasks include: analysis of code-switched language; processing ambiguous and noisy output from automated speech recognition systems; handling of names and neologisms; and analysis of clinical language samples in the context of communication disorders.

There are two underlying reasons for this issue, the first of which involves input representation. As with any computational model, LLMs require language to be encoded into a numerical form, and a model's choice of encoding technique has a profound impact on its functionality. Today's LLMs typically rely on sub-word representations such as WordPieces (Schuster and Nakajima, 2012) or Byte-Pair Encoding (BPE) (Gage, 1994; Sennrich et al., 2016), which strike a balance between vocabulary size and semantic precision while allowing unrestricted input and avoiding the limitations arising from a fixed word vocabulary. Such tokenization schemes are generally optimized for representing textual input for word-level processing, based on the distributional properties of a training corpus, and as such in practice the most prevalent segments in a tokenized input sentence represent entire words or large word fragments, with sub-word token fragment boundaries only incidentally co-occurring with morphological boundaries.

While it has long been known that neural language models are able to capture implicit information about phonology from orthography (Elman, 1990; Prince and Smolensky, 1997), the extent to which this occurs will depend on the degree to which the model's unit of representation maps to the writing system in question's representation of phonology. Furthermore, the self-supervised pre-training techniques at the foundation of LLM tend to work (roughly) at a word-level or word-fragment scope; a task like MLM has little to gain

Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 212–225

from learning the sound relationships between words, so we have no reason to expect these models to adapt to phonetic tasks as well as they do semantic ones.[1]

The second underlying reason for LLMs' phonological naïveté is data-related. Phonology is typically expressed in written form using a writing system such as the International Phonetic Alphabet (IPA) or ARPAbet. One advantage of the tokenization schemes used by LLMs is that they are, at a technical level, able to represent arbitrary character input sequences: sequences of previously-unseen characters (e.g., a sequence of IPA glyphs representing a phonemically-transcribed utterance) simply resolve to a sequence of single-character tokens (rather than lumping together into sub-word units as do more commonly-seen sequences of characters). More recently, new techniques have emerged for training "token-free" LLMs that operate at the character level (e.g. CANINE (Clark et al., 2022)) yet retain many of the semantic benefits of word-level models; however, their pre-training remains exclusively orthographic in nature, and thus will only learn phonology indirectly (and in a way that is mediated by the specifics of the writing system in question). Furthermore, and most importantly, the data used to pre-train an LLM incidentally contains little IPA content if any at all. As such, regardless of the modeling strategy used, the resulting embedding of these tokens is likely to be uninformative from the a phonological perspective.

Previous attempts to augment LLMs with phonemic information have focused on speech-centered applications, and have emphasized phonology at the expense of orthography (Jia et al., 2021; Sundararaman et al., 2021; Li et al., 2023; Zhang et al., 2022). For many applications however, particularly including clinical applications in speech-language pathology, *both* are crucial, as expressive and receptive language depend on both phonology and semantics. In this paper, we introduce BORT, an LLM that accepts a mixture of English pronunciations in IPA and English orthography, and demonstrate its use on a task motivated by a real-world clinical problem: analysis of speech errors made by individuals with aphasia following a stroke.

In §3, we create the BORT models by extending the pre-training of an existing LLM, BART (Lewis et al., 2020). Our self-supervised task focuses on a novel IPA-to-orthography translation task: given a document, we transform some words into IPA, then train the model to restore the orthography. Hypothesizing that we could bolster the pre-trained models with two additional transforms, we experiment with configurations that include spelling and noise transforms. In §4, we evaluate the utility of BORT by fine-tuning to two clinically-motivated tasks: a) an "easy" task, another mixed IPA/orthography to orthography translation task, but in the context of aphasic speech; and b) a "hard" task, in which the model must predict the intended word for several types of word errors, including phonologically and semantically related errors. We make our pre-trained models available for anyone to download and fine-tune.[2]

2 Background

2.1 Speech Language Pathology

Speech language pathologists (SLPs) work to diagnose and treat speech and language disorders. Language disorders typically include breakdowns in the linguistic system that supports the abilities to activate the semantic representation of a concept; retrieve its lexical/syntactical representation; and, encode its phonological or orthographic form. Speech disorders include deficits that may stem from underdeveloped or faulty perceptual representations of speech sounds; difficulties specifying a motor plan for the articulatory gestures required for producing a word; and/or executing the motor plan. Such disorders often co-exist, and may be developmental (affecting primarily pediatric populations, as in the case of Specific Language Impairment), or acquired and seen primarily in adult populations (e.g., dementia, aphasia, dysarthria). In addition, they might affect different modalities including spoken (e.g., anomia) or written output (e.g., agraphia).

The use case described in the present work focuses on aphasia, a disorder in which an individual has an impairment to one or both of their expressive or receptive language abilities. In expressive language, this may take the form of difficulties in word retrieval or production; these difficulties may involve the inability to produce a word, the production of an unintended word, or the mispronunciation of a produced word. Aphasia typically follows an injury to be brain (such as a stroke or a traumatic brain injury), though it may also be a sign of certain neurodegenerative conditions.

To arrive at a diagnosis, clinical professionals typically elicit productions from patients, and then, based on the relationship between the intended target and the realized unexpected or atypical production, they draw inferences regarding the nature of the cognitive-linguistic or motoric deficits of the patient. Inherent in this diagnostic process is identifying what was the intended word of a speaker. That requires a clinician to combine multiple sources of information including semantic, phonemic, and/or orthographic information.

Consider for example the response /bəgænə/ when

[1] Though we note that Itzhak and Levy (2022) have demonstrated that LLMs working at the word and sub-word level do implicitly learn a certain amount about the character-level contents of their tokens.

[2] https://github.com/rcgale/bort

a stroke patient is asked to name the picture of an apple. A clinician, naturally, will recognize the phonological similarity of the production to the candidate intended word /bənænə/ ("banana") and given the semantic similarity of "banana" and "apple," the clinician will arrive at the conclusion that most likely the speaker's intended target was "banana." This step is critical in the diagnostic process across clinical populations and disorders. Therefore, the development of a robust computational tool that can combine multiple sources of information to predict a speaker's intended words during a paraphasic speech event is of great clinical significance.

There exist several settings in which tools such these may find applications. There are a variety of highly accurate and informative assessment techniques that are regularly used in research settings but rarely used in clinical practice due to the large amount of effort they require for delivery and scoring (Edmonds and Kiran, 2006; Abel et al., 2007; Kendall et al., 2013; Minkina et al., 2015; Walker and Hickok, 2016); automation has the potential to streamline this process greatly, thereby enabling their clinical use. Additionally, automation of this sort would be a key part in many telemedicine and remote assessment scenarios, which is an area of great clinical interest (Van De Sandt-Koenderman, 2004; Kiran et al., 2014) as there exist major challenges around access to care for many individuals in need of speech and language services (Hou et al., 2023).

Use cases such as these currently are limited by two categories of technical barrier. The first is the need for robust automated speech recognition algorithms able to accurately process the disordered speech characteristic of individuals with speech and language disabilities, and produce detailed phonemic transcriptions; this is needed given the impracticality of detailed manual transcription in a fast-paced clinical setting. The second category is a lack of specialized algorithms designed to process the resulting data to identify features of clinical interest, for example in speech error classification (Casilio et al., 2023). Both types of technology are necessary, and neither on their own are sufficient, to leverage NLP in this clinical domain. The present work, by design, only addresses the second of these categories; however, it is important to note that the first is an area of very active research (Fraser et al., 2013; Le et al., 2017; Jacks et al., 2019; Perez et al., 2020; Torre et al., 2021; Gale et al., 2022), with major strides being made in recent years.

A second setting of use for automation in the analysis of language produced by people with aphasia (PWA) is that of aphasiological research. Standard research practice typically results in the creation of recorded sessions with participants (for example, in a discourse elicitation task), which are then transcribed to a very high degree of accuracy by specially-trained research staff, for use in analysis. Often, this transcription is done at a mixture of orthographic and phonemic levels, with particular phonemic attention paid to clinically-relevant phenomena such as neologisms (non-word productions), mis-pronunciations, etc. There exist large databases of such transcripts, for example TalkBank and its many sub-projects (see https://talkbank.org; MacWhinney, 2000), and automated analysis of these datasets is extremely valuable from a scientific perspective.

Notably, in this scenario, one need not assume the existence of an ASR system robust to disordered speech in order for automation to be useful, as the data are transcribed as part of their collection and data management process. However, the specifics of this transcription process tend to be very closely linked to the scientific needs of the research team conducting the study, and while the amounts of data generated tend to be far more than humans can conveniently analyze by hand, they tend to be relatively small in comparison to the datasets commonly used in natural language processing. As such, techniques such as transfer learning have become crucial tools in this space.

2.2 Automating Clinical Language Evaluation

There exists a long history of use of NLP techniques in clinical language evaluation, across a wide variety of disorders including Alzheimer's disease (Petti et al., 2020), Autism Spectrum Disorder (Virnes et al., 2015; MacFarlane et al., 2023), and various forms of aphasia (Fraser et al., 2014; Azevedo et al., 2023). From a computational perspective, this typically takes the form of a pipeline accepting language samples of some sort as input, and producing as output some sort of relevant analysis, such as a score on a validated assessment instrument. The language samples used may consist of spontaneous speech, a patient's responses to a structured interaction of some kind, or a mixture of the two, and may feature continuous speech, or single-word productions. The input may be actual audio recordings, or transcriptions thereof.

Recent work has taken advantage of LLMs for automating clinical language evaluation tasks. Balagopalan et al. (2020) fine-tuned BERT to detect Alzheimer's disease from transcribed spontaneous speech, and found that BERT performed better than a standard model based on hand-crafted features. Liu et al. (2022) evaluated transformer models for use in identifying relevant pragmatic features of transcribed speech from adults with Autism Spectrum Disorder; their analysis identified both advantages and limitations of an LLM-based approach over previous methods. Gale et al. (2021) described a system that scored tests for Specific Language Impairment in children, finding that the DistilBERT architecture was adaptable to clinical language evaluation spanning several linguistic layers. Salem

et al. (2022) fine-tuned DistilBERT for the automatic determination of semantic similarity in an aphasia test with accuracy 95.3%, improving over earlier methods that relied on word2vec (Fergadiotis et al., 2016).

LLMs demonstrate improved performance at many of these tasks, and bring two additional benefits over earlier methods: they are much more flexible with regard to their input representation, and they are well-suited for transfer learning via fine-tuning. Both attributes are crucial for work in clinical language evaluation, given the heterogeneity and limited size of the datasets used in this space. However, we note that technical approaches in this space tend to "live" in either the orthographic or phonemic space; this distinction is logical from a technical standpoint, given the nature and history of language technologies, but quite contrary to the actual clinical manifestation of speech and language disorders (and, of course, the way in which clinicians make use of language sampleswhere). From this perspective, we see BORT bridging the gap between audio recordings of spoken language tests—transcribed manually or by an automatic speech recognizer—and downstream language evaluation tasks.

2.3 Considering alternative methods

Conceptually, BORT enhances BART with phoneme-to-grapheme functionality. Existing English grapheme-to-phoneme (G2P) systems are highly accurate, with recent transformer models achieving a 5.23% character error rate and a 22.1% word error rate on CMUDict (Yolchuyeva et al., 2019). These systems are trained and evaluated on word-length samples, so integration with a contextual language model would require novel architectural adapters, lest translation errors propagate through to downstream tasks. Explicit phoneme-to-grapheme (P2G) systems are uncommon by comparison; however, the hidden Markov model and Gaussian mixture model (HMM-GMM) architecture used in last-generation ASR systems (Mohri et al., 2001) might be described as an n-gram language model with a phoneme-to-grapheme component. This architecture assumes predefined mappings between words and its known pronunciations, and thus cannot capture the open-ended variability of disordered speech (in which a production might bear little or no phonological relationship to a target word). Further, considering the benefits of transfer learning for clinical (see §2.2), HMM-GMM models lack the flexibility and ergonomics of pre-trained transformer models.

3 BORT

3.1 Model Selection

Our models are a direct continuation of a pre-trained BART model as described by Lewis et al. (2020). We chose BART for several reasons. First, it uses a BPE tokenizer, which is able to encode arbitrary Unicode characters, including the entirety of the IPA. By contrast, BERT models use WordPiece tokenizers with finite token inventories. None of the pretrained BERT-like models we considered covered the IPA symbols used in English phonology, and expanding a WordPiece inventory is a non-trivial task. We were also motivated by the denoising task behind BART, since the synthesis of word errors is at least tangentially relevant to speech language pathology. Finally, unlike most models derived from the BERT architecture, BART features a left-to-right decoder ideal for generative tasks, and its pre-training task allows a mask token to represent one or more tokens, thus enabling us to overcome key limitations we encountered while revisiting our earlier work on analysis of connected speech from aphasic speakers (Adams et al., 2017) using techniques and models developed by Salem et al. (2022).

3.2 An IPA-to-Orthography Translation Task

Our self-supervised approach was formulated as a translation task, restoring partially-transformed documents back to the originals. We experimented with three word transforms: pronunciation, spelling, and noise. Ultimately, we were aiming for a system that could convert a mixture of IPA and orthography to its all-orthographic equivalent. However, despite an abundance of orthographic text, our pronunciation dictionary (described in more detail in §3.3) was limited to only 98K words during training. In an effort to avoid overfitting, we experimented with the other two less-constrained transforms.

Pronunciation transform. Our self-supervised approach was formulated as an IPA-to-orthography translation task. We used Wikipedia articles as a resource for orthographic text. Similar to masked language modeling, we randomly obscured words in each article, but instead of a special mask token, we replaced words with their pronunciations written in the IPA. Since the IPA includes letters from the English alphabet, which the tokenizer is likely to merge, we strategically inserted bullet symbols to maintain separation between phonemes (e.g. "shakedown" was transformed into "ʃe·k·d·aʊn"). We then trained the model to restore the text to its original orthographic form.

Spelling transform. Considering how English orthography is related to its phonology (however difficult a relationship it may be), we experimented with a spelling transform that could be formulated around any word, even those outside our pronunciation dictionary. For this task, for randomly selected words, we inserted a bullet symbol before each letter, forcing the tokenizer to treat each letter as a discrete token (e.g. "pizza" becomes "·P·I·Z·Z·A"). Note that we also used uppercase letters to avoid any overlap with the English

IPA symbols.

Noise transform. We also include a denoising task like the one used in pre-training the BART models (Lewis et al., 2020), wherein input tokens are inserted, deleted, and replaced at random. We used these same kinds of transforms, except we only apply noise to pronounced or spelled words. For these words, we randomly replace, insert, and delete tokens in the word. Insertions were limited to letters from the appropriate alphabet (either English or IPA).

Experimental configurations. We experimented with several variations of the above transforms, replacing 10% of pronounceable words with IPA, 10% of words with spellings, as well as a combined pronounced/spelled configuration (at 10% each). We repeated these configurations with noise added at a 5% probability. Table 1 summarizes the pre-training configurations used in this paper.

3.3 Data Preparation

Data sources. We based our pronunciation dictionary on version 0.7b of CMUDict (Carnegie Mellon University, 2014).[3] We converted their ARPABet entries to IPA using hard-coded rules, removing stress symbols. As carrier text for our self-supervised task, we used the 20220301.en version of Wikipedia provided by Huggingface Datasets.[4]

Word/article associations. For most purposes, the synthetic dataset was designed to be as open-ended as resources allowed, applying the transforms according to the word frequency distributions of Wikipedia. However, considering how the pronunciation dictionary was by far the data bottleneck (and thus the primary focus of our validation strategies), we needed a way to intentionally and efficiently find a high quality context for a given word. To this end, we paired each word in the pronunciation dictionary with a unique article based on an algorithm based on word frequencies. Tallying how many times our dictionary words appeared in each article, we assigned each dictionary word a unique article with a simple algorithm: beginning with the rarest word, we chose the next available article with the highest count for that word. We used a few simple rules to avoid specific types of low-quality articles.[5]

Data splits. We split the final list of 122K words and their associated Wikipedia articlesinto training, validation, and test sets (80%, 10%, and 10%, respectively). The remaining 6.5M Wikipedia articles with no associated word were added to the training set. Anticipating approximately 3100 target words that we would use to evaluate fine-tuning in §4, we placed these words in the test set. Three pairs of words were found only in overlapping articles; we placed these words in the training split.

Pronunciation transforms were only allowed for words which could be found in the split's dictionary. A spelling transform was only allowed for words which were *not* be found in *another* split's dictionary.

Training and validation inputs. We iterated over each Wikipedia article, transforming and noising pronounceable/spellable words at the rates specified in §3.2. Training inputs were limited 1000 BPE tokens to fit within BART's attention limit. If the article had an associated word, the sample was trimmed so the median-position occurrence of that word was at the center, otherwise a trim window was chosen at random. To minimize computation required for the validation set, each sample was limited to 100 tokens, with the median-position occurrence of that word at the center. If the experiment included a pronunciation transform, the associated dictionary word in the center was always pronounced.

Since the training set contained the most words, and the test set contained a number of high frequency words held out for fine-tuning evaluation, we found that the validation set only transformed at a word rate of 7–8% in practice, short of the 10% observed in the training inputs.

Test data. For evaluation purposes, only a single instance of the associated dictionary word was pronounced. Inputs were limited to 100 tokens, with the median-position occurrence of that word at the center. Neither the spelling nor the noise transforms were applied during testing.

3.4 Pre-training

Model weights were initialized from the 140M parameter BART-BASE pre-trained model found on the Fairseq website.[6] We based our hyperparameters on (Lewis et al., 2020), using the fairseq toolkit (Ott et al., 2019). Training targeted a categorical cross-entropy loss, with a learn rate of 10^{-5} and a maximum batch size of 12288 tokens, resulting in 317K batches per epoch. We computed the validation loss every 1K batches, and restored the best model after validation loss failed to improve for 63K batches (20% of the Wikipedia data).

3.5 Evaluation

We evaluated our pre-trained models in terms of accuracy and character error rate (CER) of the test set.

[3] https://github.com/Alexir/CMUdict/blob/7a37de7/cmudict-0.7b

[4] https://huggingface.co/datasets/wikipedia/

[5] We noticed our algorithm favored articles like "List of people with surname Carpenter" and "Mercury (disambiguation)," which have high word frequencies for "carpenter" and "mercury," respectively. In Wikipedia, these articles function only as lists of links to other articles, and we used them only as a last resort.

[6] https://github.com/facebookresearch/fairseq/tree/main/examples/bart

Pre-trained Model	Pron.	Spell.	Noise	**Example of Transformed Text**
BORT-PR	10%	—	—	he retaliates by s·pɹɛdɪŋ false rumours
BORT-SP	—	10%	—	he ·R·E·T·A·L·I·A·T·E·S by spreading false rumours
BORT-PR-SP	10%	10%	—	he ·R·E·T·A·L·I·A·T·E·S by s·pɹɛdɪŋ false rumours
BORT-PR-NOISY	10%	—	5%	he retaliates by ɔɪs·pɛdɪŋ false rumours
BORT-SP-NOISY	—	10%	5%	he ·E·T·A·K·I·A·T·E·S by spreading false rumours
BORT-PR-SP-NOISY	10%	10%	5%	he ·E·T·A·K·I·A·T·E·S by ɔɪs·pɛdɪŋ false rumours

Table 1: The various self-supervised training configurations for our models, indicating the percentage of words replaced with either pronunciations in the IPA or spelled, with or without noise added to the replacements. Example text from Wikipedia is shown to demonstrate the transformations. Bullet characters are inserted to enforce separation between those letters which would otherwise be merged during BPE text encoding.

Recall that in the test set we only applied the transform we were most interested in for this work: pronunciation. Considering how each correct output contains nearly a hundred words of text also found in the input—a trivial task to predict—we defined CER as the number of character errors divided by the length of *only the target word*. Text case and whitespace were ignored during evaluation. Only the models applying the pronunciation transform could be directly evaluated in this manner, so we do not include a baseline for this evaluation, and some models are only evaluated indirectly in §4.

3.6 Results

Configurations which included the noise transform did better overall than those without. The one with spelling (BORT-PR-SP-NOISY) was the overall best with a 15.1% CER, compared to BORT-PR-NOISY at 19.5%. To a lesser extent, spelling improved CER in the models without a noise transform: BORT-PR-SP and BORT-PR had a CER of 23.4% and 22.4%, respectively. Accuracy followed a similar pattern, with BORT-PR-SP-NOISY, BORT-PR-NOISY, BORT-PR-SP, BORT-PR showing accuracies of 64.5%, 61.8%, 51.6%, and 55.0%, respectively.

As for our overfitting concerns, most of our models showed an increase in validation loss (i.e. early stopping was triggered) before completing a full epoch of 6.5M Wikipedia articles. The pronunciation-only model BORT-PR after about 1.0M articles. Adding only the spelling transform for BORT-PR-SP nearly tripled the training duration to about 2.8M articles, while adding only the noise transform for BORT-PR-NOISY actually shortened training to about 0.6M articles. The combination of spelling and noise for BORT-PR-SP-NOISY trained for about 4.4M articles. The only models which completed a full epoch were those trained without the pronunciation transform, with BORT-SP and BORT-SP-NOISY training for about 6.9M and 7.9M articles, respectively.

Configuration	**CER**	**Accuracy**
BORT-PR	0.234	0.550
BORT-PR-SP	0.224	0.516
BORT-PR-NOISY	0.195	0.618
BORT-PR-SP-NOISY	0.151	0.645

Table 2: Character error rates (CER) and accuracies for the pre-training task, Only those configurations which applied the pronunciation transform are shown.

3.7 Discussion

Out of the pre-training configurations we evaluated with CER, the best performance was seen with BORT-PR-SP-NOISY, the model trained on all three transforms (phonology, spelling, and noise). This indicates that all three transforms were useful for the task of restoring a word from its pronunciation. Additionally, noise was clearly helpful for these models, since the next best configuration also used the noise transform (BORT-PR-NOISY).

The evaluation at this stage was lower than we expected, but this can largely be explained in terms of how the problem and its evaluation were formulated. Our hold-out rules were unusually strict to ensure the model could not memorize any of the words used during fine-tuning, heavily biasing the test data toward common English words. Second, CER is an imperfect evaluation measure for a model which operates on subword tokens, and one which isn't strictly a P2G translator. Additionally, as we emphasize in §4, evaluation on 1-best is a poor measure of the usefulness of a model intended for use in a complex pipeline.

4 Fine-tuning BORT

4.1 Data

Fine-tuning data consisted of 2,234 transcripts from 339 people with aphasia (PWA) from the English AphasiaBank database (MacWhinney et al., 2011),

a widely-used repository of recorded and transcribed administrations of a standardized protocol consisting of a variety of tasks including discourse tasks, meant for use in studying language and cognitive sequelae of post-stroke aphasia.[7] Demographic characteristics of the participants are included in Appendix A. The transcripts used in this study were from one of nine tasks designed to elicit discourse; the tasks themselves are described in more detail in Appendix B. The tasks were transcribed by human annotators according to the CHAT transcription manual (Codes for the Human Analysis of Transcripts; MacWhinney 2000).

In the AphasiaBank transcripts, non-word paraphasias are transcribed in IPA, whereas lexical paraphasias are written in their orthographic form. For each paraphasia, whenever possible, the human annotator also identified the target word for that production (i.e., the word the person intended to say). Given the paraphasia-target pair, the annotator also categorized the paraphasia according to whether it was a real word, whether it was phonologically related to the target, and whether it was semantically related to the target.

We filtered the transcripts to just the PWA's language, and removed annotations irrelevant to the task (e.g., gestures). Then, we prepared the transcripts for training in two ways: an "easy" way and a "hard" way. In the "hard" task, the model learned the task that the human annotators performed in AphasiaBank: to predict the target word given the pronunciation of the paraphasia (and the surrounding context). In the "easy" task, we instead replaced the paraphasia pronunciation with the correct pronunciation for its associated target word. That is, in the "easy" task, we train the model to fill in correct pronunciations for words with their corresponding orthographic form, and in the "hard" task we train the model to fill in paraphasias (i.e., incorrect pronunciations) with the intended orthographic word.

For both the "easy" and "hard" tasks, we only considered paraphasias with a known target provided by the human annotator. For the "easy" task, we additionally removed paraphasias where there was not a known pronunciation of the orthographic form of the target in our pronunciation dictionary. This left us with 10,120 paraphasias. With the "hard" task, for the real word paraphasias, we instead removed paraphasias where there was not a known pronunciation of the orthographic form of the *paraphasia* in the pronunciation dictionary. This left us with 9,781 paraphasias for the "hard" task.

Each usable production was prepared with the full context of its transcript: all productions in the transcript were prepared as phonemes separated by bullets (as in §3.2), whether it be the target pronunciation or the paraphasia itself, and the production for the model to predict was marked with surrounding angle brackets. Some of these prepared samples were too long when tokenized, and thus trimmed to the maximum length (1024 tokens) in such a way that the maximum possible context on either side of the paraphasia was preserved. Part of a prepared example with the target "screwed" is shown below:

> **"Easy"**: and it <s·kɹu·d > up. but I went to to the hɑs·pɪtəl. and my brain s·kæn·d.
>
> **"Hard"**: and it <ʃkɹu·d > up. but I went to to the ɑs·pɪtəl. and my brain s·tæn·d.

For "easy," we substituted the correct pronunciation for screwed ("screwed"), while in "hard" we included the paraphasia ("shcrewed").

4.2 Training

Given either a correct ("easy") or incorrect ("hard") pronunciation, we fine-tuned each pre-trained BORT model to predict the intended word. As a baseline, we fine-tuned from the unmodified source model, BART-BASE. . As in pre-training, we adapted code from the fairseq sequence modeling toolkit (Ott et al., 2019). We used 10-fold cross validation with participant as the grouping factor, sequentially holding one fold out as valid set, a second fold as test set, and the remaining eight folds as the training set. We trained each model until early stopping occurred using loss on the validation set after 20 epochs without improvement. Training hyperparameters were the same as §3.4 but with an effective batch size of 4000 tokens.

4.3 Evaluation

We evaluated performance for the "easy" and "hard" tasks using CER and accuracy. As we did in §3.5, we calculated CER between the top model prediction and the human identified target for each paraphasia in the test set. We also calculated top 1 accuracy (the top model prediction matched the human identified target) and top 5 accuracy (the human identified target was within the top 5 model predictions) for each of the fine-tuned models' predictions on the test set. We determined whether disagreements between top 1 accuracy of the different models were significant using McNemar's test with continuity correction (McNemar, 1947) and Bonferroni correction (Haynes, 2013). We conducted this test for all six models versus the baseline, as well as the best performing model versus all other models, for both "easy" and "hard" tasks, leading to 24 comparisons in total. Accounting for multiple comparisons and using an alpha of 0.05, a p-value of < 0.00208 was retained as the level of statistical significance . Finally,

[7] Our snapshot of the transcripts was copied from AphasiaBank on March 8th, 2023, and the corresponding demographic metadata downloaded on April 23, 2023.

for the best configuration in the "hard" task, we also calculated top 1 accuracy stratified by AphasiaBank task and error type, presented in Appendix C.

4.4 Results

CER, top 1 accuracy, and top 5 accuracy on the test set is shown for the "easy" and "hard" tasks in Tables 3a and 3b respectively. For both "easy" and "hard" tasks, all of the BORT models had significantly higher performance than BART-BASE at top 1 accuracy according to McNemar's test with $p < 0.00208$ for all comparisons.

In the "easy" task, the baseline configuration (BART-BASE) led to top 1 accuracy 72.5%, and CER 22.8%. Out of the models trained with noise, best performance was seen in BORT-PR-SP-NOISY with top 1 accuracy 89.5%. However, BORT-PR-SP saw the best performance out of all models, with a 90.1% chance of correctly predicting the appropriate word for a given correct pronunciation and a CER of just 5.7%, improving on the baseline by a relative 24% and 75%, respectively. The accuracy was significantly higher than all other models with $p < 0.00208$ for all comparisons. Allowing for five chances to get the correct prediction, this model achieved 94.7% accuracy.

For the "hard" task, the baseline achieved 36.3% top 1 accuracy and 60.6% CER. Out of the configurations which did not apply noise, BORT-PR-SP achieved the highest top 1 accuracy of 45.6% and CER 44.7%. The best pre-training configuration was BORT-PR-SP-NOISY, which applied all three transforms. It achieved top 1 accuracy 46.7% and CER 42.0%, improving on the baseline by a relative 29% and 31%, respectively. The top 1 accuracy was significantly higher than most models with $p < 0.00208$, except for BORT-PR-SP ($p = 0.020$) and BORT-PR-NOISY ($p = 0.023$). BORT-PR-SP-NOISY also had 65.6% top 5 accuracy. Accuracy within top 1–20 predictions of the baseline and best performing models for the "easy" and "hard" tasks can be seen in Figure 1. Performance increases for all models as we allow more chances to find the correct target, but the order of performance remains the same. Additional results—namely those stratified by AphasiaBank task and error types—can be found in Appendix C.

4.5 Discussion

The fine-tuning configurations without the pronunciation transform (BORT-SP and BORT-SP-NOISY) were the lowest performing of the BORT models, but they still had significantly higher top 1 accuracy than the fine-tuned BART-BASE model. Moreover, for both "easy" and "hard" tasks, the best performing models were trained with both the pronunciation and spelling transforms, and either with or without noise. This pattern implies that there is enough overlap between

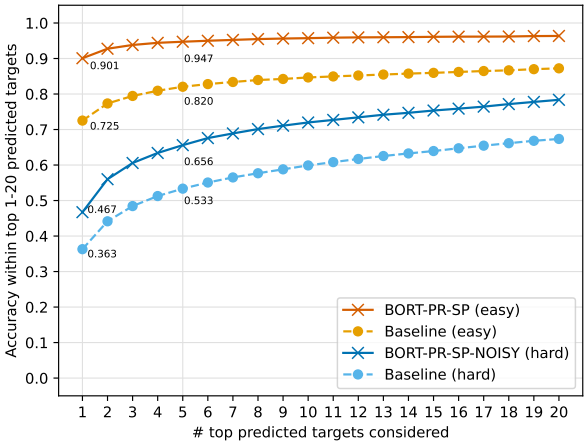

Figure 1: Accuracy within top 1-20 predictions of baseline and best performing models for "easy" and "hard" fine-tuning tasks.

orthography and phonology in the English language that pre-training the LLM to spell helped the model to perform the task at hand. Considering the G2P frame of reference—though we note again that this is a loose comparison (see §3.7)—a 5.7% CER is on par with what could be expected from a model strictly translating between phonemes and graphemes.

Aphasic speech is characterized by paraphasias, which can be considered "noisy" productions, so it stands to reason that learning to de-noise productions would help the model with the "hard" fine-tuning task. As we hypothesized, the "hard" task saw the best performance from the pre-training configuration with all three transforms (pronunciation, spelling, and noise), although its performance was not *significantly* different than the configuration without spelling (BORT-PR-NOISY) or without noise (BORT-PR-SP). Contrary to what we observed in §3.5, for the "easy" task, noise did not seem to help the models, and the best performing configuration was one that did not include noise in pre-training (BORT-PR-SP).

Moreover, the top 1 accuracy performance of this model was significantly different than all other "easy" models. This is surprising since the evaluation for pre-training and the "easy" task were quite similar, being a phoneme-to-grapheme translation task with and without context, respectively. This might be explained by the stricter hold-out rules during pre-training—we had no vocabulary restrictions during fine-tuning—or by the shift in data domain (Wikipedia vs. AphasiaBank). It is difficult to say with certainty why this discrepancy occurred, but perhaps noise was most helpful for language from a very diverse corpora (Wikipedia), while in the more constrained tasks from AphasiaBank, the more limited vocabulary did not benefit from synthetic variability.

Pre-training Configuration	CER Top 1	Accuracy Top 1	Accuracy Top 5
BORT-PR	0.083	0.869	0.931
BORT-SP	0.106	0.843	0.906
BORT-PR-SP	0.057	**0.901**	0.947
BORT-PR-NOISY	0.089	0.863	0.925
BORT-SP-NOISY	0.096	0.848	0.911
BORT-PR-SP-NOISY	0.060	0.895	0.947
BART-BASE	0.228	0.725	0.820

(a) "Easy" Task

Pre-training Configuration	CER Top 1	Accuracy Top 1	Accuracy Top 5
BORT-PR	0.462	0.451	0.634
BORT-SP	0.526	0.401	0.579
BORT-PR-SP	0.447	*0.456*	0.641
BORT-PR-NOISY	0.452	*0.458*	0.640
BORT-SP-NOISY	0.469	0.446	0.625
BORT-PR-SP-NOISY	0.420	**0.467**	0.656
BART-BASE	0.606	0.363	0.533

(b) "Hard" Task

Table 3: Accuracies for each pre-trained model after fine-tuning to our two tasks. Bold font indicates accuracy was significantly different from all other models, with the exception of those italicized, according to McNemar's test.

5 Conclusion

In §3, we pre-trained BORT to accept a mixture of orthography and IPA. During training and validation, we used one to three different transforms of words (pronunciation, spelling, noise) and trained the model to restore the words to their original form. We directly evaluated the four models trained on pronunciation using a test set of the Wikipedia data by testing the CER of their performance at restoring a word from its pronunciation. In §4, we evaluated all six models (and the baseline) by further fine-tuning them to restore words from productions in aphasic speech. This allowed us to evaluate the applicability of our mixed orthography/IPA LLM to a clinical task.

Our best BORT configurations achieved high accuracy and low CER rates for the "easy" fine-tuning task, with the fine-tuned accuracy as high as 90%. This indicates we were able to successfully produce a LLM that can accept both orthography and IPA. Moreover, observing differences between accuracy and CER revealed that even when our fine-tuned models incorrectly predicted the target word, they still may have found close-by words. For instance, considering the "easy" task, the best performing model picked the wrong target 10% of the time, but it achieved an average CER of just 5.7%. This indicates that there were likely instances where even though the LLM predicted the wrong target, it picked a similar word with overlapping letters with the target word.

Future work will improve on these pre-trained models. In §3.1 we hypothesized BART's generative architecture and denoising task were advantages for our use case. With model selection, though, we find certain tradeoffs, and we would like to test whether these advantages outweigh unique functionalities found in other models. In particular, models designed to operate at a character level (e.g. CANINE, Clark et al., 2022) could overcome other limitations of BORT (e.g. our explicit demarcation of phonemic and orthographic regions), and is perhaps generally well-suited for the task at hand.

An additional area of future work will consist of exploring alternative training strategies. In the present work, we were quite strict with regard to preparing the test split, withholding the most frequent English words because they appeared in our AphasiaBank evaluations. As our focus turns more toward downstream tasks, we will update our holdout methods to prioritize a stronger pre-trained model, holding out only as much data as needed for validation. Further, seeing how our models did not train for an entire pass through the Wikipedia data, we will adjust the training schedule (e.g. a ramp-up in the learning rate) and task configuration (e.g. word transform rates) to ensure we get the most out of the pre-training stage.

For our aphasia-specific application, we see room to improve the noise transform with more strategic approaches. Phoneme errors could be made more realistic with statistically or linguistically informed approaches (e.g. replacing phonemes with similar phonemes). To better prepare a model for semantic errors, whole-word replacements could be made with semantically similar words.

Limitations

The models presented here were trained with the basic inventory of English phonemes found in CMUDict. However, a more fine-grained phonetic analysis would require a pronunciation dictionary with more narrowly defined entries. Additionally, while this paper focused on models trained with English-only resources (pre-trained BART-BASE, English Wikipedia text, CMUDict, and the English AphasiaBank), the techniques should be applicable to non-English language models as well. Finally, from a clinical standpoint, the model we describe in this paper assumes the existence of transcribed input (from either a manual or automated source, discussed in detail in § 2.1); in its

current form, this represents a limitation to its clinical implementation, though not to its use in research settings with archival or newly-transcribed datasets.

Ethics Statement

Our use of the AphasiaBank data was governed by the TalkBank consortium's data use agreement, and the underlying recordings were collected and shared with approval of the contributing sites' institutional review boards. Limitations exist regarding accents and dialect, which in turn would affect the scenarios in which a system based on our model could (and should) be used. It should also be noted that these models and any derived technology are not meant to be tools to diagnose medical conditions, a task best left to qualified clinicians.

Acknowledgements

We thank our anonymous reviewers for their helpful insights and detailed feedback. This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award 5R01DC015999 (Principal Investigators: Bedrick & Fergadiotis). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Stefanie Abel, Klaus Willmes, and Walter Huber. 2007. Model-oriented naming therapy: Testing predictions of a connectionist model. *Aphasiology*, 21(5):411–447.

Joel Adams, Steven Bedrick, Gerasimos Fergadiotis, Kyle Gorman, and Jan van Santen. 2017. Target word prediction and paraphasia classification in spoken discourse. In *BioNLP 2017*, pages 1–8, Vancouver, Canada,. Association for Computational Linguistics.

Nancy Azevedo, Eva Kehayia, Gonia Jarema, Guylaine Le Dorze, Christel Beaujard, and Marc Yvon. 2023. How artificial intelligence (AI) is used in aphasia rehabilitation: A scoping review. *Aphasiology*, pages 1–32.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's disease detection. In *Interspeech 2020*, pages 2167–2171. ISCA.

Carnegie Mellon University. 2014. Carnegie mellon university pronouncing dictionary (CMUDict), version 0.7b. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Marianne Casilio, Gerasimos Fergadiotis, Alexandra C. Salem, Robert C. Gale, Katy McKinney-Bock, and Steven Bedrick. 2023. Paralg: A paraphasia algorithm for multinomial classification of picture naming errors. *Journal of Speech, Language, and Hearing Research*, pages 1–21.

Soojin Cho-Reyes and Cynthia K. Thompson. 2012. Verb and sentence production and comprehension in aphasia: Northwestern Assessment of Verbs and Sentences (NAVS). *Aphasiology*, 26(10):1250–1277.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Lisa A. Edmonds and Swathi Kiran. 2006. Effect of Semantic Naming Treatment on Crosslinguistic Generalization in Bilingual Aphasia. *Journal of Speech, Language, and Hearing Research*, 49(4):729–748.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Gerasimos Fergadiotis, Kyle Gorman, and Steven Bedrick. 2016. Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology*, 25(4S):S776–S787.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54, Grenoble, France. Association for Computational Linguistics.

Kathleen C Fraser, Jed A Meltzer, Naida L Graham, Carol Leonard, Graeme Hirst, Sandra E Black, and Elizabeth Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Robert Gale, Julie Bird, Yiyi Wang, Jan van Santen, Emily Prud'hommeaux, Jill Dolata, and Meysam Asgari. 2021. Automated scoring of tablet-administered expressive language tests. *Frontiers in Psychology*, 12.

Robert C. Gale, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2022. The post-stroke speech transcription (PSST) challenge. In *Proceedings of the RaPID Workshop - Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments - within the 13th Language Resources and Evaluation Conference*, pages 41–55, Marseille, France. European Language Resources Association.

Winston Haynes. 2013. Bonferroni Correction. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 154–154. Springer New York, New York, NY.

Yvette Hou, Aileen Zhou, Laura Brooks, Daniella Reid, Lyn Turkstra, and Sheila MacDonald. 2023. Rehabilitation access for individuals with cognitive-communication challenges after traumatic brain injury: A co-design study with persons with lived experience. *International Journal of Language & Communication Disorders*, pages 1460–6984.12895.

Itay Itzhak and Omer Levy. 2022. Models in a spelling bee: Language models implicitly learn the character composition of tokens. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5061–5068, Seattle, United States. Association for Computational Linguistics.

A Jacks, K L Haley, G Bishop, and T G Harmon. 2019. Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions. *Folia Phoniatrica et Logopaedica*, 71(5-6):286–296.

Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. 2021. PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS. In *Proc. Interspeech 2021*, pages 151–155.

Edith Kaplan, Harold Goodglass, and Sandra Weintraub, editors. 2001. *Boston naming test*, 2. ed edition. Lippincott, Williams & Wilkins, Philadelphia.

Diane L Kendall, Rebecca Hunting Pompon, C Elizabeth Brookshire, Irene Minkina, and Lauren Bislick. 2013. An analysis of aphasic naming errors as an indicator of improved linguistic processing following phonomotor treatment. *American Journal of Speech-Language Pathology*, 22(2):S240–S249.

Andrew Kertesz. 2012. Western Aphasia Battery–Revised. Technical report, American Psychological Association. Type: dataset.

Swathi Kiran, Carrie Des Roches, Isabel Balachandran, and Elsa Ascenso. 2014. Development of an Impairment-Based Individualized Treatment Workflow Using an iPad-Based Software Platform. *Seminars in Speech and Language*, 35(01):038–050.

Duc Le, Keli Licata, and Emily Mower Provost. 2017. Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study. In *Proc. Interspeech 2017*, pages 294–298.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions.

Duanchen Liu, Zoey Liu, Qingyun Yang, Yujing Huang, and Emily Prud'hommeaux. 2022. Evaluating the performance of transformer-based language models for neuroatypical language. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3412–3419, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Heather MacFarlane, Alexandra C Salem, Steven Bedrick, Jill K Dolata, Jack Wiedrick, Grace O Lawley, Lizbeth H Finestack, Sara T Kover, Angela John Thurman, Leonard Abbeduto, and Eric Fombonne. 2023. Consistency and reliability of automated language measures across expressive language samples in autism. *Autism research : official journal of the International Society for Autism Research*, 16(4):802–816.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Irene Minkina, Megan Oelke, Lauren P Bislick, C Elizabeth Brookshire, Rebecca Hunting Pompon, Joann P Silkes, and Diane L Kendall. 2015. An investigation of aphasic naming error evolution following phonomotor treatment. *Aphasiology*, pages 1–19.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2001. Weighted finite-state transducers in speech recognition. *Departmental Papers (CIS)*, page 11.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Perez, Zakaria Aldeneh, and Emily Mower Provost. 2020. Aphasic Speech Recognition Using a Mixture of Speech Intelligibility Experts. In *Interspeech 2020*, pages 4986–4990. ISCA.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association : JAMIA*, 27(11):1784–1797.

Alan Prince and Paul Smolensky. 1997. Optimality: From neural networks to universal grammar. *Science*, 275(5306):1604–1610.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Alexandra C. Salem, Robert Gale, Marianne Casilio, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2022. Refining semantic similarity of paraphasias using a contextual language model. *Journal of Speech, Language, and Hearing Research*, pages 1–15.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. PhonemeBERT: Joint Language Modelling of Phoneme Sequence and ASR Transcript. In *Proc. Interspeech 2021*, pages 3236–3240.

Iván G. Torre, Mónica Romero, and Aitor Álvarez. 2021. Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Applied Sciences*, 11(19):8872.

Mieke Van De Sandt-Koenderman. 2004. High-tech AAC and aphasia: Widening horizons? *Aphasiology*, 18(3):245–263.

Marjo Virnes, Eija Kärnä, and Virpi Vellonen. 2015. Review of research on children with autism spectrum disorder and the use of technology. *Journal of Special Education Technology*, 30(1):13–27.

Grant M. Walker and Gregory Hickok. 2016. Bridging computational approaches to speech production: The semantic-lexical-auditory-motor model (SLAM). *Psychonomic Bulletin & Review*, 23(2):339–352.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Transformer Based Grapheme-to-Phoneme Conversion. In *Interspeech 2019*, pages 2095–2099. ISCA.

Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. 2022. Mixed-phoneme bert: Improving bert with mixed phoneme and sup-phoneme representations for text to speech.

Appendix

A AphasiaBank Demographics

Demographic characteristics from the 339 participants are summarized in Table 4.

B Description of AphasiaBank Tasks

We used transcripts from nine tasks from AphasiaBank. Descriptions of each task are provided in Table 5. More information can be found on the AphasiaBank website.[8]

C Detailed Results

We calculated top 1 accuracy for our best performing model in the "hard" task, BORT-PR-SP-NOISY, stratified by AphasiaBank task type. These results are shown in Table 6. Performance was lowest for AphasiaBank's "Free Speech Samples" section with accuracies ranging from 33%–36%, followed by most of "Picture Descriptions" at 41.1%–44.5%, with the exception of Umbrella. The best performance was seen in the "Story Narrative" and "Procedural Discourse" sections with 53.3% and 52.2%, respectively, as well as the Umbrella picture description (61.8%). This pattern makes sense, since the fine-tuned models can use exposure to the task domain to learn what common vocabulary occurs in the tasks. Topics that are very open-ended, like the Free Speech Samples, instead could have a large range of possible targets for paraphasias.

[8] https://aphasia.talkbank.org/protocol/english/materials-aphasia

Characteristic	Value	Characteristic	Value
Age		**Aphasia Duration**	
M (*SD*)	61.8 (12.3)	*M* (*SD*)	5.4 (5.4)
Min - Max	25.6–90.7	Min - Max	0.08–44
Missing (*N*)	3	Missing (*N*)	16
Race			
White (*N*)	284	**WAB-R AQ**	
African American (*N*)	37	*M* (*SD*)	71.0 (19.6)
Asian (*N*)	2	Min - Max	10.8-99.6
Hispanic/Latino (*N*)	9	Missing (*N*)	37
Native Hawaiian / Pacific Islander (*N*)	2		
American Indian / Alaska Native (*N*)	1	**BNT-SF**	
		M (*SD*)	7.1 (4.6)
Mixed (*N*)	2	Min - Max	0-15
Other (*N*)	1	Missing (*N*)	69
Unavailable (*N*)	1		
Gender		**VNT**	
M (*N*)	199	*M* (*SD*)	14.3 (6.6)
F (*N*)	140	Min - Max	0-22
		Missing (*N*)	56
Years of Education			
M (*SD*)	15.4 (2.7)		
Min - Max	8–25		
Missing (*N*)	14		

Table 4: Demographic characteristics for the 339 participants at their first session, where available. WAB-R AQ is the Western Aphasia Battery-Revised Aphasia Quotient, and captures overall aphasia severity with higher values indicating lower severity (Kertesz, 2012). BNT-SF is the raw score from the Boston Naming Test-Short Form (Kaplan et al., 2001). VNT is the raw score from the Verb Naming Test (Cho-Reyes and Thompson, 2012). The BNT-SF and VNT are both confrontation picture naming tests, where the BNT-SF captures word retrieval deficits of object words and the VNT captures word retrieval deficits of action words.

Section	Task	Description
I: Free Speech Samples	Speech	The participant describes how their speech is currently.
	Stroke	The participant's story of his or her stroke.
	Important Event	A personal narrative with a wide range of possible topics.
II. Picture Descriptions	Window	A picture description task.
	Umbrella	A picture description task.
	Cat	A picture description task.
	Flood	A picture description task.
III. Story Narrative	Cinderella	The participant recounts a narrative of Cinderella, after reviewing pictures of central events of it.
IV. Procedural Discourse	Sandwich	The participant is asked to describe how to make a peanut butter & jelly sandwich.

Table 5: Descriptions of nine AphasiaBank tasks

AphasiaBank Task	*N*	Accuracy
I. Free Speech Samples		
Speech	265	0.340
Stroke	1620	0.357
Important Event	900	0.330
II. Picture Descriptions		
Window	711	0.414
Umbrella	1103	0.618
Cat	1197	0.445
Flood	180	0.411
III. Story Narrative		
Cinderella	3065	0.533
IV. Procedural Discourse		
Sandwich	740	0.522

Table 6: Number of samples (N) and top 1 accuracy for BORT-PR-SP-NOISY after fine-tuning to the "hard" task, stratified by AphasiaBank task.

Error Type	*N*	Accuracy
Phonological	4557	0.473
Semantic	3137	0.485
Neologism	1570	0.445
Morphological	417	0.350
Dysfluency	11	0.455
Multiple Types	82	0.488
Unknown Type	7	0.286

Table 7: Number of samples (N) and top 1 accuracy for BORT-PR-SP-NOISY after fine-tuning to the "hard" task, stratified by AphasiaBank error type annotations.