Context-Aware Lexical Stress Prediction and Phonemization for Ukrainian TTS Systems

Anastasiia Senyk[1], Mykhailo Lukianchuk[2], Valentyna Robeiko[2], Yurii Paniv[1],

[1]Ukrainian Catholic University, [2]Taras Shevchenko National University of Kyiv

Abstract

Text preprocessing is a fundamental component of high-quality speech synthesis. This work presents a novel rule-based phonemizer combined with a sentence-level lexical stress prediction model to improve phonetic accuracy and prosody prediction in the text-to-speech pipelines. We also introduce a new benchmark dataset with annotated stress patterns designed for evaluating lexical stress prediction systems at the sentence level.

Experimental results demonstrate that the proposed phonemizer achieves a 1.23% word error rate on a manually constructed pronunciation dataset, while the lexical stress prediction pipeline shows results close to dictionary-based methods, outperforming existing neural network solutions.

1 Introduction

Text-to-speech (TTS) systems are essential for enhancing human-computer interaction across various everyday applications, including virtual assistants, language learning tools, and navigation systems, while making digital content more accessible to people with visual impairments. The quality of TTS output depends heavily on accurate linguistic analysis, especially for languages with rich morphology like Ukrainian.

Effective text preprocessing is a critical step in language modeling pipelines, helping models generalize from limited data by transforming raw input into a standardized format (Oyucu and Dogan, 2023). This reduces linguistic variability and improves consistency. While similar results can be achieved without preprocessing, such approach typically requires significantly larger datasets and forces the model to learn a broader range of morphological and phonological irregularities, often at the cost of performance and interpretability. Moreover, post-training adjustments such as refining pronunciation or stress patterns become difficult without retraining or fine-tuning the entire model.

Phonemization and lexical stress prediction are two areas where preprocessing can significantly enhance TTS quality. Ukrainian, in particular, poses unique challenges due to its complex phonology and non-deterministic stress system (Moisiienko A. K., 2010; Pohribnyi, 1984). The language features rich inflectional morphology, frequent sound changes, such as consonant cluster reductions and different types of assimilation.

Moreover, Ukrainian has a non-deterministic stress system, where lexical stress may be fixed in some word forms, but in other cases varies based on syntactic or morphological context, influenced by factors such as free variation, where multiple stress placements are correct without a change in meaning (e.g., ба́йдуже vs. байду́же — “indifferently”); heteronyms, where identical spellings have different meanings depending on stress (e.g., за́мок — “castle” vs. замо́к — “lock”); and inflectional stress shifts, where morphological changes like case or number alter stress placement (e.g., низови́ни — nominative plural vs. низовини́ — genitive singular, both meaning “lowlands”).

Apart from that, phonemization is an essential preprocessing step that allows Text-to-Speech models to create speech from phoneme-based text, improving the match between text and audio data. This means that the quality of generated speech directly depends on the accurate mapping of graphemes to phonemes.

These complexities make accurate stress prediction and phonemization essential for natural-sounding speech synthesis.

In this work, we propose a framework for Ukrainian in which we introduce: a benchmark dataset for evaluating the performance of existing stress prediction systems; a context-aware model for lexical stress prediction; and a new rule-based phonemizer designed to reflect the unique phono-

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 96–104

logical characteristics of Ukrainian.

The benchmark, datasets, and source code are available at the following link: https://github.com/lang-uk/ukrainian-tts-preprocessing.

2 Related Work

2.1 Lexical Stress Prediction

Traditional methods for lexical stress prediction in Ukrainian have primarily relied on dictionary lookups and rule-based systems. One such approach, presented in (Syvokon, 2022), combines dictionary-based stress assignment with part-of-speech (POS) tagging to resolve certain ambiguous cases (e.g., доро́га (noun - "road") vs дорога́ (adjective - "expensive")). Although this hybrid approach achieves good overall accuracy, it is limited to heteronym pairs with clearly distinct grammatical features. Additionally, it does not handle out-of-vocabulary (OOV) or misspelled words.

More recently, neural network models have been applied to address stress prediction. As part of a Grapheme-to-Phoneme system, (van Esch et al., 2016) developed a lexical stress prediction approach using an LSTM-based model trained on phonemic representations of words. A similar approach, but applied to original word forms rather than phonemes, was used in (Smoliakov and Mykhailenko, 2022) for the Ukrainian language. Their method relied on dictionary-based training data for predicting stress within individual words. While these approaches effectively handle OOV words, they fail to resolve contextual stress ambiguity, as they do not consider the broader linguistic context of the sentence.

Some studies focus specifically on homograph disambiguation pairs, using contextual features or embeddings (Gorman et al., 2018; Nicolis and Klimkov, 2021; Hajj et al., 2022), though these methods target only a small set of word pairs and require extensive annotated data.

An initial attempt to incorporate contextual understanding into lexical stress prediction for Ukrainian was presented in (Mykhailenko, 2023), where a transformer-based model was trained on synthetic stress-annotated data generated using labels from (Syvokon, 2022) pipeline. While this demonstrated the potential of using synthetic data, the labeling approach was constrained by a predefined dictionary, limiting coverage for OOV words.

To improve generalization in low-resource settings, (Geneva et al., 2023) proposed a sentence-level neural model for Bulgarian, trained on synthetic data generated from an ASR-based stress detection pipeline. This strategy showed that large-scale machine annotation can be a viable alternative to manual labeling, which we similarly adopted in our approach.

2.2 Grapheme-to-Phoneme Conversion

Grapheme-to-phoneme (G2P) conversion, also known as phonemization, refers to the process of mapping written text to its corresponding phonemic representation (Prabhu and von der Wense, 2020). G2P is a crucial component in both speech synthesis and automatic speech recognition systems. Over the years, various approaches to G2P have been developed, ranging from rule-based methods (Mortensen et al., 2018; Sazhok and Robeiko, 2012) to statistical models (e.g. conditional and joint models (Chen, 2003), Hidden Markov Models (Taylor, 2005)) and modern neural architectures (e.g. LSTMs (Rao et al., 2015), CNNs (Yolchuyeva et al., 2019), Transformers (Prabhu and von der Wense, 2020)).

For the Ukrainian language, most of the available systems rely on rule-based approaches (Mortensen et al., 2018; Sazhok and Robeiko, 2012; Chaplinsky et al.). This is due in part to the limited availability of high-quality pronunciation dictionaries and the challenges in aligning phonemic and orthographic symbol sets.

Despite the relatively transparent orthography, achieving accurate grapheme-to-phoneme conversion requires careful attention to linguistic characteristics, such as assimilation. Many current solutions exhibit flaws in their approach:

- overgeneralizing rules (e.g. the rule regarding the assimilation of voiceless consonants, leading to берехти instead of the correct берегти "keep") (Sazhok and Robeiko, 2012)
- prompting the user to modify the input (e.g. adding a letter to accurately indicate a morphemic boundary in віджжилий "antiquated", підзземній "underground") (Chaplinsky et al.)
- ignoring all phonetic phenomena by applying naïve mapping between letters and phonemes (Mortensen et al., 2018)

Furthermore, many existing solutions are either not open source or are not publicly accessible for

evaluation. In this study, our aim is to address all these issues.

3 Approach to Stressifier

3.1 Benchmark Dataset: Ukrainian Lexical Stress Corpus

A standardized evaluation framework is crucial for comparing different systems with each other and estimating their performance for a task. However, to the best of our knowledge, there is no publicly available benchmark for Ukrainian lexical stress prediction, making it difficult to measure progress or compare approaches fairly.

To address this gap, we introduce the first benchmark dataset for Ukrainian lexical stress prediction. This dataset provides sentence-level context with gold-standard stress annotations, enabling consistent and meaningful evaluation across various approaches.

3.1.1 Dataset Composition

The dataset consists of 1,026 sentences manually annotated with primary stress by a native speaker. We intentionally retained OOV words and misspellings to reflect real-world language use better.

Sentence data was collected from two primary sources: 300 sentences were extracted from Wikipedia (Wikimedia), representing formal and encyclopedic language, and 438 from the Pluperfect GRAC corpus (Shvedova and Lukashevskyi, 2024), which introduces a wider variety of writing styles.

To facilitate the evaluation of contextual disambiguation for heteronyms, we identified 288 commonly used words exhibiting stress ambiguity, each occurring only once in the initial dataset. Stress pattern information for these words was obtained from the "Dictionaries of Ukraine" (Ukrainian Linguа-Information Foundation, 2008). We created an additional sentence for each ambiguous word, providing an alternate stress variant, augmenting the dataset with 288 new examples. This extension ensures a more balanced and comprehensive coverage of word pairs with the same spelling but different pronunciations.

An overview of key statistics for the benchmark dataset is provided in Table 1.

The dataset will be publicly available to encourage further research and reproducibility.

Statistic	Count
Total number of sentences	1,026
Unique word forms (including grammatical inflections, derivations, etc.)	6,439
Unique words with stress ambiguity (due to meaning or inflections)	640
Unique words with at least two stress forms in the dataset	296
Unique out-of-vocabulary words	1,005

Table 1: Overview of the Ukrainian Lexical Stress Benchmark

3.2 Model Architecture and Training

Developing a context-aware model for predicting lexical stress requires a large annotated dataset. However, there is currently no publicly available dataset for lexical stress in Ukrainian. To address this, we adopted a synthetic data generation approach inspired by (Geneva et al., 2023), enabling us to construct a scalable set of training examples without relying on manually labeled corpora.

While manual labeling remains the most accurate method, it is costly and time-consuming. To mitigate this, we utilize natural speech, which provides prosodic features such as pitch, duration, and intonation. These acoustic cues serve as a rich source of weak supervision and form the basis for pseudo-annotation.

3.2.1 Synthetic Stress Corpus

For automatic speech recognition (ASR), we selected the Wav2Vec2 model (Baevski et al., 2020), configured to transcribe audio with the Ukrainian alphabet and stress mark.

As the base for training, we used the Common Voice 19 dataset (Ardila et al., 2020), consisting of approximately 30,000 sentences, split into training, development, and test subsets. Pseudo-stress labels were generated using the Ukrainian Word Stress tool (Syvokon, 2022), configured with the OnAmbiguity.Skip option (skip the stress label when the system could not fully disambiguate a given case).

When the tool failed to assign stress, we employed a model-based fallback using Ukrainian Accentor (Smoliakov and Mykhailenko, 2022).

Once the model was trained, to refine the assigned stress labels, we applied post-correction using dictionary lookups. This approach resulted in a stress prediction accuracy of 93.81% at the word level and 72.00% at the sentence level, evaluated on a test subset. Words with fewer than two vowels

were excluded from the evaluation.

After that, we applied that pipeline to the Voice of America Ukrainian speech corpus (Smoliakov, 2022), followed by sentence cleaning and filtering, resulting in a synthetic dataset of approximately 135,000 sentences with stress marks containing around 80,000 unique words.

3.2.2 Model Setup

We trained a grapheme-to-phoneme model based on the ByT5 architecture (Zhu et al., 2022) to perform sentence-level lexical stress prediction. We selected this model because it operates on byte tokens, making it convenient to adapt to new languages without tokenizer-introduced bias. The model was trained on the annotated Voice of America dataset for 10 epochs using a learning rate of 0.0002, achieving a character error rate (CER) of 0.58%. The training was performed on normalized text to reduce noise and improve generalization.

To manage input length during model inference, each sentence was split into chunks of up to 150 characters before being processed by the model to mitigate long-context performance problems due to the encoder-decoder architecture of ByteT5. As the model operates on normalized text, the outputs were then merged with the original text to restore punctuation, capitalization, and special characters.

3.2.3 Evaluation

We evaluated the proposed model by comparing it against three established Ukrainian lexical stress systems: Ukrainian Accentor (Smoliakov and Mykhailenko, 2022), Ukrainian Accentor Transformer (Mykhailenko, 2023), and Ukrainian Word Stress (Syvokon, 2022). In the Ukrainian Word Stress system, when multiple stress options were retrieved during a dictionary lookup, disambiguation was attempted using the POS tags of the word in its sentence context and the grammatical features of the retrieved word forms. If disambiguation was not possible, two strategies were used to handle the ambiguity: `OnAmbiguity.First`, which selects the first retrieved stress variant, and `OnAmbiguity.Skip`, which skips stress labeling for that word. We tested the Ukrainian Word Stress under both disambiguation strategies.

We assess each approach using the following metrics:

- **Word-Level Accuracy**: Percentage of words with the correctly placed stress.
- **Sentence-Level Accuracy**: Percentage of sentences in which all words are correctly stressed.
- **Ambiguous Word Accuracy**: Accuracy on context-dependent words that exhibit stress ambiguity due to meaning or grammatical inflections.
- **Unambiguous Word Accuracy**: Accuracy on words with only one valid stress pattern.
- **Mean Macro F1 (Ambiguous Word Pairs)**: Macro-averaged F1 score over ambiguous word pairs, reflecting the model's ability for contextual stress prediction.

It is important to note that words containing fewer than two vowels were excluded from the evaluation.

3.2.4 Results and Analysis

A detailed comparison of the evaluation results across all systems is presented in Table 2.

The `ByT5 G2P` model demonstrates strong performance across all evaluation metrics, outperforming the `Ukrainian Accentor` baseline and reaching the dictionary-based `Ukrainian Word Stress` system in most tasks. The system also outperforms `Ukrainian Accentor Transformer`, except for unambiguous words, where the latter achieves higher accuracy, likely due to its reliance on dictionary-derived labels during training.

The highest overall performance is achieved through a hybrid approach that combines the `ByT5 G2P` model with `Ukrainian Word Stress (OnAmbiguity.Skip)`. In this setup, dictionary-based predictions are used when disambiguation is possible; otherwise, we used the ByT5 G2P model to provide the stress assignment. This hybrid strategy yields the best sentence-level accuracy (52.0%) and word-level accuracy (92.5%), highlighting the effectiveness of integrating deterministic and neural methods for stress prediction.

Among all systems, `Ukrainian Word Stress (First)` achieves the best performance on ambiguous words, reaching 64.3% accuracy and a Mean Macro F1 score of 47.3%. This is primarily due to its use of part-of-speech–based disambiguation and a consistent fallback to one of the possible listed stress variants when ambiguity is unresolved.

It is important to note that the classification of words as ambiguous or unambiguous was based

on the same dictionary used internally by the Ukrainian Word Stress tool. The system does not achieve 100% accuracy on unambiguous words due to inherent inconsistencies in the dictionary itself and the prioritization of capitalized over lowercase forms.

4 Approach to Phonemization

4.1 Motivation for a Rule-Based Approach

In this work, we present a new rule-based G2P system designed specifically for the Ukrainian language. The rule-based paradigm was selected for two primary reasons:

1. The scarcity of high-quality pronunciation data for Ukrainian, which limits the applicability of data-driven methods.

2. The relatively consistent and transparent mapping between graphemes and phonemes in Ukrainian orthography.

Despite its advantages, the rule-based approach comes with certain limitations:

1. As the number of rules increases, the system becomes increasingly complex and difficult to maintain.

2. Interactions among rules can lead to unexpected or undesired outputs.

4.2 Symbol Inventory and Phonemic Representation

The grapheme-to-phoneme conversion rules were derived from an analysis of linguistic studies on Ukrainian phonetics and phonology (Moisiienko A. K., 2010; Pohribnyi, 1984).

Internally, the system uses a custom set of transcription symbols based on the Ukrainian alphabet. After rule application, these symbols are converted into their corresponding International Phonetic Alphabet (IPA)[1] representations.

The system produces IPA phonemic transcription, with a phoneme inventory consisting of 52 symbols (see Appendix A.). These reflect the articulatory features of Ukrainian phonemes, omitting diacritics for distinctions that are not phonemically contrastive in the language (e.g., dental vs. alveolar articulation). The Ukrainian phoneme /в/ is realized with two phonetically distinct allophones, both of which are treated as separate phonemes in the system (bilabial /w/ and labio-dental /v/). Likewise, palatalized variants of hushing sibilants, labials, and velars are represented as distinct phonemes (ʃʲ, ʒʲ, xʲ, ɦʲ, tʃʲ, dʒʲ, mʲ, pʲ, bʲ, vʲ, kʲ, gʲ, fʲ).

Since the phonological status of gemination in Ukrainian remains debated (Moisiienko A. K., 2010), the system takes a neutral stance by treating all sequences of identical letters as two distinct phonemes of the same quality (tʲtʲ: життя "life" → /ʒɪtʲtʲɑ/). This approach reduces the number of unique phoneme categories without compromising transcription accuracy.

4.3 System Architecture

The algorithm is implemented in Python using regular expressions. Each rule for converting graphemes to phonemes is expressed as a regular expression of the form: <left context> <grapheme sequence> <right context> → <phoneme sequence> (e.g., <ле><г><к>о → ле<х>ко "easy"; неві<с><т><ч>ин → невіс<>чин, неві<с><ч>ин → неві<ш>чин "daughter-in-law")

Contexts are defined using lookaround assertions, allowing the system to apply rules conditionally based on surrounding characters. Rules are stored in ordered Python dictionaries and applied sequentially to the entire input without tokenization.

Because rule order can significantly affect output in rule-based systems, the rules follow a fixed and carefully designed sequence:

1. Mapping of specific graphemes (я, ю, є, ї, ь, й, щ) and grapheme combinations (e.g. дз, дж) to their phonemic equivalents (e.g. щука → шчука "pike", яблуко → jаблуко "apple", синю → синjу → син´у "blue").

2. Consonant cluster reduction (e.g. студентс´кий → студенс´кий "student", невістчин → невісчин "daughter-in-law")

3. Assimilation of voiced and voiceless consonants (e.g. борот´ба → бород´ба "fight", зсипати → ссипати "pour")

4. Assimilation of sibilants (e.g. л´отчик → л´оччик "pilot", погодишс´а → погодисс´а "agree", дочц´і → доцц´і "daughter")

5. Assimilation of palatalized consonants (e.g. с´огодн´і → с´огод´н´і "today")

6. Allophonic variation (e.g. вовк → воўк "wolf", гілка → г'ілка "branch")

An exception to the rule order is the grapheme sequence -ться (e.g. робиться "is being done"), which is converted into its phonemic representation

[1] https://www.internationalphoneticassociation.org/

Model	Sentence-Level Accuracy	Word-Level Accuracy	Ambiguous Word Accuracy	Unambiguous Word Accuracy	Mean-Macro F1 (Ambiguous Word Pairs)
ByT5 G2P	35.3%	87.7%	58.1%	94.8%	37.2%
Uk Accentor	16.6%	73.2%	41.6%	78.7%	28.7%
Uk Accentor Transformer	26.9%	83.4%	43.7%	96.3%	32.4%
Uk Word Stress (First)	41.5%	88.7%	**64.3%**	98.6%	**47.3%**
Uk Word Stress (Skip)	32.5%	86.0%	42.3%	98.6%	35.7%
ByT5 G2P + Word Stress (Skip)	**52.0%**	**92.5%**	61.0%	**98.7%**	46.7%
Uk Accentor + Uk Word Stress (Skip)	48.8%	91.9%	59.1%	**98.7**%	46.3%

Table 2: Comparison of model performance on the Ukrainian Lexical Stress Benchmark. Ambiguous words refer to those with identical spelling but different possible pronunciations, while unambiguous words have a single stress pattern per word form. All evaluations are conducted on words containing at least two vowels.

Step	Input Form	Applied Rule	Output Form
1	шістдесят	mapping of grapheme я	шістдесjат
2	шістдесjат	mapping of grapheme я	шістдес´ат
3	шістдес´ат	consonant cluster reduction (стд → сд)	шісдес´ат
4	шісдес´ат	assimilation of consonants (с → з)	шіздес´ат
5	шіздес´ат	allophonic variation (ш → ш’)	ш’іздес´ат

Table 3: Step-by-step transformation of the word "sixty" through the first five steps in the G2P pipeline.

Dataset	WER	Notes
Manually constructed dataset	1.23%	Incorrect cases
Automatically generated dataset	3.07%	Incorrect cases
Automatically generated dataset	6.15%	Incorrect + controversial cases
Baseline system	48.75%	Incorrect cases

Table 4: G2P system evaluation results.

(-ц´ц´а → робиц´ц´а), before the application of the consonant cluster reduction rule.

Each word undergoes multiple intermediate transformations, e.g. шістдесят → шістдесjат → шістдес´ат → шісдес´ат → шіздес´ат → ш’іздес´ат → ... → ʃʲɪzdɛsʲɑt "sixty" (see Table 3).

The system can be used in two modes: without word stress assignment or with word stress assigned by the automatic system or the user.

While no rules explicitly rely on stress, the position of stress must still be taken into account during rule formulation. In particular, some rules require explicit enumeration of morphemes (e.g. prefixes or roots), where the location of stress can alter the graphemic context. For example, in the case of ле́гко and легки́й "easy", the left context for the grapheme г can be either ле́ or ле.

4.4 Evaluation

The system was evaluated using two datasets, both of which were reviewed by expert linguists. The first 487-word dataset was manually constructed to maximize phonemic diversity, covering a wide range of segmental combinations. The second 553-word dataset was automatically generated from the VESUM dictionary (Rysin and Starko). The evaluation was performed using Word Error Rate (WER) as a metric. Because each word contained at most a single error type, Phoneme Error Rate (PER) was not calculated.

A baseline system implementing only simple letter-to-phoneme mappings was also evaluated. The results are as follows (see Table 4).

Incorrect transcriptions are those that violate the established rules of Ukrainian phonetics (Moisiienko A. K., 2010; Pohribnyi, 1984). For example: надзвонюватимемся "we will call" was transcribed as /nɑd͡zzwonʲuvɑtɪmɛmsʲɑ/, but the correct form is /nɑd͡zwonʲuvɑtɪmɛmsʲɑ/; ексдипломатів "former diplomats" was rendered as /ɛkzdɪpɫomɑtʲiw/, instead of the correct /ɛgzdɪpɫomɑtʲiw/.

Controversial transcriptions, on the other hand, involve cases not explicitly covered by the current rule set. For instance: Ваньчжоу "Wanzhou"

was transcribed as /vɑnʲʤʒou/, though /vɑnʲʤou/ is more accurate; Держспоживслужба "State Consumer Service" was transcribed as /dɛrzspoʒɪwsłuʒbɑ/ instead of /dɛrʒspoʒɪwsłuʒbɑ/.

Controversial cases were excluded from the first (manually constructed) evaluation dataset.

The lowest WER (1.23%) was observed on the first dataset, likely due to the exclusion of abbreviations and words with complex consonant clusters — two categories known to cause frequent errors. In the second dataset, the rates of incorrect and controversial transcriptions were equal, resulting in the second figure being twice the first.

The high WER (48.75%) of the baseline system reflects the large proportion of words with non-phonemic orthography in the evaluation datasets. Further evaluation on complete transcriptions of running text is planned.

5 Conclusion

In this work, we presented a modular approach to Ukrainian text-to-speech preprocessing that combines a rule-based phonemizer with a context-aware neural model for lexical stress prediction. Our system achieves strong results in both tasks: it reaches a low word error rate of 1.23% on a constructed phonemization dataset and shows competitive performance in lexical stress disambiguation, outperforming existing neural models and closely matching dictionary-based approaches. As part of this work, we also released the first publicly available benchmark dataset for evaluating Ukrainian lexical stress at the sentence level, providing a standardized foundation for consistent evaluation and future research.

Limitations

The proposed approach has several limitations that present opportunities for further enhancement.

First, while `ByT5 G2P` shows strong potential for context-driven disambiguation, its current performance on ambiguous words is limited by sparse coverage in the training data and the reliance on automatically labeled examples using Wav2Vec-based model. Enhancing heteronym representation in future training datasets remains a key direction for improvement.

Second, the current version of the phonemization system operates strictly at the word level and does not handle abbreviations or numerical expressions. These cases are excluded due to their irregular or ambiguous phonemic patterns, which require contextual or morphological analysis beyond the current system's scope. In the future, the system may be extended to operate on the sentence level.

Finally, neither pipeline accounts for non-standard language varieties, such as regional dialects.

Addressing these limitations could significantly enhance the coverage and applicability in real-world Ukrainian TTS applications.

Acknowledgments

We would like to express gratitude to the Talents for Ukraine project of Kyiv School of Economics for the grant on compute resources and to Tetiana Zakharchenko and Mariana Romanyshyn for their support and advice.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Preprint*, arXiv:2006.11477.

Dmytro Chaplinsky, Danylo Mysak, and Volodymyr Kyrylov. ipa-uk.

Stanley F Chen. 2003. Conditional and Joint Models for Grapheme-To-Phoneme Conversion. In *INTERSPEECH*, pages 2033–2036.

Diana Geneva, Georgi Shopov, Kostadin Garov, Maria Todorova, Stefan Gerdjikov, and Stoyan Mihov. 2023. Accentor: An Explicit Lexical Stress Model for TTS Systems. pages 4848–4852.

Kyle Gorman, Gleb Mazovetskiy, and Vitaly Nikolaev. 2018. Improving Homograph Disambiguation with Supervised Machine Learning. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Maria-Loulou Hajj, Martin Lenglet, Olivier Perrotin, and Gérard Bailly. 2022. Comparing NLP Solutions for the Disambiguation of French Heterophonic Homographs for End-to-End TTS Systems. pages 265–278.

Bondarenko V. V. et al. Moisiienko A. K., Bas-Kononenko O. V. 2010. *Contemporary Literary Ukrainian. Lexicology. Phonetics*. Znannia.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P For Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bohdan Mykhailenko. 2023. Ukrainian Accentor Transformer.

Marco Nicolis and Viacheslav Klimkov. 2021. Homograph Disambiguation With Contextual Word Embeddings For TTS Systems. *11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 222–226.

Saadin Oyucu and Ferdi Dogan. 2023. Improving Text-to-Speech Systems Through Preprocessing and Postprocessing Applications.

MI Pohribnyi. 1984. *Orthoepic Dictionary*.

Nikhil Prabhu and Katharina von der Wense. 2020. Frustratingly Easy Multilingual Grapheme-to-Phoneme Conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 123–127.

Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays. 2015. Grapheme-to-Phoneme Conversion Using Long Short-Term Memory Recurrent Neural Networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4225–4229. IEEE.

Andriy Rysin and Vasyl Starko. Large Electronic Dictionary of Ukrainian (VESUM). *Version*, 5(5):2005–2022.

Mykola Sazhok and Valentyna Robeiko. 2012. Bidirectional Text-to-Pronunciation Conversion with Word Stress Prediction for Ukrainian. In *Proc. All-Ukrainian Int. Conference on Signal/Image Processing and Pattern Recognition, UkrObraz*, pages 43–46.

Maria Shvedova and Arsenii Lukashevskyi. 2024. PluG: Corpus of Old Ukrainian Texts. `https://github.com/Dandelliony/pluperfect_grac`.

Yehor Smoliakov. 2022. Voice of America: Ukrainian ASR Dataset of Broadcast Speech.

Yehor Smoliakov and Bohdan Mykhailenko. 2022. Ukrainian Accentor.

Oleksiy Syvokon. 2022. Ukrainian Word Stress.

Paul Taylor. 2005. Hidden Markov Models For Grapheme To Phoneme Conversion. In *Interspeech*, pages 1973–1976.

NAS of Ukraine Ukrainian Lingua-Information Foundation. 2008. Dictionaries of Ukraine Online.

Daan van Esch, Mason Chua, and Kanishka Rao. 2016. Predicting Pronunciations with Syllabification and Stress with Recurrent Neural Networks. *Interspeech 2016*, pages 2841–2845.

Wikimedia. Wikimedia Downloads. Wikimedia Foundation. Accessed: 2024-12-21.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2019. Grapheme-to-Phoneme Conversion with Convolutional Neural Networks. *Applied Sciences*, 9(6):1143.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 Model for Massively Multilingual Grapheme-To-Phoneme Conversion. *Preprint*, arXiv:2204.03067.

Appendix A. Symbol inventory

Ukrainian transcription symbols	IPA symbols
Vowels	
і	i
и	ɪ
е	ɛ
у	u
о	o
а	ɑ
Nasal consonants	
м	m
м’	mʲ
н	n
н´	nʲ
Plosives	
п	p
п’	pʲ
б	b
б’	bʲ
т	t
т´	tʲ
д	d
д´	dʲ
к	k
к’	kʲ
ґ	g
ґ’	gʲ
Approximants	
в (bilabial)	w
в (labio-dental)	v
в’	vʲ
й	j
Fricatives	
ф	f
ф’	fʲ
с	s
с´	sʲ
з	z
з´	zʲ
ш	ʃ
ш’	ʃʲ
ж	ʒ
ж’	ʒʲ
х	x
х’	xʲ
г	ɦ
г’	ɦʲ
Affricates	
ц	ʦ
ц´	ʦʲ
дз	ʣ
дз´	ʣʲ
ч	ʧ
ч’	ʧʲ
дж	ʤ
дж’	ʤʲ
Trill & tap (flap) consonants	
р	r
р´	rʲ
Lateral approximants	
л	ɫ
л´	lʲ

Table 5: Symbol inventory